



**The Role of Embodiment in Human-Robot Interaction:
Modelling Attentional Behaviours through
Cognitive Neuro-Robotics**

by

Massimiliano Lorenzo Cappuccio

MA (philosophy), PhD (philosophy)

Thesis submitted to Flinders University
for the degree of

PhD by Prior Published Work

College of Science & Engineering
27 September 2025

Flinders University – Adelaide
College of Science & Engineering

HDR dissertation

**The Role of Embodiment in Human-Robot Interaction:
Modelling Attentional Behaviours through
Cognitive Neuro-Robotics**

by

Massimiliano Lorenzo Cappuccio

HDR student number: 2336584
Flinders Unique Student Identifier: capp0034

HDR advisors: Prof Richard Leibbrandt, Prof Paulo Santos
HDR Topic code: COMP9999 (Computer Science)
Thesis submitted for examination on 6 February 2025
Thesis approved on 27 September 2025

*This work is dedicated to my uncles
Antonio Galdi and Silvana Ferrigno*

*When I was a kid, they made me read Asimov's stories
...and almost 40 years later I am still dealing
with the same robot problems!*

Table of Contents

0.1 Abstract	p. 13
0.2 Declaration of Integrity	p. 15
0.3 Acknowledgments	p. 17
0.4 My Journey as a Computer Science Scholar	p. 19
0.5 List of Publications and Originality Statement	p. 25
0.6 List of Acronyms	p. 37
0.7 List of Tables and Figures	p. 39

CONTEXTUAL STATEMENT	p. 43

Chapter 1 – The Embodiment of Sensorimotor Skills.

Action-Guiding Attention in Expert Performance	p. 45

1.0 Preamble	p. 47

1.0.1 Themes and problems (p. 47) – 1.0.2 Research outcomes covered in this chapter (p. 47) – 1.0.3 Author's publications used in this chapter (p. 48)

1.1 The Interdisciplinary Background	p. 49

1.1.1 The field of Interaction Studies (p.49) – 1.1.2 Affective Computing (p. 50) – 1.1.3 Social Robotics (p. 51) – 1.1.4 Cognitive Robotics (p. 52).

1.2 A Theoretical Framework for Embodied Cognitive Robotics	p. 55

1.2.1 Embodiment as a general theory of cognition (p. 55) – 1.2.1.1 Body type (p. 55) – 1.2.1.2 Expressivity (p. 56) – 1.2.1.3 Gesture (p. 56) – 1.2.1.4 Eyes and gaze (p. 56) – 1.2.1.5 Proxemics (p. 56) – 1.2.1.6 Tactility and haptics (p. 57) – 1.2.1.7 Voice and speech (p. 57) – 1.2.2 The three basic tenets of ECT (p. 59) – 1.2.2.1 Cognition reflects the dynamic interplay of body, brain, and mind (p. 56) – 1.2.2.2 Sensorimotor experiences scaffold abstract and symbolic cognitive processes (p. 57) – 1.2.2.3 Cognition is ecologically situated and attuned to the environment in which it has evolved (p. 58) – 1.2.3 The two radical tenets of ECT (p. 59) – 1.2.3.1 First Radical Tenet: Basic cognition is non-representational (p. 59) – 1.2.3.2 Second Radical Tenet: Cognitive Processes encompass non-organic/non-cerebral vehicles (p. 62) – 1.2.4

Broadening the scope of ECT: The 4E cognition framework (p. 63) – 1.2.5 Turing's legacy (p. 64) – 1.2.6 Outcomes, limitations, and future research (p. 65) – *Core Publication 1 [CP1]*.

1.3 Modelling Attention and Control in Skilful Performance p. 67

1.3.1 Embodied cognition applied to cognitive robotics (p. 67) – 1.3.2 Embodied cognitive robotics & its key ideas (p.67) – 1.3.2.1 Physical presence (p. 67) – 1.3.2.2 Coupling (p.67) – 1.3.2.3 Adaptability (p.68) – 1.3.2.4 Practical intelligence (p. 68) – 1.3.2.5 Autonomy (p. 68) – 1.3.2.6 Empathy (p. 68) – 1.3.2.6 Cognitive parsimony (p. 68) – 1.3.3 Attention and its embodiment (p. 69) – 1.3.3.1 The embodiment of attentional cues (p. 69) – 1.3.3.2 Embodiment as a tool for attention management (p. 69) – 1.3.3.3 Optimization of cognitive load distribution via attention management (p. 69) – 1.3.3.4 Human attitudes: Trust, engagement, & relationship building (p. 70) – 1.3.4 A neuro-cognitive model of attention management in skilful activities (p. 70) – 1.3.5 Challenges for the robot implementation of Active Inference (p. 74) – 1.3.6 Outcomes, limitations, and future research (p. 76) – *Core Publication 2 [CP2]*.

1.4 Conclusions and significance for the field of HRI p. 79

Chapter 2 – The Embodiment of Social Skills.

Social Attention for Interactive Engagement p. 81

2.0 Preamble p. 83

2.0.1 Themes & problems (p. 83) – 2.0.2 Research outcomes covered in this chapter (p. 83) – 2.0.3 Author's publications used in this chapter (p. 84).

2.1 Social Attention Deficits in Autism p. 85

2.1.1 Towards an embodied understanding of social attention (p. 85) – 2.1.2 Social cognition deficits in the Autistic Spectrum Disorder (p. 86) – 2.1.3 Neurocognitive theories of ASD (p. 88) – 2.1.4 An embodied approach to autism (p. 91).

2.2 Robotic Approaches to the Diagnosis and Treatment of Autism p. 93

2.2.1 Technological approaches to the diagnosis & treatment of autism: a survey (p. 93) – 2.2.2 ASD intervention with social robots: five areas (p. 94) – 2.2.3 Main applications of robotics in ASD studies (p. 95) – 2.2.4 Outcomes, limitations, and future research (p. 97) – *Core Publication 3 [CP3]*.

2.3 A Semi-Autonomous System to Measure and Train Social Attention p. 99

2.3.1 A project conceived and developed in the UAE (p. 99) – 2.3.2 Engaging with children through a robotic avatar (p. 100) – 2.3.3 A semi-autonomous adaptive system to solicit and measure social attention (p. 103) – 2.3.3.1 Semi-autonomous & socially interactive (p. 104) – 2.3.3.2 Adaptive & professionally calibrated (p. 104) – 2.3.4 Components and methods (p. 105) – 2.3.4.1 Attention assessment factors (p. 107) – 2.3.5 Measuring attention via a cumulative score system (p. 108) – 2.3.5.1 Joint attention score (p. 108) – 2.3.5.2 Joint attention score (p. 109) – 2.3.5.3 Emotion detection score (p. 110) – 2.3.5.4 Sound response score (p. 110) – 2.3.6 Therapist-System Agreement (p. 111); 2.3.7 An adaptive and interactive system (p. 112) – 2.3.8 Challenges and solutions (p. 114) – 2.3.8.1 Adaptive calibration (p. 114) – 2.3.8.2 Variety (p. 116) – 2.3.8.3 Safety (p. 117) – 2.3.8.4 Screen interface (p. 117) – 2.3.9 Outcomes, limitations, and future research (p. 118) – 2.3.9.1 Long-term interaction progress (p. 118) – 2.3.9.2 Results (p. 119) – 2.3.9.3 Human assessment (p. 119) – 2.3.9.4 Limitations (p. 120) – 2.3.9.5 Future developments (p. 120) – *Core Publication 4 [CP4]*.

2.4 Conclusions and significance for the field of HRI	p. 121

Chapter 3 – The Embodiment of Ethical Skills.

Moralizing Attention and Virtue Cultivation	p. 123

3.0 Preamble	p. 125

3.0.1 Themes and problems (p. 125) – 3.0.2 Research outcomes covered in this chapter (p. 126) – 3.0.3 Author's publications used in this chapter (p. 127).

3.1 Attitudes towards Robots: Tolerance and Anthropomorphism	p. 129

3.1.1 Towards an embodied understanding of human-robot relationships (p. 129) – 3.1.2 Making space for Tolerance (p. 129) – 3.1.3 The ATTA-bot model of Attitudes towards robots (p. 135) – 3.1.4 The risks associated with Intolerance (p. 139) – 3.1.5 A peculiar attitude: Anthropomorphism (p. 143) – 3.1.6 The dark side of anthropomorphism: unsafe relationships (p. 145) – 3.1.7 The bright side of anthropomorphism: advantageous applications (p. 146) – 3.1.8 Anthropomorphism for robots as empathy for artificial agents (p. 150) – 3.1.9 Outcomes, limitations, and future research (p. 151) – *Core Publication 5 [CP5]* – *Core Publication 6 [CP6]*.

3.2 Ethical Attitudes towards Robots: Robots that Cultivate Character	p. 153

3.2.1 Vicious relationships with robots (p. 153) – 3.2.2 Morally corruptive robots? (p. 154) – 3.2.3 Virtuous Robotics: a socio-technical framework to become better humans (p. 155) – 3.2.3.1 The principles of Virtue Ethics (p. 156) – 3.2.3.2 Operationalising Virtue Ethics (p. 158) – 3.2.3.3 Robotics applications of Virtue Ethics (p. 160) – 3.2.4 A New Approach to Robot Ethics (p. 162) – 3.2.4.1 Objectives (p. 163) – 3.2.4.2 Roles (p. 164) – 3.2.5 Outcomes, limitations, and future research (p. 164) – 3.2.5.1 Dependency and Over-Reliance (p. 164) – 3.2.5.2 Deception and Anthropomorphism (p. 164) – 3.2.5.3 Normative Ambiguity and Cultural Bias (p. 165) – 3.2.5.4 Domain-Specific Fragility (p. 165) – 3.2.5.5 Limitations of General-Purpose AI (p. 165) – 3.2.5.6 Mitigation Strategies (p. 165) – *Core Publication 7 [CP7]* – *Core Publication 8 [CP8]* – *Core Publication 9 [CP9]*.

3.3 Growing with Robots: Autonomous Virtuous Agents in Education	p. 167

3.3.1 Virtuous applications of anthropomorphism in the classroom environment (p. 167) – 3.3.2 Smart technologies, robot coaches, and the treatment of obesity (p. 168) – 3.3.3 Robots as teachers: opportunities and risks (p. 169) – 3.3.3.1 Psychological damage (p. 170) – 3.3.3.2 Privacy (p. 170) – 3.3.3.3 De-skilling (p. 170) – 3.3.3.4 Authority (p. 171) – 3.3.3.5 Lack of emotional intelligence (p. 171) – 3.3.4 Challenges and solutions (p. 172) – 3.3.4.1 Trust issues (p. 172) – 3.3.4.2 Teacher's responsibility (p. 172) – 3.3.4.2 Recording and sharing information (p. 173) – 3.3.4.3 Licence to discipline and report (p. 173) – 3.3.5 Discipline and report: robots as inspectors (p. 173) – 3.3.5.1 Background and assumptions (p. 174) – 3.3.5.2 Task design (p. 174) – 3.3.5.3 Composition (p. 175) – 3.3.5.4 Results (p. 175) – 3.3.6 Challenges and solutions (p. 176) – 3.3.6.1 Task definition (p. 176) – 3.3.6.2 Population sample selection (p. 176) – 3.3.6.3 Robot design (p. 177) – 3.3.7 Outcomes, limitations, and future research (p. 177) – *Core Publication 10 [CP10]* – *Core Publication 11 [CP11]*.

3.4 Conclusions and significance for the field of HRI	p. 179

4. Conclusion	p. 181

4.1 Answering the research questions (p. 181) – 4.2 Original contribution to the field (p. 183) – 4.3 Policy implications (p. 185) – 4.4 The way forward: limitations and future directions (p. 186)

APPENDIX

p. 189

Support Publication 1 [SP1] – Support Publication 2 [SP2] – Support Publication 3 [SP3] – Support Publication 4 [SP4] – Support Publication 5 [SP5] – Support Publication 6 [SP6] – Support Publication 7 [SP7] – Support Publication 8 [SP8] – Support Publication 9 [SP9] – Support Publication 10 [SP10] – Support Publication 11 [SP11] – Support Publication 12 [SP12] – Support Publication 13 [SP13] – Support Publication 14 [SP14] – Support Publication 15 [SP15] – Support Publication 16 [SP16] – Support Publication 17 [SP17] – Support Publication 18 [SP18] – Support Publication 19 [SP19].

Reference list

p. 191

0.1 Abstract

Talks of “embodiment” are common in the field of robotics. But, beyond the obvious fact that robots need actuators and sensors to physically operate in the environment, what is required to create a truly embodied robot? I will adopt a Human-Robot Interaction perspective to explore the role of embodiment in artificial cognitive functions. Defending the tenets of embodied & enactive cognition theory, I will argue that a robot’s body is not merely a material implementation or cosmetic addition to its cognitive functions: it is their essential constituent. A well-designed body is indispensable to develop intelligent capabilities via rich interactions with the environment.

Since robot intelligence expresses itself through skilful behaviour, the dissertation’s three chapters examine distinct domains of skilled expertise: (1) sensorimotor skills, concerning the preconditions for competent autonomous or team-based action; (2) social skills, concerning the ability to interact meaningfully with humans, meet expectations, gain trust, and build lasting relationships; and (3) moral and pedagogical skills, concerning how robots can promote human self-improvement. The chapters’ common theme is the embodiment of attention in HRI. I analyse attentional behaviours—eye contact, gaze following, gaze aversion, joint attention—in their action-guiding, communicative, and moralizing roles. Attention, I contend, is not an abstract intellectual stance but a concrete, embodied form of engagement, whose effectiveness depends on the robot’s physical presence, poise, and behavioural dynamics.

By reviewing computational models, data presented in the empirical literature, and the results of the experimental studies with humanoid robots conducted by me and my co-authors, I show that effectively controlling human and robotic attention is crucial for designing successful interactions. I will draw on the research by experts in enactivist robotics like Minoru Asada, Angelo Cangelosi, Antonio Chella, Ron Chrisley, Luisa Damiano, Ezequiel di Paolo, Tom Froese, Giorgio Metta, Rolf Pfeifer, Giulio Sandini, Jun Tani, Paul Vernon, and Tom Ziemke. Like theirs, the approach I adopt is “bottom-up” and inspired by the evolutionary and developmental paths of natural cognition.

Ultimately, this dissertation advances a general framework to design robotic systems optimised for human interaction through embodied attentional behaviours.

0.2 Declaration of Integrity

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university;
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University;
3. and to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.
4. has been completed without the use of generative artificial intelligence tools.

No person or organization gave me or offered their help to write this dissertation.

No professional editor was used to write this dissertation.

Massimiliano L. Cappuccio,
Canberra, 27 September 2025

0.3 Acknowledgments

This research was supported by an *Australian Government Research Training Program (RTP)* scholarship doi.org/10.82133/C42F-K220.

Also, I would like to acknowledge the funding bodies that supported the research activities that led to this dissertation. Many of the publications reviewed in this work were supported by either of two grants that sponsored my research activity during different periods: for the period 2016-2018, by a UPAR grant from the UAE Government (award number 31H125); for the period 2019-2021, by a Minerva grant from the U.S. Air Force Office of Scientific Research (award number FA9550-18-1-0039).

Also, I would like to thank the persons that sustained my growth as a computer scientist and an HRI researcher, especially the colleagues who provided unfailing professional advice and intellectual guidance. Those who generously shared their expertise and offered me concrete opportunities to develop my technological skills in laboratorial and field work played a crucial role in making me the established social/cognitive roboticist that I am today.

First of all, I am in debt with the colleagues who encouraged my scientific interests during my eight years in the United Arab Emirates, starting with Dr Fady Alnajjar and all the collaborators and supporters of the AI & Robotics Lab, the pioneering, multi-awarded oasis of science and technology that Dr Alnajjar and I contributed to create at UAE University in 2016 and that Dr Alnajjar has successfully directed since then.

Also, with a profound sense of intellectual gratitude and professional friendship, I would like to mention Friederike Eyssel, Mohamad Eid, Christoph Bartneck, and –again– Fady Alnajjar, for their passionate and voluntary dedication as board members and recurrent guest speakers of the annual *Joint UAE Symposium on Social Robotics*: known by the aficionados as “the JSSR conference”, this is the tremendously engaging, diverse, and fun series of meetings with academia and industry that we established ten years ago in the Arab Gulf region with the generous patronage of UAE University and NYU Abu Dhabi. Today, after nine editions of this successful franchise (including the special edition held in 2022 at the Australian pavilion as part of EXPO Dubai), JSSR is known by our professional community as a regular appointment with the most innovative and socially-valuable applications of autonomy.

Also, I would like to mention the guidance I received from the senior colleagues of University of New South Wales Canberra as part of the activities of the Trusted Autonomy group, the AI Institute, and the AI Hub – Matthew Garratt, Michael Barlow, and Hussein Abbass– as well as my colleagues in the Sydney-based research group who were responsible with me for the interdisciplinary project on “Virtuous Robotics”: Omar Mubin, Mari Velonaki, Eduardo B Sandoval, and Mohammed Obaid. This collective project led, in 2020, to the publication of the volume *Robots in Education* and, in 2021, to the publication of a special issue of *International Journal of Social Robotics*, the discipline-defining journal that I had the honour to serve and support first as an author, then as a guest editor and a referee, and finally as an associate editor.

Also, my knowledge of robotics has grown tremendously during the periods of applied work and intensive study that I spent, as a visiting researcher, at Okinawa Institute of Science & Technology – a true sanctuary of discovery and innovation and a unique destination for the lovers of interdisciplinarity. There, I greatly benefitted from the illuminating mentorship of Jun Tani-sensei and the advice of the members of the Neuro-Cognitive Robotics Lab that he directs and from the inspiring collaboration with Tom Froese and his Embodied Cognition Science Unit. Also, I am deeply grateful to the colleagues of the Italian Institute of Technology of Genova, whose visionary work with iCub has constantly stimulated my interest for humanoid robotics during the last 20 years: in Genova, I found incredible opportunities of learning and professional development thanks to Giorgio Metta, Giulio Sandini, Agnieszka Wykowska and Alessandra Sciutti.

Last but not least, this dissertation would not have been possible without the prudent advice and the constant encouragement that I received at Flinders University from Richard Leibbrandt and Paulo Santos.

I hope to have many more intellectual exchanges and opportunities of collaboration with all of them during the years to come.

Massimiliano L. Cappuccio,
Canberra, 27 September 2025

0.4 My Journey as a Computer Science Scholar

This doctoral dissertation is the second I write in my academic career: my first PhD dissertation was in philosophy and defended a theoretical thesis framed within the interdisciplinary domain of cognitive science. Today, my second dissertation is in computer science, with a distinct focus on Human-Robot Interaction (HRI), i.e. the branch of Human-Machine Interaction research that studies the interaction between humans and robots. This work attests my decision to continue in a different capacity, as a designer and developer of autonomous technologies, the research journey that I started more than twenty years ago as a cognitive philosopher.

The decision to give an even more technological trajectory to my research activity is motivated by the urge I've progressively felt to complement the conceptual analysis of cognitive functions with the creation of empirically-informed computational models of the same functions. Complementing the work of AI theorists, computer scientists engaged in HRI research use synthetic methods to recreate the processes and systems underpinning human mental functions, like language processing (Cangelosi 2001). This resembles the way philosophers use thought experiments to examine the necessary structure, or the logical form, of such functions (Di Paolo, Noble & Bullock 2000): roboticists design, train, and optimize cognitive agents to simulate the chaotic processes and systems underlying natural cognition. The long periods I spent in the laboratories of well-established roboticists in Australia, Italy, Japan, and the UAE have persuaded me that, no less than other forms of scientific inquiry, synthetic methods offer a very effectual way to formulate, test, and validate competing hypotheses about the functioning of the mind and intelligent systems. As argued by Chrisley (2009) and Damiano & Stano (2018), by systematically adopting the synthetic methods of cognitive robotics and social robotics, HRI allows researchers to highlight the inherently embodied and social nature of the human mind, which has always been the target of my inquiry.

My project is sustained by two methodological pillars: on the one hand, the *phenomenological descriptive analysis* of human experience (which is inspired by the seminal work of major 20th century philosophers like Edmund Husserl, Maurice Merleau-Ponty, and Martin Heidegger and aspires to exhaustively examine the behaviours and the emotions that accompany our cognitive processes in this or that real-life scenario (Cappuccio, Wheeler 2012 [SP3]; Coeckelbergh 2011; Dreyfus 2005, 2007; Andler 2006;

Wheeler 2008; Tani 2017); on the other, the *synthetic methods* deployed by cognitive robotics, which systematically attempts to generate these behaviours and even emotions by artificial means (Damiano & Cañamero 2010; Stano, Kuruma & Damiano 2018). I have always assumed that both pillars are indispensable, as they legitimately illuminate complementary aspects of our cognitive life (one available through unmediated experience, the other through specific techniques and scientific analysis), thus reflecting mutual kinds of expectations for the humanists and the scientists who study human cognition.

Not only does my research project raise the same theoretical questions that inspired my graduate studies more than twenty years ago. It also responds to the same moral aspirations. My research has always been inspired by the principle that, when they are understood as opportunities for self-understanding and emancipation, robotics and AI can become vehicles of human flourishing, promoting both moral and cognitive development (Cappuccio et al 2020a [CP7], Tonkens 2012, Torrance 2011, Vallor 2016). This idea became perfectly clear to me the day I visited a facility for autistic children in Abu Dhabi to test a humanoid robot. My colleagues and I had specifically programmed it to help ASD children establish reciprocal attention and maintain robust social engagement over time (Alahbabi et al. 2017 [SP13], Alnajjar et 2020 [CP4]). Despite the initial uncertainties, we were overwhelmed with joy when the clinical personnel and the parents of the children shared with us their amazement, as the robot had started eliciting unprecedented social responses in the little students. This deeply touching experience persuaded me (probably more than the aprioristic arguments normally used in philosophical debate) that HRI research is a morally valuable enterprise because, among other things, it can have a significant impact on the mental well-being of others while helping us understand the mechanisms underpinning their cognitive processes.

Autonomous technologies –including primarily humanoid robots, conversational agents, and self-driving vehicles– have always been at the very centre of my research interests. From the earliest years of university, my humanistic education largely intertwined and overlapped with computational sciences and artificial intelligence. The training I received as a cognitive philosopher and as a philosopher of technology threw the foundations for my subsequent applied research in robotics and AI. Over the years, the progressive specialization of my scholarship equipped me with a greater and greater familiarity with the concepts, themes, and methods of HRI research. In 2001 and 2002, I was an Erasmus exchange student based at the Institute for Language Logic and Computation of the University of Amsterdam. In the Netherlands I studied under the supervision of Prof Michiel Van Lambalgen and Prof Martin Stokhof, two influential logicians whose original work has

shown effective ways to combine the study of the cognitive foundations of formal logic with the study of the logical foundations of AI. In this context, I was for the first time introduced to the thought of Francisco Varela and the philosophy of the “embodied mind” (Varela et al. 2018), while having the opportunity to learn for the first time about the theoretical foundations of computability and Alan Turing’s seminal work on artificial intelligence (Turing 1936, 1950).

In Amsterdam, I intensively studied both mathematical logic and computational theories of mind, developing a great passion for the technological side of the cognitive sciences and a firm commitment to interdisciplinarity that still defines my career’s trajectory. In 2003, in visit to Milan, Giuseppe Longo (today emeritus professor of informatics and mathematics at École Normale Supérieure de Paris) read some of my notes on the cognitive preconditions of computability and invited me to submit a more articulated account of my ideas to the special issue “Géométrie et Cognition” that he was editing for the *Revue de Synthèse*, the prestigious 130-year-old interdisciplinary journal of the École Normale Supérieure. My paper was eventually accepted and my academic advisors in Milan, Carlo Sini, Rossella Fabbrichesi and Corrado Sinigaglia, encouraged me to elaborate on my original idea to dedicate my Master dissertation to the semiotic aspects of Alan Turing’s 1936 theory of computationalism. The thesis, that I defended on 25 June 2004, was subsequently awarded by the Istituto Italiano per gli Studi Filosofici as the best philosophy dissertation of year 2004. After a thorough work of revision and restyling, in 2006 it was published as a book, receiving positive reviews and praises.

From 2005, a post-graduate scholarship allowed me to spend two years at the École Normale Supérieure de Paris as a visiting researcher. In Paris, I deepened my knowledge of computationalism and AI theory under the supervision of Prof Giuseppe Longo. Importantly, I also joined CREA at Ecole Polytechnique de Paris, the transdisciplinary research unit co-founded by Francisco Varela with his closest collaborators and some of his pupils. In this context, I developed a greater familiarity with embodied/enactive cognition theory and neurophenomenology (as advanced in the volume *Naturalizing Phenomenology*, Petitot et al 2008). During the following years, my work on mirror neurons and empathy gave me a pivotal opportunity to familiarize with the models of “enactive robotics” developed by Giorgio Metta and Giulio Sandini at the Italian Institute of Technology, often in collaboration with Paul Vernon (Sandini, Metta, & Vernon 2007). I was enthused by their line of study with iCub (Metta et al 2008), which represents one of the first attempts to model empathy by means of a cognitive architecture that replicates the function of mirror neurons (embodied simulation): a neural network trained to learn the transformative goals achieved by its actions and subsequently to use such sensorimotor knowledge to

recognize the goals of actions performed by others (Metta et al. 2006). In 2009, I spent almost a year at the University of Stirling on a post-doctoral research fellowship awarded by the Royal Society of Edinburgh and Caledonian Foundation. My work in Scotland was supervised by philosopher of robotics Michael Wheeler, who introduced me to the work on distributed cognition by Andy Clark and the study on communicative cricket robots developed by Barbara Webb at the nearby University of Edinburgh, while encouraging me to extensively study the principles of situated robotics as defined by MIT roboticist Rodney Brooks and by Berkeley philosopher of AI Hubert Dreyfus, who have deeply influenced my following interests for HRI.

Since 2011, for eight years, I was employed as an Assistant Professor (2011-2016) and then as an Associate Professor (2016-2018) at UAE University, the national university of the United Arab Emirates and one of the most prestigious research-intensive universities in the MENA region. Here, I taught and mentored several hundreds of computer science students, designing for them original course contents and creating the first interdisciplinary program in Cognitive Science. In 2014, our college's Dean asked me to become the first coordinator of the Cognitive Science Program of UAE University, as well as the founder and first director of the Cog Sci Lab. In 2016, I became the co-founder, with Prof Fady Alnajjar, of the AI & Robotics Lab of the same university, where I have conducted some of my most rewarding research projects. I was awarded several small-medium grants by governmental funding bodies and three large research grants by the federal government of UAE, including the UPAR2016 research grant I was awarded in 2016 to coordinate, as its Principal Investigator, the 3-year research project titled "Sympathy for Suhail: An experimental study on empathy and embodied cognition to design the first UAE social robot". This project aimed to create an empathic enactive humanoid robot for the diagnosis and treatment of autistic children.

At the beginning of 2019, I left the Middle-East to take on a research-only position at the prestigious University of New South Wales, ranked first in Australia and 19th in the world. For more than six years I have been employed in the School of Engineering & Technology of UNSW Canberra, where I became co-convenor of the course "Cybersecurity ethics" and contributed to design courses like "Strategic and Social Implications of AI". Also, I served as the deputy director of Values in Defence and Security Technology, an interdisciplinary group specialising in strategic implications of AI-based and autonomous technologies and collaborating with the Australian government and some important technological firms. In September 2022 I became the chief investigator of a Strategic Policy grant awarded by the Australian Department of Defence to study the attitudes and perceptions towards autonomous technologies in the Indo-Pacific countries and the cultural factors that shape

them. In 2022, I was awarded the AI Faculty Seed Funding from the UNSW AI Institute to coordinate a research project on the limitations of autonomous vehicles. Since 2022, I am a member of the Trusted Autonomy research group in my school and a founding member of the AI Institute of UNSW and its local branch in Canberra, the UNSW Canberra AI Hub. At UNSW Canberra, I regularly collaborate with the laboratory of robotics and, at UNSW Paddington (in Sydney) I am affiliated with the National Facility for Human-Robot Interaction Research and the Creative Robotics Lab founded by Mari Velonaki.

My publishing activity started in 2003, when, as a Masters student, I published my first paper in *Revue de Synthèse* (the 120-year-old peer-reviewed journal of the École Normale supérieure of Paris), I edited my first special issue of journal (on digital ethics), and I organized my first academic conference. To date, I have published 38 papers in more than 20 different peer-reviewed journals and 33 peer-reviewed chapters in edited books for leading international publishers (e.g., MIT, UPO, CUP), 10 conference proceedings indexed by technological professional associations like IEEE and AISB, 8 articles in non-refereed journals and popular press (like Harvard Business Review), 2 translations from English and French, and 1 technical report commissioned by the Royal Australian Air Force. I have edited and curated 7 collective books for 5 different publishers and 16 special issues for 7 different peer-reviewed journals. I have single-authored a monograph in Italian (about Alan Turing's theorem) and co-authored another in English (a systematic introduction to robots in education). The MIT Press, for which I published several book chapters, 1 edited collection, and prospectively a monograph (forthcoming) is the most prestigious publisher in my field. My third monograph ("the Skilled Mind: Labour, Training, and Leadership in the Age of AI") is already under contract with the MIT Press and expected to be published in 2026. During the last four years the number of my citations has been increasing steadily as my research is gaining a growing consideration internationally: my Google citations and H-index (21) are above the average of my colleagues at the same rank and my performance compares positively with most of my colleagues in the School of Engineering & Technology at UNSW Canberra.

My enduring involvement in computer science and HRI reverberates in all my projects. I am a recurrent Visiting Researcher at the Laboratory of Neuro-cognitive robotics, Okinawa Institute of Science and Technology (director: Prof. Jun Tani), and at the Laboratory of Social Robotics of the Italian Institute of Technology in Genova (scientific director: Prof. Giorgio Metta). Also, I produced several conference proceedings in the field of computer science, with a focus on HRI. I regularly offer my experienced service as a referee for peer-reviewed journals such as *International Journal of Social Robotics*, *AI and Society*, *Minds and*

Machines, International Journal of Machine Consciousness, Adaptive Behaviour, IEEE Transactions on Technology and Society, and several other IEEE and ACM journals and conferences specializing in robotics and HRI (e.g., IEEE Symposium Series on Computational Intelligence, HRI2021).

During the last fifteen years, I successfully conceived, promoted, and supervised a number of high-impact international conferences (including 8 editions of the Joint UAE Symposium on Social Robotics organized by UAE University in collaboration with NYU Abu Dhabi and UNSW) and several professional workshops (like the training workshop “Doing Things with Robots” at UAE University in 2016).

0.5 List of Publications and Originality Statement

This section lists 30 publications that are meant to be evaluated as part of my doctoral work: *Core Publications* [CPn], included as part of the Contextual Statement, and 19 *Support Publications* [SPn], included in the Appendix. My original contribution to the CPs is stated below and attested by the co-authors signed declarations submitted to the HDR office.

Core Publications

Core Publication #1 [CP1] (included at the end of Chapter 1, Part 2)

Cappuccio M.L. (2015),

“The Seminal Speculation of a Precursor: Elements of Embodied Cognition and Situated AI in Alan Turing”

in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Springer - Synthese Library, 2015.

Publication type: chapter in a peer-review edited collection.

Original contribution: I am the sole author.

Disciplinary relevance: the volume is part of well-established series in AI, computer science, logic, and epistemology (*Fundamental issues of artificial intelligence*, Synthese Library). All submitted contributions have passed a strict double-blind peer-review process.

* * * * *

Core Publication #2 [CP2] (included at the end of Chapter 1, Part 3)

Cappuccio M.L., Kirchhoff M.D., Alnajjar F.S., Tani J. (2019),

“Unfulfilled Prophecies in Sport Performance. Active Inference and the Choking Effect”

Journal of Consciousness Studies, 27(3-4), pp. 152-184.

<https://www.ingentaconnect.com/contentone/imp/jcs/2020/00000027/f0020003/art00008>

Publication type: paper in Q1 peer-reviewed journal.

Original contribution: I am the first author and I have contributed the greatest portion of the research and writing work, including conceptualization and data collection.

Disciplinary relevance: according to Scopus-Scimago, *Journal of Consciousness Studies*, is a Q2 Journal (H-index: 77) in the areas of “computer science” and “artificial intelligence” (2019-2020).

* * * * *

Core Publication #3 **[CP3]** (included at the end of Chapter 2, Part 2)

Alnajjar F.S., Cappuccio M.L., Mubin O., Arshad R., Shahid S. (2021b),
“Humanoid robots and autistic children: a review on technological tools to assess social attention and engagement”

International Journal of Humanoid Robotics, World Scientific Publishing Co.

<https://www.worldscientific.com/doi/10.1142/S0219843620300019>

Publication type: paper in Q2 peer-reviewed journal.

Originality contribution: I am the first author with Fady Alnajjar. Also, I am the corresponding author. I have contributed the greatest portion of the research and writing work, including conceptualization, data collection, and analysis.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Humanoid Robotics* is a Q2 Journal (H-index: 44) in the areas of “computer science” and “artificial intelligence”.

<https://www.scimagojr.com/journalsearch.php?q=4700152208&tip=sid&clean=0>

* * * * *

Core Publication #4 **[CP4]** (included at the end of Chapter 2, Part 3)

Alnajjar F.S., Cappuccio M.L., Renawi A., Mubin O., Kiong Loo C. (2020),
“Personalized Robot Interventions for Autistic Children: an Automated Methodology for Attention Assessment”

International Journal of Social Robotics, 13(1), Springer Verlag.

<https://doi.org/10.1007/N12369-020-00639-8>

Publication type: paper in Q1 peer-reviewed journal.

Original contribution: I am the first author with Fady Alnajjar and I have contributed the greatest portion of the research and writing work, including conceptualization, data collection, and analysis.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) in the areas of “computer science” and “human computer interaction”.

<https://www.scimagojr.com/journalsearch.php?q=19500157063&tip=sid&clean=0>

* * * * *

Core Publication #5 **[CP5]** (included at the end of Chapter 3, Part 1)

Cappuccio M.L., Galliot J.C., Lanteri A., Eyssel F. (2024),

“Autonomous Systems and Technology Resistance: New Tools for Monitoring of Acceptance, Trust, and Tolerance”

International Journal of Social Robotics, 16(1), Springer Verlag.

Publication type: paper in Q1 peer-reviewed journal.

Originality contribution: I am the first author and I have contributed the greatest portion of the research and writing work, including conceptualization, data collection, and analysis.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) in the areas of “computer science” and “human computer interaction”.

<https://www.scimagojr.com/journalsearch.php?q=19500157063&tip=sid&clean=0>

* * * * *

Core Publication #6 **[CP6]** (included at the end of Chapter 3, Part 1)

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021),
“Saving Private Robot: Risks and Advantages of Anthropomorphism in Soldier-Agent Teams”
International Journal of Social Robotics, 14(1), Springer Verlag.

Publication type: paper in Q1 peer-reviewed journal.

Originality contribution: I am the first author and I have contributed the greatest portion of the research and writing work, including conceptualization, data collection, and analysis.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) in the areas of “computer science” and “human computer interaction”.

<https://www.scimagojr.com/journalsearch.php?q=19500157063&tip=sid&clean=0>

* * * * *

Core Publication #7 **[CP7]** (included at the end of Chapter 3, Part 2)

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2020),
“Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents”
International Journal of Social Robotics, 13(1), Springer Verlag.
<https://link.springer.com/article/10.1007/s12369-020-00700-6>

Publication type: paper in Q1 peer-reviewed journal.

Originality contribution: I am the first author and I have contributed the greatest portion of the research and writing work.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) in the areas of “computer science” and “human computer interaction”.

* * * * *

Core Publication #8 **[CP8]** (included at the end of Chapter 3, Part 2)

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2021),
“Robotics Aids for Character Training: More Than Just Another Enabling Condition”
International Journal of Social Robotics, 13(1), Springer Verlag.
<https://link.springer.com/article/10.1007/s12369-021-00756-y>

Publication type: paper in Q1 peer-reviewed journal.

Originality contribution: I am the first author and I have contributed the greatest portion of the research and writing work.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) specializing in “computer science” and “human computer interaction”.
<https://www.scimagojr.com/journalsearch.php?q=19500157063&tip=sid&clean=0>

* * * * *

Core Publication #9 **[CP9]** (included at the end of Chapter 3, Part 2)

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (eds.) (2021),
“Virtuous Robotics: Artificial Agents and the Good Life”
special issue of *International Journal of Social Robotics*, 13(1), Springer Verlag.

Publication type: edited collection in Q1 peer-reviewed journal.

Originality contribution: I am the first guest editor and I have contributed the greatest portion of the editorial work.

Disciplinary relevance: according to Scopus-Scimago, *International Journal of Social Robotics* is a Q1 Journal (H-index: 85) specializing in “computer science” and “human computer interaction”.
<https://www.scimagojr.com/journalsearch.php?q=19500157063&tip=sid&clean=0>

* * * * *

Core Publication #10 **[CP10]** (included at the end of Chapter 3, Part 3)

Alnajjar F., Bartneck C., Baxter P., Belpaeme T., Cappuccio M., Di Dio C., Eyszel F., Handke J., Mubin O., Obaid M. (2021),
Chapters 3 and 10 in “Robots in Education: An Introduction to High-Tech Social Agents, Intelligent Tutors, and Curricular Tools”, Routledge, New York.

Publication type: chapters in peer-reviewed monograph.

Original contribution: The co-authors of this monograph are listed alphabetically. I have contributed the greatest part of chapter 10 (*Ethics*) and of chapter 3 (*The Interactive Mind*), in addition to many other parts throughout the whole book.

Disciplinary relevance: the volume is part of a well-established series in AI, computer science, robotics, and HMI. All submitted manuscripts need to pass a stringent double-blind peer-review process.

* * * * *

Core Publication #11 **[CP11]** (included at the end of Chapter 3, Part 3)

Mubin O., Cappuccio M.L., Alnajjar F.S., Ahmad M., and Shahid S. (2020),
“Can a robot invigilator prevent cheating?”
AI and Society, 35, pp. 981–989.
<https://doi.org/10.1007/s00146-020-00954-8>

Publication type: paper in Q1 peer-reviewed journal.

Original contribution: I am the first author with Omar Mubin and I have contributed the greatest portion of the research and writing work, including conceptualization, experimental design, data collection, and analysis.

Disciplinary relevance: according to Scopus-Scimago, *AI and Society* is a Q1 Journal (H-index: 55) in the areas of artificial intelligence and human computer interaction.
<https://www.scimagojr.com/journalsearch.php?q=23616&tip=sid&clean=0>

* * * * *

Support publications

Support Publication #1 [SP1]

Cappuccio M.L. (2009),
“Constructing the Space of Action. From Bio-Robotics to Mirror Neurons”
World Futures. The Journal of General Evolution, 65/2 (February 2009), Taylor & Francis,
London, pp. 126-132.

* * * * *

Support Publication #2 [SP2]

Cappuccio M.L., Ilundain-Agurruza J. (2019),
“Embodied Cognition”
in Dieter Hackfort and Robert Schinke (eds.), *The Routledge International Encyclopedia of
Sport and Exercise Psychology*, Routledge, New York.

* * * * *

Support Publication #3 [SP3]

Cappuccio M.L., Wheeler M. (2012),
“Ground-level Intelligence: Action-Oriented Representation and Dynamics of the
Background”
in Zdravko Radman (ed.), *Knowing without Thinking. Mind, Action, Cognition and the
Phenomenon of the Background*, Palgrave Macmillan, London.

* * * * *

Support Publication #4 [SP4]

Cappuccio M.L., Froese T. (eds.) (2014),
Enactive Cognition at the Edge of Sense-Making. Making Sense of Non-Sense
Palgrave Macmillan, London.

* * * * *

Support Publication #5 [SP5]

Cappuccio M.L. (2017),
“Flow, Choke, Skill: the role of the non-conscious in sport performance”
in Zdravko Radman (ed.), *Before Consciousness. In Search of the Fundamentals of Mind*,
Imprint, 2017.

* * * * *

Support Publication #6 [SP6]

Cappuccio M.L., Gray R., Hill D.M., Mesagno C., Carr T.H. (2019),
“The Many Threats of Self-Consciousness: Embodied Approaches to Choking Under
Pressure in Sensorimotor Skills”
in Massimiliano L. Cappuccio (ed.), *Handbook of Embodied Cognition and Sport
Psychology*, MIT Press, Cambridge MA.

* * * * *

Support Publication #7 [SP7]

Cappuccio M.L. (2023), “Dreyfus is right: knowledge-that limits your skill”, *Synthese* 202, 85.
<https://doi.org/10.1007/s11229-023-04248-6>

* * * * *

Support Publication #8 [SP8]

Sandoval E.B., Sosa R., Cappuccio M.L., Bednarz T (2022)
“Human–robot creative interactions: Exploring creativity in artificial agents using a
storytelling game”
Frontiers Robotics and AI - Section Human-Robot Interaction 24 August 2022
<https://doi.org/10.3389/frobt.2022.695162>

* * * * *

Support Publication #9 [SP9]

Sandoval E.B., Shi J., Sandoval D.C., Li B., Cappuccio M.L., Rosenbaum S. (2021),
“A Prototype of a Robot Memory Game: Exploring the Technical Limitations of Human Robot
Interaction in a Playful Context” In *Companion of the 2021 ACM/IEEE International
Conference on Human-Robot Interaction (HRI '21 Companion)*, March 8–11, 2021, Boulder,
CO, USA.ACM, New York, NY, USA, 5 pages.
<https://doi.org/10.1145/3434074.3447158>

* * * * *

Support Publication #10 [SP10]

Cappuccio M.L., Shepherd, S. (2013),
“Pointing Hand: Joint Attention and Embodied Symbols”, *The Hand, an Organ of the
Mind: What the Manual Tells the Mental*, Zdravko Radman.

* * * * *

Support Publication #11 [SP11]

Shepherd V.S., Cappuccio M.L. (2012),
“Sociality, Attention, and the Mind’s Eye”
in Axel Seemann (ed.), *Joint attention. Perspectives and Developments*, MIT Press,
Cambridge MA, pp. 205-242.

* * * * *

Support Publication #12 [SP12]

Alnajjar F., Renawi A.M., Cappuccio M.L., & Mubin O. (2019),
“A low-cost autonomous attention assessment system for robot intervention with autistic
children”, *Proceedings Of The Tenth Ieee Global Engineering Education Conference
(Educon 2019)*, 9-11 April 2019, American University in Dubai, Dubai, UAE, 787-792.
<https://doi.org/10.1109/EDUCON.2019.8725132>

* * * * *

Support Publication #13 [**SP13**]

Alahbabi M., Almazroei F., Almarzoqi M., Almeheri A., Alkabi M., Nuaimi A.A., Cappuccio M.L., Alnajjar F. (2017),
“Avatar Based Interaction Therapy: A Potential Therapeutic Approach for Children” In X
IEEE International Conference on Mechatronics and Automation (ICMA), 480–496.
<http://dx.doi.org/10.1>

* * * * *

Support Publication #14 [**SP14**]

Cappuccio M.L. (2015), “Inference or Familiarity? The Embodied Roots of Social
Cognition” *Synthesis Philosophica*, 29/2, (April 2015), Croatian Philosophical Society,
Zagreb, pp. 253-272.

* * * * *

Support Publication #15 [**SP15**]

Alotaibi M., Alnajjar F., Cappuccio M.L., Khalid S. (2023),
“Efficacy of Emerging Technologies to Manage Childhood Obesity: A Systematic Review”
Diabetes Metab Syndr Obes. 2022; 15: 1227-1244 <https://doi.org/10.2147/DMSO.S357176>

* * * * *

Support Publication #16 [**SP16**]

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021),
“Mapping Meaning and Purpose in Human-Robot Teams: Anthropomorphic Agents in
Military Operations”
Philosophical Journal of Conflict and Violence, Vol. V (1), Trivent Publishing.
[https://trivent-publishing.eu/img/cms/5-%20Massimiliano%20L-
%20Cappuccio%20et%20al-.pdf](https://trivent-publishing.eu/img/cms/5-%20Massimiliano%20L-%20Cappuccio%20et%20al-.pdf)
DOI: 10.22618/TP.PJCV.20215.1.139005

* * * * *

Support Publication #17 [**SP17**] (redacted)

Cappuccio M.L., Devitt K.S., Cara Wrigley (2020),
Attitudes and Beliefs toward the Introduction of a Logistic Robot in a RAAF Base Warehouse Unpublished report - January-March 2020, 18 pages, 7547 words.
Commissioned by Royal Australian Air Force.

* * * * *

Support Publication #18 [**SP18**]

Cappuccio M.L., Peeters A., McDonald W (2019),
“Sympathy for Dolores. Moral consideration for robots based on virtue and recognition”
Philosophy & Technology, 1-23 (2019), Springer.
DOI: 10.1007/N13347-019-0341-y

* * * * *

Support Publication #19 [**SP19**]

Sandoval E.B., Billingham M., Cappuccio M.L. (2024),
“Robot Natives: Ethical Design and Regulations for Baby-Robot Interaction”
in *Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction*.

0.6 List of Acronyms

4E = Embodied, Enactive, Embedded, Extended Cognition
AI = Artificial Intelligence
ASD = Autistic Spectrum Disorder
ATTA = Acceptance, Trust, Tolerance, & Anthropomorphism
BMH = Broken Mirror Hypothesis
CNN = Convolutional Neural Network
CP = Core Publication
EAA = Embodied Account of Autism
ECT = Embodied Cognition Theory
EDH = Executive Dysfunction Hypothesis
EECR = Embodied/Enactive Approach to Cognitive Robotics
EU = European Union
FEP = Free Energy Principle
GOFAI = Good Old Fashioned Artificial Intelligence
HCI = Human-Computer Interaction
HMI = Human-Machine Interaction
HRI = Human-Robot Interaction
IJSR = International Journal of Social Robotics
iTOM = Impaired Theory of Mind
MTRNN = Multiple Time-Scale Recurrent Neural Network
NSW = New South Wales
PP = Predictive Processing
SOH = Sensory Overload Hypothesis
SP = Support Publication
RAAF = Royal Australian Air Force
ToM = Theory of Mind
UAE = United Arab Emirates
UAEU = United Arab Emirates University
UN = United Nations
UNSW = University of New South Wales
VDST = Values in Defence and Security Technology group
WCC = Weak Central Coherence Theory

0.7 List of Tables and Figures

The 39 tables and pictures included in this dissertation are reproduced with the respective publishers' permission. Most of them are taken from the Core and Support Publications.

[**FIG1**, p. 59] “Representational / Internal-Model Architectures”

[**FIG2**, p. 60] “Non-Representational / Direct Sensorimotor Coupling Architectures”

[**FIG3**, p. 69] “Choking effect as precision-modulation imbalance” (from **CP2**)

[**FIG4**, p. 71] “Active inference and Free Energy minimization” (from Pezzulo, Parr & Friston 2024)

[**FIG5**, p. 76] “Nao robot performing golf putting” (research in progress).

[**FIG6**, p. 88] “Theories of Autism: Comparing the Key Features”

[**FIG7**, p. 97] “Robot Interaction Room setup” (from **CP1**).

[**FIG8**, p. 99] “Real-time motion captures” (from **SP13**).

[**FIG9**, p. 100] “Standard robot versus Interactive Avatars Training” (from **SP13**).

[**FIG10**, p. 101] “Averaged scores of interaction subjects” (from **SP13**).

[**FIG11**, p. 102] “Therapy and interaction room setup” (from **SP12**).

[**FIG12**, p. 104] “Summary Table I: Participants”

[**FIG13**, p. 104] “Summary Table II: Procedures”

[**FIG14**, p. 104] “Summary Table III: Robot Features and Technical Setup”

[**FIG15**, p. 105] “Summary Table IV: Outcomes ”

[**FIG16**, p. 106] “Attention score increments and decrements” (from **CP1**)

[**FIG17**, p. 107] “Normalized interaction scores” (from **CP4**).

[**FIG18**, p. 108] “Real-time plot of attention score” (from **CP4**).

[**FIG19**, p. 111] “Titles of the scenarios” (from **CP4**).

[**FIG20**, p. 112] “Scoring system interface” (from **CP4**).

[**FIG21**, p. 113] “Flowchart of human-robot interaction” (from **CP4**).

[**FIG22**, p. 117] “Attention score progress in multiple patients” (from **CP4**).

[**FIG23**, p. 128] “Interactions between attitudes, norms, and beliefs” (from **SP17**).

[**FIG24**, p. 130] “The Logistic bot” (from **SP17**).

[**FIG25**, p. 135] “Comparing three attitudes towards robots” (from **CP5**).

[**FIG26**, p. 136] “Interactions between attitudes, beliefs, and behaviours” (from **CP5**).

[**FIG27**, p. 138] “The ATTA-bot model: four attitudes” (from **SP17**).

[**FIG28**, p. 140] “Designing a scale for autonomy estrangement” (from **CP5**).

[**FIG29**, p. 143] “Increment 1 Platform unmanned ground vehicle” (from **CP6**).

[**FIG30**, p. 145] “Three approaches to anthropomorphism” (from **CP6**).

[**FIG31**, p. 146-147] “Summary table: Anthropomorphic affordances and ethical behaviour”

[**FIG32**, p. 159-160] “Notable Applications of the Virtuous Robotics Paradigm”

[**FIG33**, p. 166] “Shelly the turtle” (from **CP9**, courtesy of Ku et al 2018a).

[**FIG34**, p. 167] “Interconnection of cloud technologies” (from **SP15**).

[**FIG35**, p. 172] “Setup of testing session” (from **CP11**).

[**FIG36**, p. 173] “Screenshot from the video feed” (from **CP11**).

[**FIG37**, p. 173] “Results: duration of participants’ behaviour” (from **CP11**)

[**FIG38**, p. 179] “Structure of the dissertation”

[**FIG39**, p. 180-181] “Addressing the research questions”

CONTEXTUAL STATEMENT

Chapter 1 – The Embodiment of Sensorimotor Skills. Action-Guiding Attention in Expert Performance

*This chapter introduces the key concepts of HRI through the lens of ECT. The first part argues that devising and delivering effective solutions in social robotics depends on how cognitive functions are embodied by humans and robots. The second part, dedicated to **CP1**, refers to the seminal work of the father of AI, Alan Turing, to explain how ECT necessarily informs Cognitive Robotics. The third part, focussing on **CP2**, discusses the embodiment of attention in skilful action and argues that expert sensorimotor activity is damaged by attentive self-monitoring and presupposes pre-reflective control.*

Chapter 1, Part 0: Preamble

1.0.1 Themes and problems

Chapter 1 lays the theoretical foundation of my research project in Human-Robot Interaction (HRI). It aims to identify the core principles of Embodied Cognition Theory (ECT) and explain how they apply to HRI, cognitive robotics, and social robotics. While theoretical in nature, these principles define the practical approach I have applied in my empirical research, informing the methodology and the questions that motivate my study. These principles emphasize that cognition arises from (and is constituted through) a history of interactions between an agent and its world-environment; that intellectual faculties are scaffolded by and emerge over more fundamental sensorimotor processes; that cognitive processes closely depend on the details of their physical implementation; that cognitive faculties, like attention, are geared toward action and inherently tailored to support interaction with the world; and that cognitive vehicles encompass non-neural material components and relational dynamics.

1.0.2 Research outcomes covered in this chapter

Chapter 1 clarifies the terms, concepts and models that inform my research in cognitive robotics and social robotics. It builds on some of my theoretical works in the domains of AI and HRI. Taken together, these publications outline a set of general methodological principles and a consistent foundational approach to intelligent robotics inspired by ECT. My analysis of embodiment primarily focuses on the intelligent role of sensorimotor processes that enable skilful activities that are adaptive yet pre-reflective (automatized) yet not representational (Schneider & Chein 2003). Skilful performance perfectly illustrates how bodily features contribute to practical knowledge (habitual know-how, expertise) and cognition (understanding, judgment). By adopting the models advanced by *Predictive Processing* (PP), and consistently with the principles of Active Inference under *Free Energy Principle* (FEP), I will expound how the adaptive behaviour of a robot trained with machine learning techniques can be skilled in a pre-reflective sense while suffering from the same limitations of expert human performers, as per ECT. Thus, I will argue that skilful performance in both expert humans and intelligent robots must gradually evolve from pre-reflective, embodied forms of attention that are anticipatory in nature, enactive, and outcome driven (Robertson & Kirchhoff 2024), as opposed to merely spectatorial and analytic.

1.0.3 Author's publications used in this chapter

In Chapter 1, I will review and discuss the contents of 11 of my publications (2 “Core Publications” and 9 “Support Publications.”)

Core Publications

[CP1] Cappuccio M.L. (2015)

[CP2] Cappuccio M.L., Kirchhoff M.D., Alnajjar F.S., Tani J. (2019)

Support Publications

[SP1] Cappuccio M.L. (2009)

[SP2] Cappuccio M.L., Ilundain-Agurruza J. (2019)

[SP3] Cappuccio M.L., Wheeler M. (2012)

[SP4] Cappuccio M.L., Froese T. (eds.) (2014)

[SP5] Cappuccio M.L. (2017)

[SP6] Cappuccio M.L., Gray R., Hill D.M., Mesagno C., Carr T.H. (2019)

[SP7] Cappuccio M.L. (2023)

[SP8] Sandoval E.B., Sosa R., Cappuccio M.L., Bednarz T (2022)

[SP9] Sandoval E.B., Shi J., Sandoval D.C., Li B., Cappuccio M.L., Rosenbaum S. (2021)

Chapter 1, Part 1: The Interdisciplinary Background

1.1.1 The Field of Interaction Studies

In this section, I will present the methodological and foundational background of my publications, contextualizing them within the domain of Interaction Studies. From a taxonomical standpoint, HRI represents a branch of Human-Computer Interaction (HCI), a broad area of research that falls under the jurisdiction of Human-Machine Interaction (HMI), which –in turn– represents a sub-field of Human-Technology Interaction (HTI) and, through it, of Interaction Studies at large. HMI investigates how intelligent technological artifacts and automated systems can meaningfully coordinate with humans in various settings to make these interactions smoother and more efficient (Boy 2017)¹.

HRI accounts for the unique needs arising from the complex relationships between humans and autonomous robots (Rodogno 2016). These needs were not considered when automated systems were confined to the factory environment but became more perspicuous when mobile interactive robots became integrated into everyday life, including informal settings and unstructured environments like home, office, classroom, playing field, etc., (Chen 2016). Notably, *autonomous* systems represent a more sophisticated and specific version of *automated* systems, as *autonomy* derives from, but also exceeds, the quality of *automation*: instead of mechanically reproducing pre-established patterns of activity, autonomous systems intelligently adjust to and anticipate the circumstances, actively modifying their behaviour accordingly (Abbass, Scholz, Reid 2018; Asada 2020). They navigate the online environment through iterative cycles of perception-decision-action operations accompanied by constant predictions of the consequences of their own actions. Compared to other automated technologies, autonomous technologies display a higher degree of self-determination, i.e. the capability to operate with little or no human supervision (Formosa 2021). This is accounted for by HRI's protocols, which offer the coordinates necessary to construe robots as co-workers (co-bots) and companions (social robots), while it is not necessarily a concern for HCI and HMI, which encounter automated systems as passive tools.

¹ HMI research draws from disciplines such as industrial design, psychology, cognitive science, computer science, engineering, and social sciences to understand and improve the interaction between humans and various automated systems in all its desirable aspects: communicative, collaborative, usable, safe, secure, natural, and ethical interaction.

HRI focuses on embodied dimensions of interaction that did not represent a primary concern for HCI researchers (Asada 2021a). While HCI mainly focuses on designing and improving interactions between humans and digital interfaces (software or websites), HRI deals with the complexities arising from the embodiment of robots: *physical presence* (unlike traditional computers or screens, robots occupy physical space and interact with the environment, demanding HRI to address issues of spatial awareness, physical laws, and body language that aren't typical concerns in HCI, see He et al. 2024; Reichenbach, Bartneck & Carpenter 2006); *mobility and behaviour* (unlike static or passive interfaces in HCI, robots are designed to complement human actions and intentions, see Cappuccio 2009 [SP1]; Damiano, Dumouchel & Lehmann 2015b; Daza et al. 2021); *safety and trust* (with physical robots, reliability becomes a crucial element, as they operate in close proximity to humans, see Cappuccio et al. 2024 [CP5] Gaudiello et al. 2016; Ullman & Malle 2018); *time-sensitive interactions* (robots act in real time to adjust to the human pace, thus HRI emphasizes timing, responsiveness, and velocity, see Ku et al 2018b; Milne et al. 2018; Wirkuttis & Tani 2021); *emotional engagement* (robots may engage with humans on a social or emotional level, creating meaningful pseudo-relationships with humans, see Dautenhahn 2007; Cauchard et al. 2016; Mazzei et al. 2014). To introduce my project, I need to explain how HRI intersects the fields of *Affective Computing* and *Social Robotics*.

1.1.2 Affective Computing

The need to produce a fulfilling emotional engagement with machines explains why HRI importantly overlaps with Affective Computing, the discipline that designs systems and devices capable to recognize, interpret, and simulate human emotions (Asada 2014; Damiano, Dumouchel & Lehmann 2015b, Lombardi et al. 2024). The goal of Affective Computing is to create technology that interacts intuitively with humans by adjusting to their emotions. Affective Computing is defined by three key components: first, *Emotion Recognition* (affective computing uses sensors, algorithms, and machine learning techniques to detect human emotions. This can involve analysing facial expressions, body language, tone of voice, text, or physiological signals –such as heart rate or skin conductance– to determine a person's emotional state, see Nezami 2020); second, *Emotion Synthesis* (this involves creating systems that can simulate or express emotions in ways humans can recognize. For example, a virtual assistant might express empathy in its voice or display a concerned face on screen when responding to a user's distress, see Lombardi et al. 2023); third, *Personalization and Adaptation* (systems using affective computing can adapt their responses based on the user's mood or personality – which is especially valuable in education, customer service, or healthcare, where adjusting to a person's emotional state

can improve outcomes and satisfaction, see Milne 2018; Alnajjar et al 2020 [CP4]). By integrating these capabilities, Affective Computing aspires to enhance HRI and make technology human-centered, especially in sectors where understanding and expressing emotions are crucial (healthcare, education, customer service, and entertainment).

1.1.3 Social Robotics

Social Robotics is the field of robotics that focuses on designing autonomous agents that can engage with humans in *socially meaningful ways* (Nørskov 2017a; Fortunati & Edwards 2024; Palinko et al 2025). These robots' capability to interact and communicate ultimately aims to *form relationships* (or, arguably, pseudo-relationships) with people by recognizing, interpreting, and responding to human social cues, emotions, and behaviours. Social Robotics arises from the union of HRI and Affective Computing, as they both aim to create robots that function effectively in human-centered environments, such as the classroom, the hospital, the office (Breazeal, Dautenhahn & Kanda 2016). Key aspects of social robots are: (i) *Human-Centered Design* (social robots are designed to fit naturally into human environments like companions, assistants, or service providers, see Breazeal 2002; Marchesi & Wykowska 2024); (ii) *Social Interaction and Communication* (social robots need to produce and interpret social cues such as eye contact, facial expressions, body language, and tone of voice – involving Affective Computing, see Onumi 2014; Nezami et al 2020; Lombardi et al 2023;); (iii) *Autonomy and Decision-Making* (social robots may need autonomy to adapt their responses to the context, past interactions, or individual user preferences, see Pfeifer & Gómez 2005; Pfeifer, Iida & Gómez 2006).

Like Affective Computing, Social Robotics is a growing area of interest for industries where social engagement can enhance user experience and quality of life. But Social Robotics presents designers and makers with specific challenges (Nørskov 2017b). First, achieving natural, human-like interaction is difficult because it presupposes significant advances in robotics hardware (for movement, expression) and software (especially for social cognition – including, importantly, behaviour reading and intention understanding), see Neau et al (2022). Additionally, creating robots that can adapt to rich human contexts (with robots often operate in roles that involve trust and social intimacy), raises ethical questions about privacy, dependency, and the potential for deception (Richards, Vythilingam & Formosa 2023).

Social Robotics aims to create robots that function as "social agents," interacting with humans not only *instrumentally* (interaction for an external goal) but also *relationally* (interaction aiming to improve the quality and the significance of the relationship itself), see

Seibt, Nørskov & Hakli (2014). Both disciplines aim to optimize the interactions between the users and the machines, however Social Robotics prioritizes emotional and psychological factors, while HRI focuses on improve performance effectiveness and safety, with or without social elements².

1.1.4 Cognitive Robotics

Cognitive Robotics focusses on endowing Agents with cognitive abilities (such as perception, learning, and problem-solving) similar to the ones displayed by natural cognitive agents (such as individual animals or swarms of animals), see Nakamura et al (1999), Clark & Grush (1999), Cangelosi & Asada (2022). Its success depends on figuring out how natural cognition works to effectively replicate its functional architecture and its underlying processes. Cognitive Robotics incorporates ecological insights from cognitive psychology, biology and neuroscience to design robots that reason and learn like humans, enabling robots to operate in dynamic, unstructured environments created by humans for humans, i.e. environments in which human interaction is pervasive, complex, unprincipled, and thus hardly predictable by machines that act based on fixed sets of explicit instructions and finite, formal representations of the world (Sandini, Sciutti & Vernon 2021).

For a long time, cognitive robots were controlled only by pre-programmed instructions expressed in logico-symbolic terms, which confined the robots' activity within formal reasoning tasks (playing chess) and artificial environments characterized by a limited number of degree of freedom and an artificially imposed low-variability (i.e., protected, closed factory environments where interaction is constrained by railings and mechanical barriers). This is often referred to as the "Good Old Fashion Approach to AI" (GOF AI). GOF AI robots may display scarcely adaptive behaviours when operating outside of the closedness and formality of ad hoc artificial environments: in open-ended, unprincipled contexts, their intelligence is limited because it is governed by "if...then" rules or explicit ontological descriptions that can only encode simple toy-model representations of the world (Verschure, Kröse & Pfeifer 1992).

However, in the age of machine learning and deep-learning, robots are trained with large amounts of data so that they can learn to autonomously interpret, navigate, manipulate, and adapt to real-world environments that were not specifically designed for them. They can do

² Methodologically, social robotics borrows primarily from psychology and affective computing, while HRI leans on engineering, usability, and cognitive psychology to address diverse interaction needs. Both HRI and Social Robotics build on cognitive science and Cognitive Robotics.

this because they are not designed anymore to rely on pre-defined sets of rules, formal ontologies and internal representations of the world (Tani & Nolfi 1999). These AIs rely on advanced connectionist systems, such as convolutional-neural-networks (CNNs), specialized in identifying and recognizing objects by interpreting visual cues, and multiple-time-scale-recurrent-neural-networks (MTRNNs), specialized in creating hierarchically organized goal-oriented actions emerging from interwoven constituent movements and predicting their most likely effects (Tani 2022). In a manner that appears much more autonomous and adaptive when compared to the traditional logico-symbolic systems, connectionist systems seem better equipped to process natural language. Also, like animals and humans, they are capable to learn from past experiences, plan complex tasks based on the desired goal, recognize the intended goal of an observed action, and predict an action's intended effects (Fitzpatrick et al 2001; Erlhagen et al 2006). It is especially through machine learning and deep learning techniques that Cognitive Robotics is going to inform Social Robotics and HRI, advancing both of them. My work is inspired by the particular approach to Cognitive Robotics informed by Embodied Cognition Theory: the Embodied and Enactive Approach to Cognitive Robotics (EECR), which makes extensive use of dynamical systems based on machine-learning techniques and MTRNNs. I shall deepen the principles of ECT and EECR in Part 2.

Chapter 1, Part 2: A Theoretical Framework for Embodied Cognitive Robotics

This section is centred on Core Publication #1 [CP1]

1.2.1 Embodiment as a general theory of cognition

Embodied Cognition Theory (ECT) defends the idea that mental functions, including high-level “thought” processes (rational decision-making, evaluative judgment, cognitive appraisal, linguistic communication, and symbolic-formal reasoning) are influenced by the details of the body’s physical instantiation (Varela, Thompson & Rosch 2018; Gallagher 2011; Clark 2016; Cappuccio & Ilundain-Agurruza 2019 [SP2]). This idea applies to both natural-biological bodies and artificial-synthetic bodies (Chrisley 2003; Pfeiffer & Lida 2004; Hoffman 2012). The embodiment of a robot is understood primarily as *mechanical structure*, secondarily, as *apparent animacy and presence* and, finally, as *capability to physically interact with the environment*, including other agents (Asada 2021a). But ECT does neither simply emphasize the functional role of the robot’s actuators and sensors in HRI (as an interface, instrument, or medium) nor does it simply bear out that social and cognitive tasks rely on the external outlook of a robot. Everybody agrees that, even in a naïve sense, embodiment determines how the robot engages, communicates with, and is perceived by users and bystanders, for example how children interact with a robot in story-telling games based on creativity and narration (Sandoval et al 2021 [SP9], 2022 [SP8]). In this basic sense, a robot's embodiment reduces to features like body type, facial expressivity and body language, gestures, eyes and gaze, and proxemics, which all contribute to how effectively the robot can fulfill its intended roles and engage with humans.

1.2.1.1 Body type: (i) *Morphology and Structure* – A robot's shape, size, and structure should align with its concept and purpose. Humanoid robots might be more suitable for social interaction tasks, while other forms, quadrupeds or wheeled robots, are better for tasks requiring mobility across challenging terrain (Pfeifer 2000); (ii) *Anthropomorphism* – Designing robots with human-like features can improve social interaction by making them more relatable and engaging (for example, humanoid robots with a torso and arms are better suited for tools and spaces, see Vernon & Sandini 2024; Nehaniv et al 2013); (iii) *Physicality and Materiality* – The materials (hard vs. soft) and weight distribution of a robot impact its physical interactions. A soft-bodied robot is often more approachable and safer, especially in caregiving or educational contexts, as it appears more accessible (Nakajima 2013).

1.2.1.2 Expressivity: (i) *Facial Expressions* – In humanoid robots, a face that can express emotions through changes in “features” (such as a moveable mouth or eyebrows) can improve interaction by conveying emotions, intentions, and states (Lombardi et al 2023); (ii) *Body Language* – Posture and stance can also convey moods, attitudes, and attention (relaxed posture can indicate friendliness, while a more rigid one may signal formality or readiness, Cauchard et al 2016); (iii) *Colour and Lights* – Some robots use lights or colour changes to convey emotional states or inform the user about the robot status (Palinko et al 2025).

1.2.1.3 Gesture: (i) *Hand and Arm Movements* – Gestures convey information, enhance communication, and add expressiveness (for example, pointing can direct attention, while waving can greet or dismiss, see Salem et al 2011); (ii) *Iconic vs. Deictic Gestures* – Deictic gestures like pointing help in directing attention, while iconic gestures mimic the shape or action related to a topic, enhancing understanding and engagement (Cappuccio & Shepherd 2013 [SP10]; Franco 2013); (iii) *Expressive vs. Functional Gestures* – Some gestures serve functional purposes (e.g., pressing a button) and facilitate the execution of a cognitive task (Chu & Kita 2011) (e.g., counting, mental rotations), while others are purely expressive (e.g., waving hello). Including both types in a robot’s repertoire makes the robot more responsive in social contexts and facilitates joint tasks.

1.2.1.4 Eyes and gaze: (i) *Eye Contact* – Eye contact can be powerful for establishing rapport, soliciting self-awareness (Mubin et al 2011 [CP11]), and creating a sense of connection. Robots making eye contact are perceived as more “present” (Shepherd & Cappuccio 2012 [SP11], Kompatsiari 2021); (ii) *Gaze Following* – Robots that follow human gaze can improve collaborative activities, such as shared tasks or instruction-following (Lombardi et al 2022). If a robot looks in the direction someone is pointing, it signals social awareness by orienting attention, a key phenomenon that is known as “joint attention” (Seemann 2012; Willemse & Wykowska 2018); (iii) *Blinking and Eye Movements* – Small details like blinking and subtle eye movements can make robots appear more lifelike and socially engaging, reducing the “uncanny valley” effect by making their behaviour seem less mechanical (Onuki 2014).

1.2.1.5 Proxemics: (i) *Personal Space Awareness* – Proxemic awareness, i.e. understanding and respecting comfortable social distances, can improve a robot’s acceptability in social contexts, because robots that respect personal space are often perceived as more considerate (Daza et al 2021); (ii) *Approach Behaviours* – The way a robot approaches people, for example moving slowly and stopping at a respectful distance,

can make interactions feel more natural, as the ideal proxemic distance may vary by culture and context (Ball et al 2014); (iii) *Adaptability to Contexts* – Proxemics may change based on the setting, such as sitting closer in a medical context but keeping more distance in a retail space, considering that robots that adapt to proxemic expectations are better accepted and integrated into varied settings (De Graaf & Ben Allouch 2013).

1.2.1.6 Tactility and haptics: (i) *Touch Responsiveness* – For robots designed to work closely with humans, the ability to respond to touch (haptic feedback or tactile sensors) adds a new dimension to embodiment (robots “recoiling” gently if touched); (ii) *Surface Texture and Material* – Soft, skin-like materials may make robots appear more approachable and safer, while hard, metallic surfaces could convey durability or reliability, depending on the role of the robot (Pfeifer 2000; Pfeifer, Iida, Gómez 2006).

1.2.1.7 Voice and speech: a robot’s voice and speech characteristics are essential for interactions, influencing how users perceive its personality, trustworthiness, and competence: a friendly, soft tone is more appropriate in social settings, while a more authoritative tone may suit informational or instructional tasks (Cangelosi et al 2010).

These features contribute to create a robot that is more approachable and capable to carry out its physical duties with its material body. While uncontroversial, this is hardly enough to characterize the role played by embodiment in the development of intelligence, which is central for ECT.

1.2.2 The Three Basic Tenets of ECT

The design features described in section 1.2.1 contribute to determine how naturally a robot interacts with people. But endorsing ECT implies something more specific. A truly embodied robot is *more than a computer attached to some peripherals*: its body is not only an input device or an actuator, but a defining, constitutive component of what the cognitive agent is and what it does (Pfeifer 2006; Vernon, Metta & Sandini 2010).

Bodily features contribute to cognitive information processing even if they do not partake in the core computational activity of the robot (which is carried out by its electronic components, like its micro-processors and memory banks: for example, the proponents of soft robotics (Pfeifer, Lungarella & Iida 2012; Pfeifer, Marques & Iida 2013). Morphological computation maintain that not only the sensors and the actuators, but even the material constitution and the geometrical configuration of the robot’s body (e.g., the flexibility of

artificial tendons, the joints' degrees of freedom) actively contribute to simplify the calculus of the movements necessary to walk or perform an action (Pfeifer & Gómez 2009; Pfeifer & Scheier 1997; Pfeifer, Iida & Gómez 2006). This notion is leveraged by the cognitive scientists of enactivist orientation (see section 1.2.5) and the computer scientists working within the paradigm of FEP under *Active Inference*, a particular approach to PP that considers bodily movement as a legitimate part of the process of prediction error minimization that every agent continuously performs to adjust to their ever-changing environment (see section 1.3.5). With its inherent responsiveness to environmental solicitations, the body realizes a living model of the world. So, what is really key of ECT is that it does not treat embodiment as a mere realiser or accidental implementer of cognitive functions. The body is not a contingent and instrumental aspect concerning only the terminal edges of the agent's intelligent processes (input and output). Rather, ECT understands embodiment as a constitutive, necessary, and defining component of those very processes (Gallagher 2012; Ziemke 2022).

In cognitive science's broader domain, the key idea of ECT is that, far from being reducible to cerebral or neuronal activity, cognitive processes are constituted by dynamical and interactive systems that are spread over the body, brain, and environment (Clark 1998; Pfeifer & Gómez, 2009), thus encompassing physical processes and causal relations situated beyond the central nervous system. I proposed a systematic account of ECT in Cappuccio and Ilundain-Agurruza (2019) [SP-1B], where the reach of embodied cognitive science is analytically examined and illustrated with reference to sensorimotor abilities used in sport activity. Sport is one of the domains of human expertise in which sensorimotor skills most clearly demonstrate the body's capability to scaffold and shape higher-forms of intelligence such as strategic evaluation, decision-making, and anticipation (Cappuccio 2015, 2017 [SP5]). This perspective encourages a broadened understanding of the mind as a situated, interactive system operating in the world skilfully, rather than an isolated computing/information processing system performing calculi and inferences (Sciutti & Sandini 2021). ECT challenges traditional views of cognition as a purely formal and disembodied process (a-modal information exchange, computational or algorithmic processing abstracted from its contingent material bases, purely brain-bound mechanistic functions).

ECT essentially revolves around three basic tenets (mentioned most frequently) and two radical tenets (that are often underplayed or neglected). Let's review the basic tenets, first:

1.2.2.1 Cognition reflects the dynamic interplay of body, brain, and mind – ECT posits that cognitive processes are not isolated within the brain but arise from bodies and their activities (Clark 1998). Our bodies, including our sensory and motor systems, play a crucial role in shaping cognitive functions, both in ontogeny (development of the individuals) and in phylogeny (evolution of the species).

1.2.2.2 Sensorimotor experiences scaffold abstract and symbolic cognitive processes – Our perceptual and motor experiences provide us with scaffolds for the development of abstract and symbolic ways of thinking, which however are not part of our basic cognitive endowment (Ghiglino et al 2023). Language, including abstract concepts, is grounded in physical experiences and sensory-motor activity: concepts of spatial understanding (like "up" and "down") are rooted in bodily experiences and spatial orientation, as demonstrated by many metaphors ("grasping an idea," "falling in love," or "upbeat mood", see Lakoff 1999).

1.2.2.3 Intelligence is attuned to the environment in which it has evolved – The environment is not just a backdrop for cognitive processes but their essential component. Intelligence is a fluid, adaptive, and context-dependent dynamic process, unlike a mechanical system of inferences governed by rules for the manipulation of semantic representations (Pfeifer & Scheier 2001; Pfeifer 1995). Perception and action are not only instrumental to cognition but are part of it, like decision-making and problem-solving.

1.2.3 The two radical tenets of ECT

Embodiment cannot be fully understood without adequately appreciating two additional tenets, which are considered radical by many cognitivists: anti-representationalism and the anti-neuro-reductionism.

1.2.3.1 First Radical Tenet: Basic cognition is non-representational – ECT posits that the fundamental cognitive processes (the processes we share with other animal species and even simple organisms), do not rely on internal representations or symbolic/semantic models of the external world (Brooks 1991). Basic cognition emerges from *direct interactions* with the environment, facilitated by the body's sensorimotor and organic dynamics (Di Paolo 2003; Pfeifer, Lungarella & Iida 2007; Vernon 2010).³ Even without the mediation of

³ Of course, stored heuristics, contentful rules and instructions, logico-symbolic descriptions and propositional models are an important component of our most advanced and intellectual cognitive processes. But they are hardly

representations, our mind can effectively, adaptively engage with the world through perception and action. Hubert Dreyfus proposed this view through his theories of “absorbed” and “background” coping (Dreyfus 2005, 2007) to deepen how skilled activity is based on continuous pre-reflective engagement rather than internal models of the external world, an idea reprised and expanded in Cappuccio and Wheeler (2012) [SP3] and Cappuccio (2023) [SP7].

The non-representational account of cognition promoted by ECT can be summarized in three points: (i) *History of Interactions Over a Repository of Contents* – Basic cognitive processes are unmediated, as they directly emerge from bodily perception and action loops in real time (Clark 1998; Cangelosi 2011): catching a ball doesn’t involve an internal model of the ball’s trajectory but a process of ongoing engagement through the online response to the ball’s motion (Runswick et al 2019). (ii) *Sensorimotor Patterns before Contentful Beliefs* – knowledge at its most fundamental level relies on sensorimotor patterns and perceptual know-how which allow organisms to detect saliency and respond to environmental changes, without symbolic knowledge of facts and rules (Dreyfus 2005; Cappuccio 2023 [SP7]). (iii) *Environment, not the Brain, as the Main Informational Resource* – The environment acts as an active component of cognition, providing the organism with online cues that it uses directly rather than forming an internal model. This approach aligns with roboticist Rodney Brooks’ “world-as-its-own-best-model” hypothesis: the world out there functions as a real-time guide to itself, bypassing the need for internal representations (Brooks 1999). Instead of manipulating semantically rich models and contentful representations of the world, organisms effectively *use* the environment itself as a cognitive aid and a source of information (Friston 2018)⁴.

the most basic or fundamental component, which is why the project to build minds in a bottom-up fashion cannot start from them (Hutto & Myin 2013).

⁴ The anti-representationalism of ECT has important implications in cognitive science. First, this approach emphasizes the mutual dependency of cognition and agency (Ziemke 2023): exploring the world and extracting relevant information are reciprocal elements of the task at hand (Bruineberg & Stone 2024). Basic cognitive processes are world-directed rather than mediated, enacted rather than theorized, performative rather than contemplative, governed by pragmatic principles (conditions of success, improvement) rather than epistemic, truth-fulfilling principles (conditions of accuracy, correspondence). Most important, cognition is fine-tuned to produce and control skilful activity, not abstract intelligence. Secondly, skilful activity relies on the embodied know-how developed through exercise and practice (expertise is not reducible to rules-of-thumb and instructions, it must be learned by doing, concretely experiencing the effort necessary to successfully complete goal-oriented actions). Thirdly, skilful activity resembles a state of flow, i.e. continuous undisturbed sensorimotor engagement (Swann 2012; Cappuccio 2017 [SP5]): embodied know-how derives from habitual control, not formal rules or verbal instructions, which often interfere with the spontaneous execution of familiar sensorimotor patterns that support embodied know-how (Cappuccio 2017 [SP5]; Cappuccio 2023 [SP7]; Cappuccio et al 2019a [SP6]; Cappuccio et al 2019b [CP2]).

ECT's anti-representationalism has important implications for the field of HRI, as robots based on a non-representational architecture differ significantly from the robots designed according to classical representationist principles. The two tables below summarize some of the main features of both for comparative purposes.

Representational / Internal-Model Architectures				
System / Project	Authors & Year	Task & Environment	Core Mechanism	Canonical references
SHKEY (SRI)	N.J. Nilsson (ed.), 1984; R.E. Fikes & N.J. Nilsson, 1971	Indoor mobile manipulation with navigation & simple rearrangement tasks	Symbolic world model + STRIPS planning; perception updates internal map; plan-act loop	Nilsson, <i>Shakey the Robot</i> (Tech Note 323); Fikes & Nilsson, "STRIPS" (AI, 1971).
EKF-SLAM (foundational)	R. Smith & P. Cheeseman, 1986	Mapping & localization in unknown environments	Probabilistic state-space model; Extended Kalman Filter jointly estimates pose + landmark map	Smith & Cheeseman, "On the Representation and Estimation of Spatial Uncertainty" (1986).
D* (Dynamic A*)	A. Stentz, 1994/1995	Real-time replanning in partially known terrain	Incremental graph search that repairs optimal paths as costs change	Stentz, <i>The D* Algorithm*</i> (1994); <i>Focussed D*</i> (IJCAI 1995).
FastSLAM	M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit, 2002; book 2007	Large-scale SLAM with many landmarks	Rao-Blackwellized particle filter: particles for robot trajectory + EKF per landmark	Montemerlo et al., AAAI-02; Montemerlo & Thrun, <i>FastSLAM</i> (Springer, 2007).
Potential Fields / Elastic Bands	O. Khatib, 1986; S. Quinlan & O. Khatib, 1993	Reactive collision avoidance & local path refinement	Continuous configuration-space cost fields; gradients guide motion while respecting constraints	Khatib (IJRR, 1986); Quinlan & Khatib (ICRA, 1993). (SRI , Stanford AI Laboratory)
Task & Motion Planning (TAMP)	L.P. Kaelbling & T. Lozano-Pérez, 2011→	Long-horizon mobile manipulation	Hierarchical integration of symbolic task plans with geometric motion plans ("HPN")	Kaelbling & Lozano-Pérez, ICRA-2011; follow-ups/tutorials. (people.csail.mit.edu)
Stanley (DARPA Grand Challenge winner)	S. Thrun et al., 2006	Autonomous off-road driving (128 mi, Mojave)	Probabilistic perception & mapping fused with model-based planning	Thrun et al., <i>JFR</i> 2006. (sci.brooklyn.cuny.edu)

Non-representational / Direct Sensorimotor Coupling Architectures				
System / Project	Authors & Year	Task & Environment	Core Mechanism	Canonical references
Behaviour-based “Creatures” / Subsumption	R.A. Brooks, 1991 (and layered control 1986)	Robust mobile navigation in real buildings	Parallel, layered subsumption behaviours; “use the world as its own model” (minimal internal content)	Brooks, “Intelligence without representation” (AIJ, 1991).
Braitenberg Vehicles	V. Braitenberg, 1984/1986	Minimal agents that “look intelligent”	Direct sensor-to-motor couplings; emergent behaviour without internal content	<i>Vehicles: Experiments in Synthetic Psychology</i> (MIT Press).
Passive Dynamic Walkers	T. McGeer, 1990	Biped walking down slight slope	Exploits natural body–environment dynamics; no actuation/control needed	McGeer, “Passive Dynamic Walking” (IJRR, 1990)
Raibert Hoppers/Running Machines	M.H. Raibert, 1986	1-leg, biped & quadruped dynamic balance	Simple feedback laws that stabilize limit cycles; minimal world modelling	Raibert, <i>Legged Robots That Balance</i> (MIT Press, 1986/2000).
Evolutionary Robotics (ER)	S. Nolfi & D. Floreano, 2000	Phototaxis, navigation, collective tasks	Controllers evolved in/for embodiment; behaviour emerges from agent–environment loop	Nolfi & Floreano, <i>Evolutionary Robotics</i> (MIT Press, 2000).
Kismet (social coupling)	C. Breazeal, ~2001–2003	Face-to-face affective HRI, turn-taking	Multimodal, affect-driven behaviour selection & attention; interaction-first design	Breazeal, “Emotion and Sociable Humanoid Robots” (IJHCS, 2003); <i>Designing Sociable Robots</i> (MIT Press).
iCub (developmental enactive platform)	G. Metta, G. Sandini, D. Vernon, 2007–2010	Humanoid exploratory learning & social interaction	Embodied developmental learning; skills built via sensorimotor exploration and social cues	Metta et al., <i>Neural Networks</i> (2010); Sandini–Metta–Vernon (AI@50 / LNCS chapter).

1.2.3.2 Second Radical Tenet: Cognitive processes encompass non-organic/non-cerebral vehicles – Cognition does not merely *arise* within specific brain structures (supervening or emerging from the interactions of myriads of simpler components); rather, cognition is *realized* and *constituted* by the brain harmonic activity as a whole (Thompson & Varela 2001) and by the entire body the brain is connected with. The body is not a mere servant of the brain but a co-decision-maker. The body's shape, motor functions, and sensory capabilities are not just contributing factors but essential substrates for cognitive processes. Also, the musculature, the internal organs (the digestive and the endocrine systems), the bone structure, and the autonomic responses contribute to cognition (Pfeifer 2020): they *are part of the cognitive process itself*. The body as a whole, with its internal and external features, including the extra-neuronal body and even extra-bodily extensions (tools,

etc.), realize cognition by implementing information processing functions. The body is (i) a *constraint for cognitive functions*; (ii) a *distributor of cognitive processes*; (iii) a *regulator of cognitive activity* (Wilson & Golonka 2013). Also, the body arguably is (iv) an *enabler and a scaffold for emerging cognitive faculties*.

1.2.4 Broadening the scope of ECT: The 4E cognition framework

ECT emphasizes the role of the body's physical structure as a vehicle of cognition. However, not only our body's characteristics (e.g., morphological structure, biochemistry, material constitution, kinematic organization) but also our bodily actions, our bodily situation within the environment, and our bodily relationship with our tools contribute to shape our perception, judgment, memory, etc., possibly altering our cognitive performances (attention span, problem-solving, cognitive capacity, working memory usage). The label of "4E cognition" was proposed to highlight that cognition is not confined to the brain but encompasses body, action, environment, and tools and other vehicles outside the individual: cognition is thus *embodied, enactive, embedded, and extended*, with the last three characterizations specifying the implications of the first one (Newen, De Bruin & Gallagher 2018). *Emotional Cognition* is often cited as a fifth "E" component (Damiano, Luisa & Cañamero 2010).

Enactive cognition posits that cognition arises from active goal-directed, sensorimotor engagement with the world and emphasizes the history of body-world interactions and their dynamical coupling (Varela, Thompson & Rosch 1991; Ziemke 2003; Vernon 2010). Intelligence develops through doing, understanding through exploration, knowledge through practice. Cognition is a process of bringing forth ("enacting") the world through interaction, as detailed in Cappuccio & Froese (2014) [SP4]. *Embedded cognition* proposes that cognitive processes are situated in a specific environment and context, which provides resources and structures that shape and guide intelligence and knowledge (Clark 1998; Pfeifer & Gómez 2009). *Extended cognition* argues that cognitive processes can be distributed across the brain, body, and external tools or devices: artifacts and external symbols are not merely aids but can, when used, become integral parts of cognitive processes (Clark & Chalmers 1998; Clark 2008). Finally, *Emotional cognition* argues that feelings and passions are not just epiphenomena of mind but processes supporting the initiation and regulation of cognitive functions⁵.

⁵ ECT became popular in the 1990s, yet the principles of embodied cognition had already shaped AI and robotics since their inception. 4E cognition has implications in various areas, such as education (learning enhanced by movement and sensory engagement), therapy (neuro-rehabilitation based on the controlled execution of movements), and AI (machines benefit from robotic bodies to develop practical know-how).

1.2.5 Turing's legacy

How did ECT impact autonomous technologies, including robotics and AI? To grasp the body's role in intelligence we must revisit Alan Turing, often called the father of computer science and AI: embodiment is not an accidental trait of intelligence but a precondition for its development and the material medium through which it arises. ECT owes much to Turing for advancing this seminal idea⁶.

In his 1950 paper *Computing Machinery and Intelligence*, a cornerstone of AI theory, Turing introduces the “imitation game”, a question–answer test with a human interlocutor meant to assess whether a machine has achieved human-like intelligence. But how could a machine reach such intelligence? His answer was a “child-machine” (Turing 1950, p. 456) that learns from attempts at manipulation and circumspection, monitoring results and broadening its repertoire of actions and goals. Like today's developmental roboticists (Asada 2013, 2015; Vernon, von Hofsten & Fadiga 2014; Cangelosi & Schlesinger 2015), Turing's proposal captures the role of embodiment in cognitive development: learning through physical exploration of the environment, whether supervised, unsupervised, or via reinforcement.

This idea now inspires iCub, a tetrapodic anthropomorphic machine equipped with infant-like perceptual, motoric, and interactional features that scaffold the development of increasingly complex action schemata (Sandini, Metta & Vernon 2007; Metta et al. 2008; Sciutti & Sandini 2021). Constant sensorimotor refinement of iCub's manual expertise—supported by grasping and force control systems—enables it to expand skills of hand-eye coordination, tool use, and object manipulation. These manipulative and explorative attempts improve its recognition of actions, affordances, objects, and faces, while updating routines to comprehend causes and predict effects of interactions, also in conjunction with language processing (Cangelosi et al. 2010), responsive gazing (Lombardi et al. 2014), and basic joint attention.

⁶ But isn't Turing traditionally associated with cognitivism and GOFAL, due to his early interest in machines that perform sequential symbol manipulation in accord with formal logic? By showing that all mathematical functions can be represented through the rule-based behaviour of a discrete-state machine, Turing (1936) has been wrongly cast as a precursor of the computational theory of mind, the cognitivist view that the mind functions like a logico-symbolic device. Yet Turing never endorsed this claim; his work also offers grounds to resist computationalism and classical cognitivism as general theories of cognition. In my work [CP1], I have argued that Turing's provided a foundation to both the cognitivist (logico-symbolic) approach to cognition exploring the connectionist (parallel, distributed) approach. Additionally, he anticipated the third, embodied/enactive and situated approach, intuiting the necessary evolutionary and developmental preconditions of intelligence.

1.2.6 Outcomes, limitations, and future research

Turing was, if not a hero of ECT, at least an important precursor of its key principles: he anticipated that general AI could have hardly been programmed through rules and instructions. Rather, it had to be trained through embodied engagement, like a child's intelligence. The only effective method to synthesize human-like intelligence consists in providing a "child-like" artificial cognitive agent with a body capable to explore the world, manipulate the surroundings and interact with other agents.

The conclusion I defended in **CP1** and other works is that an artificial agent is either embodied from the beginning or it will never be truly intelligent: in accord with Turing's intuition, real intelligence cannot be programmed in a computer but needs to evolve through bodily exploration by independently seeking the information it needs to make sense of its world. The rise of developmental robotics made this notion familiar to us (Lungarella et al 2003), but it was first proposed by Alan Turing, who often is mistakenly considered the defender of a rationalist, disembodied approach to AI. Certainly, Turing was not a theorist of embodiment and did not propose a fully-fledged theory of embodied cognition, but its speculation is useful to deepen the theory that Varela and colleagues proposed several decades later, observing how it has been applied by iCub researchers like Cangelosi, Metta, Vernon, Sandini, Tani, Wykowska, and others.

Core Publication #1 **[CP1]**

Cappuccio M.L. (2015),
“The Seminal Speculation of a Precursor:
Elements of Embodied Cognition and Situated AI in Alan Turing”
in Vincent C. Müller (ed.), *Fundamental Issues of Artificial
Intelligence*, Synthese Library 376, Springer.

Reproduced with permission from Springer. All rights reserved.

Chapter 28

The Seminal Speculation of a Precursor: Elements of Embodied Cognition and Situated AI in Alan Turing

Massimiliano L. Cappuccio

Abstract Turing's visionary contribution to cognitive science is not limited to the foundation of the symbolist approach to cognition and to the exploration of the connectionist approach: it additionally anticipated the germinal disclosure of the embodied approach. Even if Turing never directly dealt with the foundational speculation on the conceptual premises of embodiment, in his theoretical papers we find traces of the idea that a cognitive agent must develop a history of coupling with its natural and social environment, and that primitive bodily stimuli like pain and pleasure drive this coupling and elevate it to real learning by setting its normative preconditions. Turing did not consistently defend the centrality of embodiment, and ended up confounding or deemphasizing in various occasions the critical importance that he had himself implicitly recognized to the body. In line with the anti-representationist, radically enactive approaches to basic cognition, I believe that if Turing eventually failed to fully value the cognitive-developmental role played by the body, this was not because he proposed a computational and functionalist model of the mind, but because he tacitly assumed the content/vehicle dichotomy as a primitive of that model: in fact, he still believed that intelligence is realized by decontextualized contents that can be detached and transmitted regardless of their mode of physical implementation.

Keywords AI • Alan Turing • Developmental robotics • Embodied cognition • Connectionism • Cognitivism

28.1 Turing's Legacy and Cog Sci

Alan Mathison Turing is regarded as one of the fathers of computer science, and since computers represented for decades the main paradigm for the sciences of the mind (Pinker 2002; Boden 2006), his name is inextricably tied to the very

M.L. Cappuccio (✉)
United Arab Emirates University, Al Ain, UAE
e-mail: massimilianocappuccio@hotmail.com

birth of cognitive science as a research program. To evaluate the importance of this legacy, one must situate it in a historical perspective, considering the three prototypical approaches that (according to Varela et al. 1991) inspired cognitive science through its evolution. First, the cognitivist – or classical – approach, which is logico-symbolic in nature and based on the idea that a cognitive system is centrally realized by the sequential manipulation of discrete units of information in accord with formal rules of combination and substitution (Miller 1956; Neisser 1967; Minsky 1974; Haugeland 1978). Second, the connectionist approach, tied to the idea that information, realized by superpositional and distributed representations, is stored and manipulated through parallel systems implemented by neural networks (McCulloch and Pitts 1943; Rumelhart et al. 1986; van Gelder 1991). Finally, the embodied approach (Varela et al. 1991; Clark 1997; Gallagher 2005), stressing that the details of the material, organismic, and experiential implementation of bodily skills and behaviors play an active causal role in structuring cognitive functions (for example, fine-tuning the sensorimotor capabilities of a system, cfr. Noe and O'Regan 2002).

The first approach defines intelligent decisions as controlled by stored sets of 'rules of thumb' that reflect an internally localized representation of the behavior to be delivered by the system in specific circumstances. In opposition to this, the second approach says that the contents of these representations are memorized through stochastic reinforcement mechanisms when the weights of the network nodes are modified by environmental feedbacks. The third approach maintains, contrary to the first and the second, that the coupling between the system and the environment is not stored as a contentful representation of states of affairs, but a dynamical (Dreyfus 2002; Freeman 2000) and emergent (Petitot et al. 1999) know-how that both modifies and is modified by the system's perceptual and motoric dispositions while adapting to contextual circumstances. Some versions of the embodied approach stress that the cognitive processing responsible for this adaptive capability can be partly outsourced to bodily (non-neural), or even extra-bodily, elements (Clark and Chalmers 1998); others underline the necessary contribution of situated interaction and mutual unprincipled engagement in development and evolution (De Jaegher et al. 2010; Hobson 2002; Reddy 2008).

The history of computer science saw in Turing one of the heroes of the first approach. Turing's computational machine (also called "discrete-state machine", Turing 1948) provides the prototypical model to envision how the algorithms underlying certain intelligent functions can be centrally controlled and implemented by logico-symbolic operations realized through mechanical means. However, while most of his colleagues were busy trying to replicate human intelligent functions through the logico-symbolic machines that he had theorized, Turing was designing connectionist intelligent systems, as he had already started working on some of the earliest models of parallel distributed processing through neural networks (which he called "unorganized machines" in 1948). In this paper, I would like to point out that Turing's visionary contribution to cognitive science is not limited to the foundation of the first and to the exploration of the second approach to cognition, as it additionally anticipated the germinal disclosure of the third approach.

In some of his most speculative works (Turing 1948, 1950), Turing began pondering upon at least some of the key insights that a few decades later inspired the embodied approach: namely, we find traces of the idea that a cognitive agent must develop a history of coupling with its natural and social environment, and that primitive bodily stimuli like pain and pleasure drive this coupling and elevate it to real learning by setting its normative preconditions. Additionally, even if this chapter will not deal with Turing's contribution to theoretical biology, it is important to mention that Turing's views on the development of cognition go on par with his ground-breaking work on morphogenesis (1952), where he characterized living functions as developing from bio-chemical patterns that are not stored or represented as sets of rules, but generated by the stochastic equilibrium of reaction-diffusion processes determined by the fine-grained interplay of organismic predispositions and aleatory environmental circumstances. This is particularly important for Turing's treatment of self-organizing learning machines: as pointed out by Lassègue (1998, p. 113), their "self-organization progressively constitute for him 'the simplest model of a nervous system having a random assemblage of neurons'. There is the outline of what later will become the domain of the 'neural networks' and of their adaptive properties."

Attributing to Turing the paternity of the embodied-embedded approach to cognition would be excessive and misleading, as he never directly dealt with the foundational speculation required by the theoretical premises of embodiment, and he certainly didn't have sufficient time and technological means to explore it empirically. Nonetheless, re-reading his works 100 years after his birth, one realizes that some of the key conceptual points seminally addressed in his papers are the same that, much later, motivated the analyses on the embodied nature of cognition. I would like to try to show that, if Turing was certainly not the father of embodied cognition, he was not a frontrunner of cognitivism either, even if it is true that his work turned out to be instrumental to support the rise of cognitivism.

28.2 Cog Sci and Turing Machine

In 1936, Turing proposed the first speculative model of a "logical computing machine" (later simply referred by others to as "Turing machine") to demonstrate what later would be called "Turing thesis" (or "Church-Turing thesis", Kleene 1952): any computable number or sequence is computable by such a machine. This means that any algorithmic function can be realized through a finite number of operations conducted on discrete symbolic elements and decided by a pre-given set of rules of manipulation, or "tables of behavior", which are both logically (formally) necessary and physically (mechanically) deterministic. But this thesis does not imply that the functioning of the human mind is computational in nature or realized by algorithms. We cannot find traces of this idea in the 1936s paper, which aimed at demonstrating the inverse assertion: each algorithmic process of thinking that can be carried out by a human being through formal procedures can be exactly reproduced by computational and mechanical means. These two assertions are compatible, yet their difference is substantial: the latter describes the powers and the constraints

of the recursive operations that we can effectively use to formalize mathematical processes, and does not imply the former, which is ontologically stronger, and relates to the deep constitution of psychological processes. The latter states that a certain practice (calculus) shapes a certain form of thinking (calculative), and that a key feature of that practice can be mechanized; the former asserts that if that practice is mechanizable it is because the underlying form of thinking is inherently mechanical.

This difference was destined to be overlooked, after Turing, especially because computer science started accumulating striking successes in replicating human formal procedures by means of sophisticated electronic implementations of the Turing machine. Since then, the universal Turing machine has been seen more and more an irresistible metaphor to describe in mechanical terms the fundamental processes that govern the functioning of the human mind, reviving the long-standing rationalist belief that thought is, after all, nothing else than calculus on mental contents (Longo 1999).

Skepticism towards this expectation has been growing especially in the last decades, and today we are better inclined to appreciate the primary and truest sense of this metaphor. Lassègue (1998, p. 202) observed that the great epistemological novelty brought by Turing's machines is "the biological aspect of this characterization of the mechanism: therefore, it is more the machine that resembles the organism than the organism resembling the machine". Ludwig Wittgenstein (1980, § 1096) had already warned us, with his dry comment, long ago: "Turing machine. These are humans who calculate". These comments, stressing the biologic and anthropomorphic background of Turing's idea, invite us to reject the cognitivist interpretation of Turing thesis, i.e. the claims that – after all – human minds are nothing other than calculating machines, and replace it with the inverse, constructivist interpretation: calculating machines – after all – are nothing other than devices built to surrogate a human practice.

The latter interpretation turns out to be not only plausible, but compelling, if we take seriously the fact that Turing thesis, exactly like its equivalent formulations (e.g., lambda-calculus, cfr. Church 1936; recursion theory, Kleene 1936), is not a theorem demonstrable by the power of mechanical calculus, but an intuitive evidence ("a direct appeal to intuition", Turing 1936, p. 249) revealed by the phenomenology of the human practice of calculus. The importance of this phenomenology appears clearly in Sect. 9 of *On computable numbers*, where Turing reveals what his machine was originally meant to be: a stylized illustration of the embodied writing/reading operations that have to be conducted by a "computer" in flesh and bone, when (s)he is busy at a desk to solve some mathematical problem with pencil and paper.

28.3 Turing Machines and Human Computers

Looking closer at how Turing describes the structure of his machine in Sect. 9 of his 1936 paper, we see how he transfigures the bodily elements of the human computational practices, replacing them with their mechanical counterparts: the

bi-dimensional squared paper sheets used by humans as physical supports of computation are substituted by the mono-dimensional tape of the machine, whose only functionally relevant characteristics are its unconstrained linear extension, its double direction of movement, and the fact that it is segmented in a series of equal “sections”; the body of the human computer, in turn, with its organs and appendices dedicated to manipulation and vision, is substituted by the moving head of the machine, whose only functions are sliding on the tape and scanning, typing, or erasing a symbol at once; finally, the attentional, dispositional, and ideational capabilities of the human computer are reduced to the machine’s moving head’s power to be “directly aware” of a single symbol at once and “keep in mind” the functional “configuration” that determines its following action (p. 231). The explanatory power of this illustration, which offers an abstractly universal, yet fully operative account of the mechanical nature of algorithmic practices, derives precisely from being both extremely concrete and, at the same time, effective in obliterating the original source of such concreteness. In fact, while it clearly shows that the human computer’s body is the necessary and constitutive precondition for any computational operation, it concurrently reduces the role of the body to a redundant (functionally irrelevant) material accident of that very operation.

This ambivalence, which preludes to Turing’s incomplete investigations of the embodiment of intelligence, is hinted at by the grammatological norms that define the machine’s behavior. On the one hand, the operations illustrated by the Turing machine are instantiations of a writing practice, i.e. the simplest embodied acts that a human computer must carry out to calculate, “so elementary that it is not easy to imagine them further divided” (p. 250). As constituents of a graphic procedure, they can be sharply distinguished as either sensory (scanning a series of symbols on a sheet, one at a time) or motoric (erasing the currently scanned symbol or printing a new one, and moving to scan another symbol), their goal being, respectively, to recognize and produce a series of written graphic elements that belong to a well-defined alphabet.

On the other hand, just because the practice of mathematical calculus is analyzed into allegedly atomic graphic manipulations reducible to logical operations in a dimension of virtual necessity, Turing’s illustration is stylized and decontextualized. What it captures of the living practice of writing is only its formal features, i.e., the discrete transitions that generally look (to a mathematician) strictly indispensable to account for the logical role played by writing in the creative task of solving a mathematical problem. Anti-formalist mathematicians (Longo 2002, building on Weyl 1985) and psychologists of rational processes (Lakoff and Núñez 2000) stress that there is much more in mathematical writings than blind and empty rule-based mechanical procedures; at the same time, they should recognize that, if such formalistic reduction is virtually possible, it is primarily because it is allowed by certain writing practices whose norms are discrete and mechanistic in nature.

Note: Turing machine is not just a bodiless replication of the body of a living mathematician; it is also an idealized explication of the norms that guide her intelligent skills, transposed into an abstract space of typographic exactness (Cappuccio 2005, 2006; Herrenschmidt 2007; Lassègue and Longo 2012). Turing’s

insight into the structural preconditions of the mechanization of computation matches very well with Turing's late reflection, where – among other things – the British polymath investigated the theoretical and practical possibility to reproduce the entirety of human thought by mechanical means. We should, therefore, refer to *Intelligent machinery* (1948) and to *Computing machinery and intelligence* (1950), two influential texts in which Turing's speculative enquiry into the possibility to build machines capable of human-like intelligence is famously addressed. “Can machine's think?”, the question that specifically stimulates his 1950 paper confirms that the ambitions of his investigation were still in line with his earlier work on the nature of computation: once again, the question is whether (and how) machine intelligence could one day become so complex to rival human thinking, not whether the processes of human thought are inherently mechanical.

Turing famously refuses to address the question in this ontologically demanding form, which – according to him – typically relies on very speculative and too abstract notions of “thought”. Rather, the proposed criterion to evaluate whether intelligence has been achieved or not by machines is entrusted to an “imitation game”, a procedure to empirically determine whether a human player's interaction with a machine appears to him indistinguishable from the usual interaction with other humans. Discussing Turing controversial choice on this regard, the commentators often stressed that this method leads to collapse the distinction between truly intelligent decisions and stereotyped routines that mimic their exterior appearances. This reflection also raised well-motivated concerns about some of the strongest explanatory expectations on AI (Searle 1980).

But there is another issue that deserves to be deepened: why do we need a game based on the actual human-machine interaction (whether competitive or cooperative in nature) to test the successful implementation of intelligence by artificial means? Interestingly, what Turing proves in this famous paper is not that in order to reproduce intelligence a machine needs a certain logico-symbolic computational architecture, but that it must be capable of adapting to the phenomenology of a social encounter with the human (Gallagher 2009, 2012). That is why one can't find any ultimate commitment to a logico-symbolical account of human intelligence in Turing's work (Longo 2008). On the contrary, the imitation game shows that a certain behavior looks credibly intelligent to a human observer only if she recognizes it as familiar; and that this familiarity is measured by the degree of interactive attunement that the observer spontaneously tends to establish with the AI through conversation.

But we know that such attunement largely depends on experiencing a corporeal relation of reciprocity (De Jaegher et al. 2010) – a bond that might be mediated by the capability to empathize through one's own body with the body of another (Gallese 2009). At a first glance, in Turing's description of the imitation game the role of the body of the participants (whether human or robotic) seems systematically removed: by design, their physical identity is hidden, and their communication is restricted to perceptually standardized and emotionally neutral exchanges of strings of symbols. The player can only infer the interlocutor's intentions, which can't be directly manifested through bodily expression, unlike in the typical human forms of social engagement. According to Turing, the imitation game is reliable precisely

because it measures the intelligence of the machine by drawing “a fairly sharp line between the physical and the intellectual capacities of a man.” He seems convinced that there would be little point in trying to make a “thinking machine” more human by dressing it up in [. . .] artificial flesh”, even if “engineers and chemists” were one day able to provide a machine with such bodily features.

This seems irremediably at odds with any interactionist and embodied approach to cognition and to sociality. Nevertheless, I would like to stress that, if Turing pointed to a social interactive context as the most reliable scenario to test the successful development of intelligence, this is because he had clear in his mind that direct, unprincipled embodied forms of intersubjective engagement are constitutive *at least* of the earliest development of their intelligence, and that the early bodily forms of interactive engagement with others are a precondition for the most developed forms of intelligence. Indeed, Turing takes seriously the psychological link between embodied sociality and cognitive performance, though he doesn’t thoroughly discuss its implication: he is always tempted to replace the defining role played by the body in its contingent modes of implementation with context-independent heuristics, formal rules, and mechanical procedures; nonetheless, rather than stating that the human thought is entirely reducible to logico-symbolic processes, he carefully describes how, in order to achieve intelligence, even a machine needs a body.

Two facts suggest that Turing was struggling to move beyond the cognitivist framework. In the first place, he never claims that, in order to think, a cognizer necessarily needs to implement a cognitivist cognitive architecture (logico-symbolic, sequential, discrete, centrally controlled, syntactically organized, fully-representational, internally localized); on the contrary, as I mentioned before, in 1948 he had carefully considered the functionality of parallel-distributed connectionist models, and was well aware that brains are “unorganized machines”, highly plastic systems sensitive to limit conditions (Longo 2002, p. 3). In the second place, in 1950, Turing claims that, even if human cognitive functions can in principle be *imitated* by logico-symbolic processes (and this is why he keeps considering discrete-state machines as the standard for his game), the *emergence* of these functions (i.e., the developmentally necessary preconditions of their genesis and actual implementability) must be rooted into an embodied process of learning based on situated interactions and unprincipled exploration. Indeed, he admittedly does not have any positive arguments to convince his readers that digital computers can fully imitate human intelligence (p. 443) and, in a constructivist fashion, he can only rely on the persuasiveness of the description of a child-machine as a genetic model of the educational and experiential conditions that could lead a computer to become intelligent. Significantly, in his account, the question into the possibility to recreate human intelligence by mechanical means becomes a question into the possibility to recreate human development by artificial means.

Now, Turing did not consistently defend the centrality of embodiment, and ended up confounding or deemphasizing in various occasions the critical importance that he had himself implicitly recognized to the body. But, this unfinished speculation can’t cancel the fact that the 1950 paper is not meant to support a mechanist reduction of intelligence to computation. This point is ignored by his commentators

more often than it should: Turing knew that an intelligent system requires a fine-grained calibration to contextual circumstances, and that only a living organic body can achieve it through a history of complex adaptive interactions with the world and with other agents.

28.4 Human Computers and Child-Machines

Today, contextual circumstances of the embodied/embedded kind are considered key by the most promising approaches to A.I., e.g. situated (Hendriks-Jansen 1996), evolutionary (Nolfi Florean 2000) and developmental robotics (Cangelosi and Schlesinger 2015) which often go hand in hand with the new anti-cognitivist and post-connectionist wave in cog sci: according to these approaches, it is the body's material predisposition to balance the unpredictable fuzziness and massive complexity of the sensorimotor circumstances that establishes implicit and pre-categorical norms for the stability, effectiveness, and flexibility of intelligence; in turn, these norms can inform an environmentally entrenched intelligence by moulding the agent's bodily dispositions to action and perception in the ways that best allow the system to fluidly respond to concrete situations so to reduce its distance from a foreseen optimal balance with the contingencies (Freeman 2000; Dreyfus 2002). An embodied system learns by enacting through its actions complex feedback loops that continuously inform the system of its own sensorial, proprioceptive, and motoric possibilities and constraints (Brooks 1991; Clark 1997). Such body – with its intrinsic adaptive and predictive functions – is what specifically determines the form and the direction of the machine's cognitive processes, as it acts as a distributor, a constraint, and a regulator of the underlying mechanisms of decision and control (Shapiro 2011). But, most importantly, what enables and coordinates these three functions is the fact that the body constitutes the most fundamental background of the sense-making activity of a cognitive system: flexibly adaptive dispositions truly able to realize intelligence can only emerge if they are materially embedded as a tacit (irrepresentable) know-how that is directly responsive to concrete contextual contingencies (Dreyfus 2008). Bodily sensitivity to real contexts is what enables the immediate intelligibility of relevance and the prioritization of the teleological dispositions of the living system.

Turing certainly contemplated the possibility that embodiment could be a tacit precondition of intelligence, but suspected that the success of an embodied AI would be contingent upon substantial future advancements in bio-medical and bio-mimetic engineering, i.e. the fields of artificial biology that aim to produce fully working bodies equipped with a sensorimotor system dynamically coupled with the environment: all machines truly capable of developing intelligence must be implemented in such system; however, as the technological reproduction of these bodily functions was not even supposable at the end of '40s, Turing had to concentrate his interests to the formal architecture of the algorithms that could better approximate and facilitate it (cfr. Turing 1948, p. 9).

The pioneering solution that Turing envisages in order to develop real embodiment is still based on a mechanistic representation of life and intelligence, but groundbreaking if seen as a developmental prototype: machine learning through embodied exploration of the environment, both free and under human supervision. The key model is a “child-machine” (Turing 1950, p. 456) that learns from attempting various acts of manipulation and circumspection, monitoring their results, and broadening the repertoire of available actions and desired goals. This model is inspired by the evolutionary adaptation of biological systems to their environmental niche: in fact, the “structure of the child machine” (assigned by the engineer and built-in), the “changes of the child machine” (dynamically developed through interaction), and the “judgment of the experimenter” (crucial to select and reinforce the successful conducts of the child-machine), correspond respectively—in Turing’s proposal – to the “hereditary material”, the “mutation”, and the “natural selection” that determine the evolution of a real human infant (*idem*). Therefore, the living body is not only and not necessarily the result of natural evolution, as its characterizing function is to be the crucial mediator to learn from casual experience during development, i.e. (in Turing’s terms) to dynamically update the contents of the table of behavior of a machine (here also called “the book of rules”, cf. pp. 437–438) by fully exerting its very explorative/adaptive behavior. And this is a function that can be studied in a laboratory and possibly reproduced by artificial means.

Strikingly, this is the idea that today successfully inspires iCub, a tetrapodic anthropomorphic machine fully equipped with infant-like perceptual, motoric, and interactional bodily features, which in turn scaffold its capability to develop more and more complex action schemata (Metta et al. 2008; Cangelosi and Schlesinger 2015). The constant sensorimotor refinement of iCub’s manual expertise, empowered by an advanced grasping control and force control systems, enables it to improve and indefinitely broaden its skills of hand-eye coordination, tool use, and fine object manipulation. Manipulative and explorative attempts, in turn, are used by iCub to perfect its recognition capabilities (actions, affordances, objects, and faces), and to update its routines to comprehend the causes and predict the effects of its physical interactions, also in conjunction with verbal commands, responsive gazing behavior, and elementary capabilities of joint attention.

It is probably while Turing was caressing the dream to father a similar artificial child that he started realizing that the body is the fundamental constraint/organizer/regulator in all the evolutionary and developmental processes of adaptation. However, as stressed by Cappuccio (2006) and Lassègue and Longo (2012), Turing’s dream is still irremediably in debt with the metaphor of writing, a metaphor that once again he uses in these pages to describe the essence of the computational processes of the child-machine: “Presumably”, says Turing, “the child brain is something like a notebook as one buys it from the stationer’s. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.)” The persistence of this analogy (intelligence as a mechanical transfiguration of a writing system operating with contentful symbols) is precisely what brought him to misunderstand the role of the body in cognition.

28.5 Child-Machines and Multiple Realizability

Sixty years before the creation of iCub, Turing had already foreseen the importance of bodily functions for the development of a child-machine. Describing the child-machine's education, he regrets that "it will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition." (Turing 1950, p. 456) Even if his tone seems humorous, Turing is seriously concerned that not having a body and a satisfactory social life might ultimately undermine the educational path of the child-machine project.

In spite of his farseeing vision, he surprisingly tended to deny their unique contingent causal role. Eventually, he decides to postpone the final struggle with this problem, trusting that one day engineering sciences will find a way to surrogate the relevant functional role of the body parts that the child-machine is not (yet) entitled to possess. "We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other" (idem). Turing's reference to Helen Keller is both inspiring and puzzling, from the perspective of embodied cognition. Helen Keller was famously born deaf-blind, and it is only thanks to an intensive educative program and constant assistance throughout her early childhood that she was able to overcome her incapacitating isolation, get a university degree, and eventually become successful as a writer and political activist. Turing brings this example to argue that even a serious lack of sensorial and communicative capability can be overcome by a functionally equivalent information processing apparatus (i.e., reducing real communication to exchanges of strings of written symbols from/to the isolated mind of Miss Keller). What he fatally overlooks is that the educational process that led Miss Keller to comprehend her instructor, and thus learn the first basic set of deaf-language signs, was actually intensely physical, at least in the beginning. This process was scaffolded by the participatory social modulation of affective and sensorimotor stimuli through informal hand gestures, guided by tactile feedback and motion. What is crucial in this methodology is precisely the bodily engagement of the child with the instructor (Ford 2011). In the quoted paragraph, it seems that Turing misunderstands the role of the body, depicting it as a set of modular and interchangeable peripherals whose function is merely instrumental and supplementary to the algorithmic work of a centralized computational agency.

However, if we look back at his 1948 work, we discover that at a certain stage he was ready to bet on a rather different cognitive role for the body. Foreshadowing the idea of a child-machine, in this paper, Turing had clarified even better that, in order to learn, a machine should implement a body capable of reproducing the entirety of the human organs and a dynamic neuronal network that is "unorganized", i.e. not regulated by pre-given sets of rules, but ready to progressively "auto-organize"

while learning from its experience of actual interaction with the world. “The machine should have a chance of finding things out for itself” (Turing 1948, p. 9) and this means “it should be allowed to roam the countryside” and, so to speak, make its own experience in the external world: by this reason, it should be endowed with legs and eyes (Turing also notes, with a sarcastic remark that probably targets the repressive conformism of his society, that “the danger to the ordinary citizen would be serious”). “However”, complains the mathematician, “even when the facilities mentioned above were provided, the creature would still have no contact with food, sex, sport and many other things of interest to the human being”. Interestingly, Turing lists among the pedagogical desiderata a series of bodily stimuli that humans typically consider enjoyable, as he considers them decisive to set the pre-categorical parameters that define desirable goals and normative conditions for the machine’s dispositions to act. In operative terms, this is the primitive system of rewards that the machine’s instructor can use to condition the machine’s behavior. Indeed, Turing explicitly attributes a notable importance to pain/pleasure reinforcement mechanisms to guide the machine’s learning experience and evaluate the success of its attempted actions, as these are the stimuli that the machine itself can use to measure its own behavior’s distance from an competent performance. For the happiness of a behaviorist, to train a machine sensible to these stimuli would be more like conditioning an animal than compiling a computer program. But it is not indispensable to be a behaviorist to appreciate the phenomenological primacy of these intentional bodily experiences in scaffolding the development of intelligence.

Today we are inclined to believe that the conditioning force of bodily affects such as pleasure and pain is strictly inherent to the history of body-environment negotiations enacted through our experience (Bower and Gallagher 2013). This ecological view emerged only partially in Turing’s conception (as highlighted by Wells 2005), which is still struggling to overcome the resilient belief that intelligence is envatted, i.e. reducible to input signals passively received by some sort of solipsistic Cartesian central module. In his conclusive remarks, Turing does not get to the point of declaring the body as a stringently (a priori) requisite for all kinds of intelligence, but admits that embodiment is an authentic and reliable strategy to actually reproduce intelligence, and moreover it is the most reliable one. In fact: “this method [unprincipled learning through spontaneously reinforced embodied interactions] is probably the ‘sure’ way of producing a thinking machine”, although – adds Turing – “it seems to be altogether too slow and impracticable”, probably also due to the technological limitations of his era. I think these observations prove that Turing never overlooked how real life contexts can directly incite action in actual bodies even when we do not have any internal model of those situations.

If we look at Turing’s speculation on AI in the light of his work in theoretical biology we realize that he knew that actual bodies are necessary preconditions to make sense of real life contexts. According to the embodied approaches to cognition, if humans and other animals are able to adapt to the always exceeding complexity of real life circumstances, this can only happen through the negotiation offered both passively and actively by the body, through the details of physical

constitution, the retentive/protentive structure of its temporally extended immanent presence, and its embedded adaptive dispositions and skills. Turing himself was probably close to make a similar conclusion, against an instructionist view of biological and cognitive functions: his morphogenetic models can account precisely for the way the infinitely variable complexity of the contextual circumstances modulates the fine-grained constitution of the bodily functions even at the molecular level, adjusting them to the stochastic fluctuations of the environmental conditions. According to Longo, these “composite systems of action, reaction, and diffusion” allow a form of “calculus” that “entirely reside in the continuous dynamics of the forms. It is not anymore about an immaterial software, but constantly evolving unpredictable plasticity and materiality.” (Longo 2009) In other words, the mechanical interactions between the material components of the system are not guided by some superimposed immaterial mental content; it is the very capability of the system to organize itself that provides it with intelligent adaptivity and flexibility. There is no meaning in life other than life itself, and this is what guides biological entities to develop thoroughly intelligent dispositions.

28.6 Multiple Realizability and the Frame Problem

However close to making this conclusion Turing was, he actually failed to spell it out – as he was clearly concerned about how “slow and impracticable” the embodied approach to AI could have been. The surrogate solution eventually proposed by Turing in 1948 (p. 9) to overcome the technological impossibility to endow a machine with a living body is “to try and see what can be done with a ‘brain’ which is more or less without a body, providing at most, organs of sight, speech and hearing”. Turing has clear in his mind that the disembodiment of cognition can, at best, produce some formalistic approximation, or simplified prototype, of an intelligence dedicated to pre-regulated synthetic environments. Otherwise, he wouldn’t have stressed repeatedly, and with such emphasis, the importance of implementing the right bodily features. It is true, however, that eventually Turing does not overcome the idea that the body is merely an extension, a replaceable appendice, of cognition, rather than its inseparable backdrop and its condition of possibility.

After more than sixty years spent “trying and seeing” to build a body-less “electronic brain”, most AI theorists abandoned this idea, turning themselves to the creation of intelligent functions that are deeply integrated, directly modulated, and actively scaffolded by their material constitution, deemphasizing more and more the role of encapsulated decision-making algorithms and centralized control systems: for example, the “soft” robotics (Pfeifer et al. 2012) and the bio-mimetics (Vincent 2009) approaches to AI, privilege solution in which the material background of cognition is not just a realizer of its logico-causal configurations, but its defining component. A theoretically relevant aspect of these trends is that they deemphasize the principle of multiple realizability (Putnam 1967a), which in philosophy is

typically associated to machine functionalism (Putnam 1967b) and a certain version of supervenience theory, and often exemplified by the distinction between the computational, the algorithmic, and the implementational levels of a cognitive architecture (Marr 1981): a cognitive process, exactly like a computer software, is multiply realizable when it depends on, but doesn't functionally identify with, the material details of its contingent implementation, which can therefore be realized by different configurations of "hardware" components (transistors, microchips, neurons...) as long as their causal relations bring about the same algorithmic and logical operations. Turing's proposal to substitute the living experience allowed through legs and eyes with equivalent sets of symbolic information manipulated through centralized algorithms clearly suggests that he was captivated by some version of the principle of multiple realizability, one that is at odds with the very idea of embodiment. In truth, at least some interpretations of multiple realizability (e.g., Wheeler 2010) don't exclude embodiment, and actually aim at accounting for the fine-grained details of the causal co-involvement of body and world in shaping cognitive functions at a sub-declarative level. That is why I believe that Turing's oversight is imputable to a particular conception of the body that is possibly derivable from the multiple realizability principle, but that is not implied by it.

This conception, which has often been the main philosophical alibi for the disembodiment of intelligence, is the one that postulates that the dichotomy of contents (the meaningful information represented by the system and manipulated by its functions) and vehicles (the meaningless realizers of the representational function) is a distinctive mark of cognition. In fact, this view, which is still mainstream in the cognitive sciences, maintains that the content/vehicle distinction is necessary in any explanatory discourse on mental functions because it assumes that the specific normativity of the mental can only be accounted for in terms of contents. Today, some anti-representationist trends in embodied/enactive cognition reject this view: i.e., the Radically Enactive Cognition, or REC (Hutto and Myin 2013). REC stresses that in no way this distinction could account for the basic structure of cognitive processes and that it actually is a categorical mistake to describe the functional realizers of cognitive processes as material containers of intangible informational contents, because any attribution of intentionality (mental states) to physical (biological or mechanical) processes inevitably violates the naturalistic assumptions behind the scientific study of the mind. Also other arguments, relevant to our discussion of Turing's legacy, could be used to undermine the belief that detached representational content is a necessary component of cognition.

For example, the arguments related to the "frame problem", a philosophical term used to describe an artificial system's incapability to sense contextual relevance (Dennett 1987; Dreyfus 1992; Gallagher 2012; Cappuccio and Wheeler 2012). Dreyfus imputes this deficiency to the fact that, ignoring how any truly intelligent decision occurs against a bottomless background of preconditions, classical AI still relies on self-contained, limited representations or heuristic models of the world: but reality transcends any attempt to capture the complexity of a situation through its representations, i.e. content-bearing models or sets of instructions. In Turing machines, these representations are the symbolic models stored in the

tables of behavior of the machine. With regard to Turing's theoretical approach to computation, the frame problem is exactly the incapability of a discrete-state system (which uses finite sets of representations/instructions) to make sense of a transcendent continuous reality and to respond to its intrinsic relevance for a certain task.

Turing is aware of the existence of a similar problem, that he somehow addresses while discussing the "Argument of Informality of Behavior" against mechanical intelligence (1950, p. 452). This argument claims that a machine will never be able to reproduce the complexity of human behavior because "it is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances". Turing confesses to "agree" with the skepticism towards the idea of reducing intelligence to a rule-based system of decision making procedures, and since he also agrees that discrete-state machines are governed by rules, he finally admits that the human mind is not a machine of a logico-symbolic kind: this is sufficient to definitively erase Turing's name from the list of the alleged supporters of classical cognitivism. However, his overall judgment is nuanced. In fact, he denies that all human cognitive functions are irreproducible by mechanical means, and that all rule-based procedures should be excluded from consideration: Turing proposes a neat distinction between "rules of behaviors", which define the normative procedures used in explicit deliberation, and "laws of conduct", which describe the dispositions that prepare a cognizer to pre-reflectively respond to real-life complex circumstances. Conceptually, this distinction matches Dreyfus' separation between representational intelligence (mediated by contentful heuristics) and skillful coping capability (enacted through embodiment by contentless dispositions to real relevance). Therefore, one could well figure that Turing had already intuited that true intelligence could only be achieved by abandoning a rule-based representational architecture to avoid the Frame problem. At the same time, because the concept of machine that Turing had in mind was different than a deterministic step-by-step apparatus, he is adamant in asserting that even Frame problem-avoiding minds could be effectively imitated by machines of some type, though, probably, not of the discrete, logico-symbolic type: "we believe that it is not only true that being regulated by laws of behavior implies being some sort of machine (though not necessarily a discrete-state machine), but that conversely being such a machine implies being regulated by such laws".

28.7 Conclusions: The Frame Problem and the Symbol-Grounding Problem

The key issue that Turing never addresses is that the Frame problem should *necessarily* arise in discrete-state machines if we assume that the symbols used to instruct their tables of behavior form a system of contentful representations using principles of logical isomorphism/projection to capture the relevant aspects of a typical situation. The deeper problem then is how these representations could be

meaningful at all, as we don't have any general account of how representations could acquire their contextual content or be coupled to the practical background of real life.

This question discloses another problem that affects all the paradigms that rely on internally stored informational contents to explain intelligence, regardless that these paradigms assumed the transparency (symbolic approach) or the opacity (connectionist approach) of content: the "Symbol-grounding problem" (Harnad 1990). As illustrated by the 'the Chinese room' thought experiment (Searle 1980), this problem concerns the difficulty of explaining the origin of the meaning of the decisions taken by an intelligent system, provided that the representations manipulated to take decisions are conventional in nature, and therefore not meaningful at all to the system itself. The problem of the non-derivability of meaning (and thus relevance) from mere syntactical procedures, is even better illustrated by the historical circumstances in which understanding the meaning of an encoded message requires, first of all, to know the broader real-life context and the underlying communicative intention, rather than blindly apply a set of algorithms for symbolic substitution. Turing probably had to intensely struggle with the intrinsic semantic emptiness of conventional symbols, during his code breaking experience at Bletchley, which certainly solicited him to appreciate the collaboration with linguists, anthropologists, and experts of German culture, beside logicians and mathematicians: the peculiar participatory experience of trying to make sense of an encrypted message probably made him aware that the pragmatic meaning of a string of symbols does hardly come only from their syntactic rules of combinations, as it first of all emerges from the broader context (historical, cultural, and bodily) of their actual use. Symbols are never sufficient to recreate their context of meaning, if the agents who manipulate them are not already embodied in it: only a staunch formalist mathematician would not recognize it. Turing was not a formalist of this kind but, if he occasionally failed to remember this lesson, in his late speculation on AI, this is because – still influenced by the heritage of a formalist thought – he was never entirely capably to get over the belief that semantic content is a commodity that could be conventionally attached to symbols and freely transferred from a string of symbols to another.

Summing up, in line with the anti-representationist approaches to basic cognition like REC, I believe that if Turing eventually failed to fully value the cognitive-developmental role played by the body, this was not because he proposed a computational and functionalist model of the mind, but because he tacitly assumed the content/vehicle dichotomy as a primitive of that model: in 1950, he still believes that intelligence is a matter of accessing contents, allegedly detachable and transmittable from a sender to a receiver through a string of conventional symbols. However, because Turing revolutionary approach to the mechanization of symbolic processes was not aware of this implicit metaphysic, his belief was not dogmatic. On the contrary, as the British mathematician was keen on valuing the situated context of real intelligence, in his work we can find certain premises to think cognition as embodied and representation-less. However, as we have seen, Turing's approach to this issue is uncertain: on the one hand, its computational machines promoted

a tacit symbolization (representational and formal inscription) of the mechanisms of our minds, suggesting that this functioning is content-bearing in nature; on the other hand, his proposal of child-machines provided us with important reflections to abandon the representational paradigm and rethink the meaning of the grammatical metaphor.

References

- Boden, M. A. (2006). *Mind as machine: A history of cognitive science* (Vol. 2). Oxford: Oxford University Press.
- Bower, M., & Gallagher, S. (2013). Bodily affects as prenoetic elements in enactive perception. *Phenomenology and Mind*, 4(1), 78–93.
- Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1–3), 139–159.
- Cangelosi, A., & Schlesinger, M. (2015). Developmental robotics. From babies to robots. Cambridge, MA: MIT Press.
- Cappuccio, M. (2005). Un'ipotesi controfattuale: la macchina di Turing ideografica. In L'eredità di Alan Turing. 50 anni di intelligenza artificiale (pp. 231–244). Milano: Alboversorio.
- Cappuccio, M. (2006). *Alan Turing: l'uomo, la macchina, l'enigma. Per una genealogia dell'incomputabile*. Milano: Alboversorio.
- Cappuccio, M., & Wheeler, M. (2012). Ground-level intelligence: Action-oriented representation and the dynamic of the background. In Z. Radman (Ed.), *Knowing without thinking*. London: Palgrave Macmillan.
- Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2), 345–363.
- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 7–19.
- De Jaeger, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10), 441–447.
- Dennett, D. (1987). Cognitive wheels: The frame problem in artificial intelligence. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence*. Norwood: Ablex.
- Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.
- Dreyfus, H. L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, 1, 367–383.
- Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland, & M. Wheeler (Eds.), *The mechanical mind in history*. Cambridge, MA: MIT Press.
- Ford, J. (2011). Helen Keller was never in a Chinese room. *Minds and Machines*, 21(1), 57–72.
- Freeman, W. J. (2000). Neurodynamics: An exploration in mesoscopic brain dynamics (Perspectives in neural computing).
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.
- Gallagher, S. (2009). The key to the Chinese room. In K. Leidlmaier (Ed.), *After cognitivism* (pp. 87–96). Dordrecht: Springer.
- Gallagher, S. (2012). Social cognition, the Chinese room, and the robot replies. In Z. Radman (Ed.), *Knowing without thinking*. London: Palgrave Macmillan.
- Gallese, V. (2009). Mirror neurons, embodied simulation, and the neural basis of social identification. *Psychoanalytic Dialogues*, 19, 519–536.
- Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42, 335–346.

- Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences*, 1, 215–260.
- Hendriks-Jansen, H. (1996). *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*. Cambridge, MA: MIT Press.
- Herrenschmidt, C. (2007). *Les trois écritures. Langue, nombre, code*. Paris: Gallimard.
- Hobson, P. (2002). *The cradle of thought*. Oxford: Oxford University Press.
- Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without contents*. Cambridge, MA: MIT Press.
- Kleene, S. C. (1936). Lambda definability and recursiveness. *Duke Mathematical Journal*, 2, 340–353.
- Kleene, S. (1952). Introduction to metamathematics. Amsterdam: North-Holland.
- Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.
- Lassègue, J. (1998). *Turing*. Paris: Les Belles Lettres.
- Lassègue, J., & Longo, G. (2012). What is Turing's comparison between mechanism and writing worth? In S. B. Cooper, A. Dawar, & B. Loewe (Eds.), *Computability in Europe* (pp. 451–462). Berlin/Heidelberg: Springer.
- Longo, G. (1999). The difference between clocks and Turing machines. In A. Carsetti (Ed.), *Functional models of cognition* (pp. 211–232). Dordrecht/Boston: Kluwer.
- Longo, G. (2002). The constructed objectivity of mathematics and the cognitive subject. In M. Mugur Schacter (Ed.), *Proposals in epistemology. On quantum mechanics, mathematics and cognition* (pp. 433–463). Dordrecht/Boston: Kluwer.
- Longo, G. (2008). Laplace, Turing and the “imitation game” impossible geometry: Randomness, determinism and programs in Turing's test. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing test* (pp. 377–413). Dordrecht: Springer.
- Longo, G. (2009). From exact sciences to life phenomena: Following Schrödinger and Turing on programs, life and causality. *Information and Computation*, 207, 545–558.
- Marr, D. (1981). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: The MIT Press.
- McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
- Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition, PerMIS '08. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (pp. 50–56).
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Minsky, M. (1974). A framework for representing knowledge. MIT Lab Memo # 306.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Noe, A., & O'Regan, J. K. (2002). On the brain-basis of visual consciousness: A sensorimotor account. In A. Noe & E. Thompson (Eds.), *Vision and mind: Selected readings in the philosophy of perception*. Cambridge, MA: MIT Press.
- Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.
- Petitot, J., Varela, F., Pachoud, B., & Roy, J. M. (Eds.). (1999). *Naturalizing phenomenology: Contemporary issues in phenomenology and cognitive science*. Stanford: Stanford University Press.
- Pfeifer, R., Lungarella, M., & Lida, F. (2012). The challenges ahead for bio-inspired ‘soft’ robotics. *Communications of the ACM*, 55(11), 76.
- Pinker, S. (2002). *The blank slate*. New York: Penguin.
- Putnam, H. (1967a). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
- Putnam, H. (1967b). The nature of mental states. In D. M. Rosenthal (Ed.), *The nature of mind* (pp. 197–203). New York: Oxford University Press.
- Reddy, V. (2008). *How infants know minds*. Cambridge, MA: Harvard University Press.

- Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Volume 1: Foundations). MIT Press: Cambridge, MA.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417–457.
- Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.
- Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem: A correction. *Proceedings of the London Mathematical Society*, 43, 544–546.
- Turing, A. M. (1948). Intelligent machinery. Report for national physical laboratory. In: *Collected works* (Vol. 1).
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 50, 433–460.
- Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London, Series B*, 237, 37–72.
- van Gelder, T. (1991). Classical questions, radical answers: Connectionism and the structure of mental representations. In T. Horgan (Ed.), *Connectionism and the philosophy of mind*. Dordrecht/Boston: Kluwer Academic Publishers.
- Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: The MIT Press.
- Vincent, J. F. V. (2009). Biomimetics, a review. *Journal of engineering in medicine. Proceedings of the Institution of Mechanical Engineers. Part H*, 223(8), 919–939.
- Wells, A. (2005). *Rethinking cognitive computation: Turing and the science of the mind*. London: Palgrave Macmillan.
- Weyl, H. (1985). Axiomatic versus constructive procedures in mathematics. In T. Tonietti (Ed.), *The mathematical intelligence* (7)4. New York: Springer-Verlag.
- Wheeler, M. (2010). In defence of extended functionalism. In R. Menary (Ed.), *The extended mind*. Cambridge, MA: MIT Press.
- Wittgenstein, L. (1980). *Remarks on the philosophy of psychology* (Vol. 1). Oxford: Blackwell.

Chapter 1, Part 3: Modelling Attention and Control in Skilful Performance

This section is centred on Core Publication #2 [CP2]

1.3.1 Embodied Cognition applied to Cognitive Robotics

Inspired by Turing, we need to discuss how ECT influences Cognitive Robotics. In this section, I present in greater detail the EECR, focussing on embodied attention as a key feature of enactive cognitive robots. EECR emphasizes the importance of a robot's physical body and its environment in shaping its intelligence and behaviour. Unlike GOFAL, which prioritizes scripted models and abstract reasoning, EECR creates intelligent robots via training techniques such as ML, focusing on open-ended interactive capabilities and environmental adaptivity. This field accepts that robot intelligence is “situated”: learned during specific interactions with specific surroundings rather than being pre-programmed.

1.3.2 Embodied Cognitive Robotics and its key ideas

EECR rests on seven theoretical principles, consistent with the notion that a robot's body and environment are not mere instruments but active constituents of cognition, generating intelligent behaviour.

1.3.2.1 Presence – Embodiment is integral to how the robot interacts, learns, and responds. It enables natural communication with humans via gestures, eye contact, and spatial positioning. Physical embodiment creates a tangible presence that fosters trust and intuitive interaction (Lyons & Nam 2021). A robot able to grasp can also use “grasping” to recognize similar actions in others (Metta et al. 2006). In collaborative or social tasks, humanoids can indicate directions or make eye contact to show attentiveness, enhancing engagement (Lombardi et al. 2022; Raggioli & Cangelosi 2022).

1.3.2.2 Coupling – Embodied behaviour arises from a continuous feedback loop between perception and action, with each shaping the other (Spivey & Huette 2024; Di Paolo 2003). The environment is sensed as affordances for action, while actions are anticipated as potential perceptual changes. These recurrent sensorimotor loops (Clark 2001b; O'Regan 2001) enable “sense-making” in living organisms (De Jaegher & Di Paolo 2007; Cappuccio & Froese 2014 [SP4]). Even unicellular organisms sense environmental features relevant to

survival, interpreting them as opportunities for action (Hoffmann 2011; Pfeifer & Gómez 2005). Natural and artificial cognizers do not merely discover a pre-existing world: they enact it by discerning sense and value (Cappuccio & Froese 2014 [SP4]).

1.3.2.3 Adaptability – Enactive robots learn from direct interaction rather than relying solely on pre-programmed data (Kirsh & Maglio 1994). Adaptability requires tacit knowledge (Georgeon, Wolf & Gay 2013), key for expert skills yet challenging to synthesise (Richards & Busch 2003). Embodied robots use prior tacit experience to handle new situations, producing emergent behaviours shaped by engagement with users and environments. This allows them to adjust to novel conditions and react fluidly (Tani 2017), enabling humans to predict their actions.

1.3.2.4 Practical intelligence – Environment-embedded intelligence stresses that cognition is shaped by immediate surroundings. Robots use physical context to simplify tasks, leveraging environmental cues (Brooks 1991). The body itself functions as a pragmatic model of possible interactions, predicting effects of actions, and explaining environmental processes (Hoffmann et al. 2010; Nakajima 2013). EECR designs bodies for exploring, experimenting, and learning from open-ended interaction.

1.3.2.5 Autonomy – EECR-based robots possess greater autonomy (Abbas, Scholz & Reid 2018): embodiment enables them to form decision patterns based on environmental attunement and emergent goals. They can develop their own goals from interactive sensorimotor experience (Kirchhoff et al. 2018). Embodiment scaffolds goal formation by mapping space in terms of affordances (Cappuccio 2009 [SP1]; Bruineberg & Rietveld 2014). Unlike pre-programmed robots, they adjust actions to changing conditions, aligning with human goals without constant supervision (Verschure, Kröse & Pfeifer 1992).

1.3.2.6 Empathy – Embodied robots can recognise and enact social behaviours (e.g., approaching, waiting, backing off) in line with expectations (Laitinen 2016), responding to cues, personal space, and non-verbal signals (Guaidello et al. 2016; Nørskov 2017a). Using embodied experience, robots can “empathise” with humans (Asada 2016; Damiano, Dumouchel & Lehmann 2015a)—for instance, recognising focus and waiting before interrupting, or noticing distress and offering help.

1.3.2.7 Cognitive parsimony – Embodied robots adapt to individual users through repeated interactions, optimising personalised support and reducing user cognitive load. By embedding intelligence in the environment and adapting autonomously, they minimise the

need for human supervision (Damiano, Dumouchel & Lehmann 2015b; Dumouchel & Damiano 2017).

1.3.3 Attention and its embodiment

Attention is central to embodied cognition (D'Angelo 2020; Franco 2013). In HRI, a robot's embodiment shapes both users' attentional resources and the robot's own attentional capacities, showing that active and perceived attention are intertwined with key implications for cognitive robotics (Raggioli & Cangelosi 2022).

1.3.3.1 The Embodiment of Attentional Cues: (i) *Embodied Cues as Attention Guides* –

A robot's physical presence directly affects where and how users focus. Robots with expressive faces and gestures use cues such as gaze, orientation, and movements to direct user attention, which is particularly beneficial in collaborative or educational tasks. (ii) *Attention Demands of Physical Presence* – Embodied robots capture more attention than disembodied agents. Their salience enhances engagement and intuitiveness but may also distract if attention is poorly managed (Ferreira & Dias 2014). Effective design requires balancing engagement with cognitive economy. (iii) *Attentional Anchoring* – Embodiment anchors user attention in physical space (Chu & Kita 2011). Distinct bodies and movements help users track robots in dynamic environments, sustaining focus on interaction despite distractions.

1.3.3.2 Embodiment as a Tool for Attention Management: (i) *Physical Gestures and Social Cues* –

Gestures, gaze, and non-verbal signals allow robots to attract, direct, or release attention (Kompatsiari et al. 2021; Shepherd & Cappuccio 2012 [SP11]). Direct gaze signals listening or request for input, while gaze aversion signals disengagement (Desideri et al. 2020; Koller et al. 2023). (ii) *Embodiment and Multimodal Attention* – Embodied robots integrate multiple sensory channels. Combining visual gestures with vocal intonation supports user engagement without overload (Milne et al. 2018).

1.3.3.3 Optimization of Cognitive Load Distribution via Attention Management: (i) *Reducing Cognitive Load Through Embodiment* –

Clear embodied cues lower cognitive demands by directing user focus. In hospitals, for example, a robot may orient toward a patient or tool to guide a nurse's attention (Cappuccio & Shepherd 2013). (ii) *Balance Between Salience and Subtlety* – Robots that demand too much attention risk fatigue, while overly subtle cues may go unnoticed. Effective social robots calibrate embodiment to remain engaging but not distracting (Lanillos, Ferreira & Dias 2014).

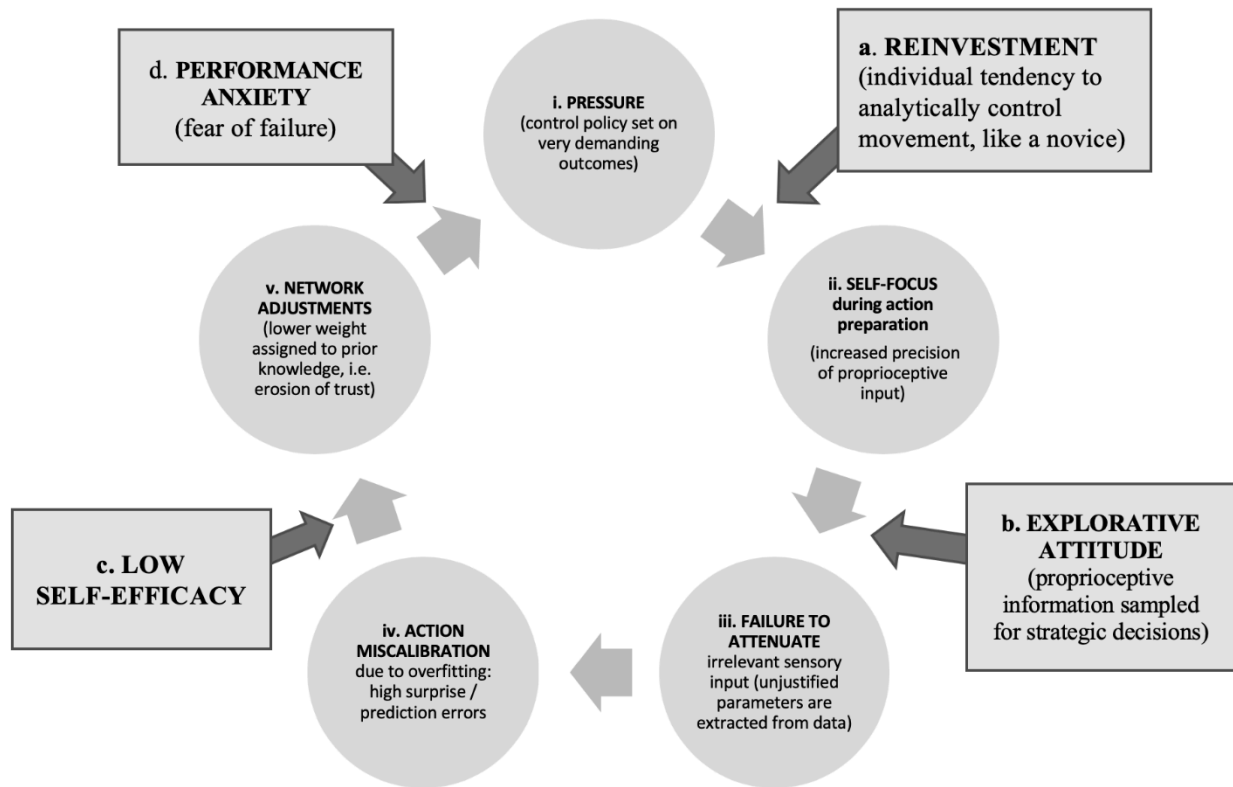
1.3.3.4 Human Attitudes and Relationship Building: (i) *Building Trust Through Embodiment and Attentional Awareness* – Robots perceived as attentive to user limits appear more competent and trustworthy. For instance, a robot that releases attention after a task or avoids interrupting a busy user fosters trust (Chesher & Andreallo 2022; Marchesi & Wykowska 2024). (ii) *Supporting Social Engagement and Empathy* – Embodiment allows robots to recognise attentional shifts and respond accordingly, fostering shared experiences essential for companionship or therapy (Ferreira & Dias 2014). Attentional alignment enhances empathy and perceived engagement. (iii) *Facilitating Natural Social Cues* – Embodied attention supports the rhythm of turn-taking in human communication, enabling smoother exchanges. Robots that embody attentional cues function more as partners than tools (Seibt, Nørskov & Schack Andersen 2016).

1.3.4 A neuro-cognitive model of attention management in skilful activities

EECR is important not only to design robots that effectively interact with humans but also to model human cognitive functions for scientific purposes: for example, a robot that performs as skilfully as a human in a certain domain of human expertise can provide cognitive scientists with very useful indications about the cognitive mechanisms underpinning the control of human action in contexts of human activity based on practical know-how and performative familiarity. Conversely, when a robot underperforms like humans typically do: a robotic model that fails to perform skilfully, exhibiting the same limitations of a human, has a remarkable epistemic function because it mimics the cognitive processes of humans. Similar patterns of failure indicate that the robot's cognitive functions have the same limitations and constraints of the human cognitive functions that it tries to imitate, hence its cognitive architecture represents a prototype of the corresponding architecture functions in humans.

Inspired by these considerations, this section is dedicated to a paper that I co-authored and published with cognitivist scientist Michael Kirchhoff and roboticists Fady Alnajjar and Jun Tani in the *Journal of Consciousness Studies* (Cappuccio et al 2019b [CP2]). In this paper, we propose an account of “choking effect” based on active inference — an approach to PP that is informed both by ECT and by the notion of FEP in cognitive neuroscience. The choking effect (also known as “choke”) is the tendency of expert athletes to underperform in high-stakes situations (Baumeister 1984; Beilock & Gray 2007; Beilock 2011; Cappuccio et al 2019a [SP6]). PP explains the choking effect in terms of precision-modulated imbalance between sensory input and higher-level predictions (Brown, Friston & Bestmann 2011; Clark 2013). It predicts that choke arises when the cognitive system fails to attenuate the error

signal generated by proprioceptive sensory input (Friston 2011). In our paper [CP2], we expand the previous formulations of this model to integrate the contribution of other factors – such as confidence erosion – that have been empirically studied by sport psychologists (Gray & Allsop 2013). Our model uses empirical research on cognitive functions to study those functions through embodied robots designed to perform skilled sensorimotor actions like expert humans. By combining the principles of ECT and PP, our expanded model provides a unified framework capable to accommodate the two main theories of performance disruption in the sport psychology literature: the self-monitoring/execution focus theory (Beilock et al 2008) and the distraction/overload theory (Wine 1971; Christensen, Sutton & McIlwain 2015).



The first theory imputes performance disruption to a *misallocation of attentional resources* and builds on the evidence that experts tends to underperform when they (either deliberately or compulsively) control their action reflectively, monitoring the component processes of action execution, singling out and macro-managing movements that they have already automatised as part of their training (Beilock et al 2002; Beilock & Gray 2007). This theory predicts that self-monitoring is detrimental to performance in competitive environments like tournaments, when experts should rely even more on the automatization of their low-level sensorimotor components of action (that normally sportsmen don't explicitly pay attention to during the performance). The second theory imputes performance disruption to the *depletion*

of *attentional resources*, due for example to distraction or overload (Montero 2016): cognitive resources (typically measured in terms of working memory) tend to be depleted when the mind of athletes is filled with recurrent worries and thoughts of failures, which is likely to happen in pressure-filled environments like tournaments and competitions (MacIntyre & Moran 2007).

Not only does our model allow us to describe both kinds of performance disruptions in terms of precision-modulation imbalance, but it also proves that the distinctive manifestations of choking, a phenomenon that significantly differs from the other forms of performance disruption, are best accounted for in terms of misallocation (not depletion) of attentional resources: choke in sport competition is best accounted for by the self-monitoring/execution focus theory because it involves the experts' apparent incapability to perform simple actions that they normally perform automatically. We use a particular approach to PP, called Active Inference, because it incorporates some key aspects of embodied and enactive cognition (Kirchhoff et al 2018; Parr & Friston 2018). Active Inference allows us to explain some distinctive experiential aspects of choke: it is a skill-level specific phenomenon (Beilock & Gray 2012); it is alleviated by ritual-like pre-performance routines (Schippers & Van Lange 2006); aggravated by contextual factors such as self-confidence erosion and performance anxiety (Beilock & Carr 2001); and accompanied by a diminished attenuation of the sense of agency normally associated with high performance and flow states in experts (Voss et al 2005; Friston 2014)⁷.

Within the broader PP framework, FEP and its associated *Active Inference* theory provide a more specialized, biologically grounded account of how organisms embody predictive processing (Friston 2010; Chen et al 2019; Kirchhoff et al 2018)⁸. Active Inference does not

⁷ PP is a broad theoretical framework in cognitive science based on the key idea that the brain's primary function is to generate predictions about incoming sensory inputs and to continually update these predictions based on observed errors or mismatches (Hohwy 2013). The brain maintains generative models of the environment to predict the sensory input it expects to receive (Hohwy 2016). This framework provides elegant accounts of cognitive functions and their disruption utilizing the principle of prediction error minimization through continuous feedback. By bridging the gap between artificial and natural cognition, PP sees both human and robotic minds as complex "prediction machines" that continuously generate inferences about how the world is going to change and forming inferences about the hidden causes of these changes (Lanillos et al 2021). This activity aims to minimize prediction error and bring the mind's internal model in line with the external world. Discrepancies between the predicted input and the actual sensory input yield prediction errors: these errors drive a process of updating the generative model (the prior beliefs held by the cognitive system) to reduce future mismatches (Parr & Friston 2018)⁷.

⁸ FEP deepens the vision put forward by PP by highlighting the inseparability of perception and action, offering a richly embodied and enactive approach to understanding how minds—be they biological or artificial—continuously create and reshape their worlds to minimize surprise (Tani 2023). Advanced by Karl Friston and colleagues, the FEP

merely address how the brain updates beliefs about the world, but also integrates how organisms *act on* the world to fulfill predictions, thereby minimizing free energy.⁹ As it recognizes the integral role of the body in shaping cognitive processes (unlike purely computational or representational views of cognition), active inference under the FEP is considered an *embodied* account of cognition, in the sense of ECT and 4E cognition. Specifically, it is *enactivist* in suggesting that cognition is something organisms *do*, not just something that happens in their heads (Ramstead, Kirchhoff & Friston 2019; Safron, Hipólito & Clark 2023). Organisms actively participate in bringing forth the world they experience.

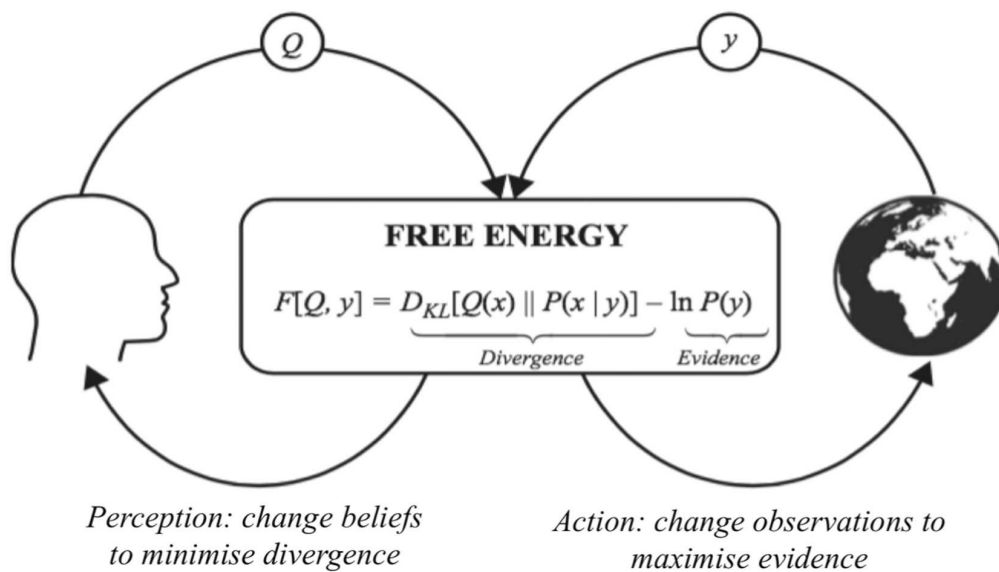


Fig. 4 - Active inference and Free Energy minimization (Pezzulo, Parr & Friston 2024)

This perspective challenges strictly representational accounts by highlighting the fluid interplay between an organism and its environment. Instead of passively representing a

posits that biological systems strive to resist disorder (or reduce “surprise”) by minimizing a quantity known as *free energy*, an information-theoretic measure of entropy that is closely related to prediction error (Friston 2010, 2013). Classical PP emphasizes perception—updating internal predictions in light of sensory information. Active Inference extends this principle by including *action* as a means of reducing prediction error: organisms do not only change their internal models to match the world; they can also *act on the world* to make it align with their predictive models (Adams, Shipp & Friston 2013). Concretely, if an organism predicts a certain sensory outcome, it might move its body or manipulate the environment so that this predicted outcome is realized, thus reducing sensory surprise. Free energy is a formal measure of both informational and energetic entropy that bounds “surprise” (or unexpectedness). In essence, living systems “stay alive” by minimizing this bound on surprise (Friston 2013). Active Inference incorporates the principles of ECT, as both perception and action serve to minimize free energy: perception updates the internal model, while action changes external conditions.

⁹ This view naturally lends itself to embodied and enactivist interpretations, where cognition emerges from the dynamic interaction between an organism and its environment, challenging traditional, disembodied models of representation: through this lens, the system’s predictions are deeply intertwined with the system’s ongoing physical engagement in the world (Cappuccio & Froese 2014 [SP4]).

static world, agents continuously mold the environment to fit their own predictive models (and vice versa)¹⁰. This aligns with enactivist ideas that “knowing” is inseparable from “doing,” and it situates representations as emergent properties in the ongoing loop of action-perception couplings. Our study on performance disruption explains this phenomenon in terms of embodiment of attention.

“In action generation, active inference takes the form of making inferences about counterfactual outcomes (the trajectories of one’s limbs and bodies), specified in terms of proprioceptive consequences. Hence, under active inference, movement comes about by predicting the proprioceptive states that bring such movement about.”

This is very important to understand the FEP mechanisms underpinning choke:

“To allow sensory attenuation in the context of high sport performance (i.e., very accurate self-paced sensorimotor tasks), the agent may need to withdraw attention from their body’s occurrent state and rather focus on the predicted sensory effects of the intended end-state. They need to attend the distal target or expected outcome of one’s action, rather than the movements that compose that action, using an external focus rather than internal (i.e., execution) focus [...]. Choke occurs because, instead of providing the system with additional information useful to correctly calibrate action, over-scrutinizing is effectively the same as not being able to trust one’s predictions about movement induced sensations. [...]”¹¹

1.3.5 Challenges for the robotic implementation of Active Inference

Despite its theoretical appeal, attempts to implement active inference in autonomous and interactive robots have faced significant obstacles. At least four interrelated challenges stand out:

¹⁰ While predictive processing can be interpreted in representational terms—internal models “represent” the external environment—some versions of Active Inference push against traditional notions of representation (Ramstead, Kirchhoff & Friston 2019), especially they model action control and the decisional processes of expert agents to dynamically attune to their environment. Authors working in the Active Inference paradigm instead emphasize *pragmatic* or *action-oriented* notions of “representations,” focusing on how internal states guide interaction rather than encode mirror-like copies of the external world (Tani & Nolfi 2019).

¹¹ Note the correspondence of phenomenological/psychological and cognitive/neuro-functional analyses: “Experientially the agent is attending to her movements too much. Computationally, this kind of over-attending rests on prediction errors being precision-weighted higher than predictions about movement induced sensations, which leads to the agent having to over-sample her movements. [...] Active inference unpacks the claim that over-attending one’s actions as they are executed is detrimental to performance, in terms of a generative model for action that is overfitted because of the disproportionate weight attributed to the proprioceptive sensory input.”

1.3.5.1 Computational Tractability - While variational free energy minimisation is mathematically elegant, its practical computation is often intractable for embodied systems operating in real time. Approximations (e.g., variational message passing, amortised inference) are computationally expensive, and scaling them to high-dimensional sensory input (e.g., vision, multimodal perception) pushes beyond the limits of current embedded hardware. Unlike the brain—which leverages massively parallel, evolved architectures—robotic systems struggle with the sheer resource demands of online Bayesian inference.

1.3.5.2 Model Specification and Complexity - Active inference presupposes the availability of a generative model of the world, specifying how hidden states cause observations. Constructing such models for robots is non-trivial: while humans rely on richly structured, evolutionarily tuned priors, artificial systems must rely on engineered models that are inevitably simplified. Overly abstract models fail to capture ecological detail, while overly complex models render inference unstable or computationally infeasible. This “modelling bottleneck” severely limits the applicability of FEP-inspired robotics.

1.3.5.3 Learning and Adaptation - Machine learning excels by exploiting large datasets to optimise parameters in high-dimensional function approximators (e.g., deep neural networks). Active inference, by contrast, emphasises online adaptation through prediction error minimisation. Although attractive for life-long learning, this strategy is fragile: robots frequently become trapped in poor local minima, or fail to adapt to environments not anticipated by their generative models. This tension between robustness and adaptability remains unresolved.

1.3.5.4 Action Selection and Embodiment - Active inference treats action as a means of fulfilling predictions, yet in robotics, action must also serve pragmatic goals such as task efficiency, safety, and coordination with humans. Mapping prediction error minimisation directly to robotic control often yields behaviours that are either trivial (e.g., minimising uncertainty by avoiding interaction altogether) or maladaptive (e.g., sampling environments in ways that humans find unintuitive or unsafe). The richness of human action emerges from a web of social, cultural, and developmental constraints that current active inference formulations do not capture.

1.3.5.5 Towards a hybrid approach - Why does active inference struggle where machine learning succeeds? Both machine learning and active inference view cognition as *prediction-driven*, but they diverge in practice. Machine learning employs powerful function approximators to detect statistical regularities and optimise prediction accuracy, often at the

cost of interpretability. Active inference, instead, grounds prediction in the optimisation of generative models, aiming for biological plausibility and normative coherence. Yet this principled grounding becomes a liability in artificial systems: the need for explicit generative models, the high computational burden of approximate Bayesian inference, and the fragility of online adaptation all contrast unfavourably with the empirical success of machine learning techniques. But the failure of “pure” active inference implementations does not undermine its theoretical significance. Rather, it suggests the need for *hybrid approaches* where active inference informs high-level decision-making and uncertainty modelling, while machine learning handles perception and low-level control. Such integration mirrors the human case, where predictive processing is supported by neural architectures honed through evolution and experience. For robotics, the challenge is to identify where the normative elegance of active inference adds unique value, and where more pragmatic data-driven approaches remain indispensable.

1.3.6 Outcomes, limitations, and future research

In Active Inference terms, Choke arises from substandard precision-modulation relative to predictions and error signals (Voss et al 2005; Clark 2016; Limanowski & Friston 2018). At the computational level, the cause of this imprecision can be described as an insufficient attenuation of proprioceptive sensory signal which prevents the system from inferring the policies required to predict proprioceptive input to optimize performance. The problem arises because the execution of automated action routines is inappropriately assigned to deliberate, reflective control, not because of a depletion of available cognitive resources (Schneider & Chein 2003). That is why Active Inference that the theory of self-monitoring is a better candidate to explain choking effect, compared to the distraction/overload theory.

A deliberate, reflective form of attention that aims to single out and analytically map the components of an otherwise organic, fluid stream of sensorimotor activity, is informed by explicit representation. The fact that suboptimal performance is caused by cognitive representation indicates that skilled action, in turn, strictly relies on non-representational and pre-reflective or implicit forms of attention (Mullen, Hardy & Oldham 2007; Jackson & Beilock 2008), i.e. forms of attention that are “embodied” (Gray & Cañal-Bruland 2015) in the specific sense that the agent does not analyse action into its constitutive components, but rather track the global aspects of the situation. Experts exert a form of intuitive, embodied control over the circumstances (analogous in many ways to a form of “flow”, see Cappuccio 2017 [SP5]), rather than making a step back to single out this or that element. This importantly

corroborates the 4E approach to cognition and the enactivist approach to PP, because it demonstrates that basic forms of cognition are inherently pre-reflective and contentless.

The concrete implications for HRI are significant. Our model provides important indications for the design of robots that not only use attention but actively adjust their attentional patterns using their body to meet the demands of the task at hand. Our model indicates that, in order to perform like an expert in a skilled sensorimotor activity, a robot must be designed to carefully calibrate its attentional resources: when they are still in the process of learning, collecting relevant data useful to training, the attentional resources available to the robot must be explicit and self-directed (focussing on the action component processes) as they serve the purpose of generating accurate models of the most effective movements (and how the perceived world changes as an effect of those movements); conversely, as it progressively incorporates the training data into a consistent model, becoming ready to mimic an “expert” stance, the same robot must divert its attention from the component processes of action, and concentrate on the outcomes or intended effects of their action, which is supposed to have been already automatized (Limanowski 2017). This selective action does not need to be “programmed” into the robot by means of explicit instructions. In the same way, a human does not need to be told how to direct their attention towards an external focus, as they acquire the necessary expertise in the target sensorimotor activity (Wulf et al 2001): importantly, the learning process itself spontaneously produces this attentional shift from internal (proximal, reflective) to external (distal and outcome-oriented) foci, by reconfiguring the whole cognitive system.

This is corroborated by neuro-cognitive models of humanoid robots that use MTRNNs to learn precise action control and to perform skilfully (Paine & Tani 2004; Yamashita & Tani 2008). These models rely on experts neural networks trained by repeatedly perform a goal-oriented action and preserving the sensorimotor memory of the instances of actions that achieved the given goal: the nodes in the lowest layers of the network (which immediately receive information from the sensory device and are more directly sensitive to mismatches between the input signal and the expected perception) become less and less active (i.e., they perform more parsimonious and simplified computations) as the network becomes more expert, while the nodes in the highest levels of the network (more directly related to the general configuration of intended actions, that is the abstract level of the action’s “concept”) become progressively more active, to reflect how the burden of action control moves from the small details of action execution to the high-level coordination of action as a harmoniously integrated whole (Tani & Nolfi 1999; Bruineberg & Rietveld 2014). This reflects how, during peak performance and habitual action, expert sportspersons don’t focus

too much on the details of movements they have already automatized but rather use their resources to strategize or conduct higher-level considerations: in terms of attentional focus re-directions, expert sportspersons perform rapid, preparatory eye movements to collect only the minimum of sensorimotor information indispensable to calibrate action, instead of performing broad, exploratory or fixating ocular actions, like novices and less experienced sportspersons (Beilock & Gray 2012). From a bio-mimetic perspective, this provides an important indication to correctly design artificial agents that allocate attentional resources in a manner that is consistent with the level of their sensorimotor familiarity with the task.



With my research group in UAE I have personally conducted some exploratory studies consistent with existing FEP model to confirm our hypotheses on skilled action and its disruptions: inspired by similar studies with robots controlled by neural networks (Tani 1999; Paine & Tani 2004, 2005; Yamashita & Tani 2008; Lanillos et al 2021) in our laboratory we used a Nao controlled by MTRNN manually trained by our laboratory assistants to perform golf putting through a long process of opportunely supervised reinforcement learning: this way the robot was trained to successfully complete the task of sending a golf ball into a hole positioned 4 meters away using a golf stick attached to the robot hand. The preliminary results of this study show that, in accord with the results of the studies conducted by Tani and colleagues, the increase in expertise coincides with the spontaneous evolution of the neural network in a way that progressively determines a greater role for the upper layers of the network while increasingly suppressing the prediction error coming from the lower layers. This suggests that expert robots, like expert human performers, tend to rely more on the sensorimotor schemata developed through performative experience (habitual action) than on the scrutiny of the perceptual contingencies, modulating the weight of the perceptual input to collect just the right amount necessary to perform. The study is at the stage of data interpretation and thus it is still unpublished, therefore I mention it here only as an indication of a concrete way forward to apply the theoretical results achieved by our models.

1.4 Conclusions and significance for the field of HRI

This chapter contextualized my research within the relevant disciplines, approaches, and frameworks: social robotics, neuro-cognitive robotics, affective computing, ECT and 4E cognition. It also explained why my research offers an original and useful contribution to the advancement of computer science, expounding the reasons why AI structurally needs embodiment and why, in a sense, embodied robotics is more primordial than -and constitutes a fundamental precondition of- general AI: intelligence cannot arise without embodied experience and physical implementation comes before abstract knowledge representation.

The first part of the chapter has introduced the key notions, the models, and the principles that I have adopted in my work, consistently with the most significant trends in the contemporary literature and the practices used by the communities of researchers I refer to. I have explained that, since my approach to HRI is informed by interaction studies, neuro-cognitive robotics, social robotics, and affective computing, the role of embodiment in robotics is central and represents a major concern and a unifying feature of all my publications.

The second part has investigated in greater detail the notion of embodiment, which is taken for granted and not examined critically by the great majority of roboticists. My analyses have established that, in order to use this notion with the right level of awareness in the field robotics, it is necessary to understand the concept of embodiment from the perspective of embodied cognition theory: this means realizing that the body's role is not only instrumental to, but constitutive of and essential for cognition, as cognitive functions are shaped by the details of their bodily implementation and the concrete modality of body-world interaction. The most significant outcome for HRI is represented by the principle that basic forms of cognition are not representational, but rather based on direct sensorimotor engagement, and the principle that cognition is a circular form of adaptive relationship between an agent and its environment: something an agent does rather than a merely spectatorial or inferential process. The analysis of Turing's seminal paper confirms that this idea can be found at the beginning of the history of AI and that they inform the very project of creating artificial intelligently interactive systems: AI cannot be programmed into a disembodied computer, but it needs to progressively evolve within an embodied robot capable to explore the environment and adapt to it.

The third part has discussed the implications of these ideas in the fields of neuro-cognitive robotics, with a specific focus on the embodiment of attention. The model of skilled performance disruption proposed by me and my co-authors has implications that reach beyond the explanation of human cognition, as it indicates a possible way forward to model skilled attention through robots capable to perform expert sensorimotor actions. Expert attention is not a universal commodity or an abstract capacity (it is not reducible to work memory or information bandwidth), but it is a specific dynamic form of relationship between an intelligent agent and its habitual environment.

A new approach to robotics arises together with the possibility to create such robots based on this principle, i.e. robots that perform like expert sportspersons: their intelligence is determined by their bodily familiarity with the performance environment and their habitual engagement in it, i.e. a sensorimotor engagement that builds on the fine-grained calibration of perception, control, and anticipatory systems, without relying on some detailed and accurate representation of the external world. All the obvious advantages and the possible limitations associated with skilled behaviours deriving from habitual action and repetitive exercise can be replicated and amplified by means of autonomous technologies, using the details of the robot embodiment as a dynamic, adaptive, and predictive vehicle of computational intelligence, rather than a cosmetic or merely implementational feature. In the next two chapters I will explore the implications of embodied attention in the clinical domain (social attention, Chapter 2) and in the educational domain (moralizing or disciplining attention, Chapter 3).

Core Publication #2 **[CP2]**

Cappuccio M.L., Kirchhoff M.D., Alnajjar F.S., Tani J. (2019),
“Unfulfilled Prophecies in Sport Performance. Active Inference and the Choking
Effect” *Journal of Consciousness Studies*, Volume 27, # 3-4, pp. 152-184, Imprint.

Reproduced with permission from Imprint Academic Publishing.

All rights reserved.

Massimiliano L. Cappuccio,^{1,2}
Michael D. Kirchhoff,³
Fady Alnajjar² and Jun Tani⁴

Unfulfilled Prophecies in Sport Performance

Active Inference and the Choking Effect

Abstract: *Choking effect (choke) is the tendency of expert athletes to underperform in high-stakes situations. We propose an account of choke based on active inference — a corollary of the free energy principle in cognitive neuroscience. The active inference scheme can explain certain forms of sensorimotor skills disruption in terms of precision-modulated imbalance between sensory input and higher-level predictions. This model predicts that choke arises when the system fails to attenuate the error signal generated by proprioceptive sensory input. We aim to expand the previous formulations of this model to integrate the contribution of other causal factors, such as*

Correspondence:

Dr. Fady Alnajjar, Director, AI & Robotics Lab, Al Ain, College of Information Technology, United Arab Emirates University, Maqam Campus, Al Ain city, 15551, The United Arab Emirates. Mobile phone: +971-(0)55-7366375.
Email: fady.alnajjar@uaeu.ac.ae

-
- ¹ University of New South Wales Canberra, School of Engineering and Information Technology, Canberra, Australia.
 - ² United Arab Emirates University, AI & Robotics Lab, Al Ain, United Arab Emirates.
 - ³ University of Wollongong, Faculty of Law, Humanity, and the Arts, Wollongong, Australia.
 - ⁴ Okinawa Institute of Science and Technology, Neuro-cognitive robotics lab, Okinawa, Japan.

confidence erosion, taking into account the empirical evidence emerging from the psychological research on performance disruption in sports. Our expanded model allows us to unify the two main theories of performance disruption in the sport psychology literature, i.e. the self-monitoring/execution focus theory and the distraction/overload theory, while recognizing that the typical manifestations of choke in sport competitions are best accounted for by self-monitoring/execution focus theory. We illustrate how active inference explains some experiential aspects of choke that are familiar to sport psychologists and practitioners: choke is a skill-level specific phenomenon; alleviated by ritual-like pre-performance routines; aggravated by personal and contextual factors such as self-confidence erosion and performance anxiety; accompanied by a drop in the attenuation of the sense of agency normally associated with high performance and flow states.

Keywords: sport skill; performance disruption; free energy principle; active inference; choking effect; proprioception; attentional focus; execution focus.

1. Introduction

Choking effect (henceforth, choke) is the tendency of experts to underperform in high-stakes situations — paradigmatically, athletes during tournaments (Beilock, 2011). Choke is just one particular form of performance disruption, i.e. the significant drop in the quality and accuracy of skilled sensorimotor action due to cognitive and motivational factors. Choke represents a concrete concern for sports practitioners and an interesting puzzle for cognitive scientists and performance psychologists, who have proposed opposing explanations of its causes and diverging intervention protocols to mitigate its effects (Cappuccio *et al.*, 2019).

In this paper, we propose an account of choke based on *active inference* — a corollary of the free energy principle (FEP) in cognitive neuroscience (Clark, 2016; Friston, 2010; Ramstead, Kirchhoff and Friston, 2019). The FEP starts from the assumption that for any living system to remain alive it must avoid crossing its terminal phase boundaries. Under the FEP, a system (or agent) can do so only by minimizing the dispersion of its sensory observations, given that it is only observations that are informationally available to an agent. Information-theoretic free energy bounds surprise (i.e. disorder) because the former can be shown to be either greater than or equal to

the latter (Kirchhoff *et al.*, 2018; Parr and Friston, 2017). This means that any agent that works to minimize its free energy necessarily reduces surprise (or statistical uncertainty), where surprise is the deviation between predictions about causes of sensory input and the actual input. Minimizing free energy has also been aptly named *self-evidencing*, since the result of free energy minimizing is better evidence for an organism's predictions of the causes of its sensorium (Hohwy, 2016). Under the FEP, sensory observations are generated by hidden states in the environment (i.e. external causes of sensory outcomes), which, in turn, are causally influenced by action policies the agent pursues (i.e. infers as most likely to minimize the expected free energy associated with future sensory outcomes conditioned on action). This brings the notion of inference (or expectation) generation to the forefront, for active inference crucially implies that for any action to ensue it must be inferred as the most optimal action to pursue (Clark, 2016; Robertson and Kirchhoff, this issue).

Active inference is relevant to sport psychology because it construes skilful performance (e.g. expert sensorimotor control in sport) to consist entirely in inferring predictions that best capture the causal regularities in the (bodily or worldly) environment generating sensory input. Specifically, active inference yields sport psychology with a computational framework by which to understand the generation of bodily movement conditioned on predicting the movements themselves and their resulting sensations (Clark, 2016; Friston, 2011a). Conversely, the same active inference account allows us to model choke in terms of the system's inability to reduce the mismatch between sensory expectations and movements produced to fulfil those expectations. Over short timescales, such as involved in learning, agents become better at exploiting their actions to keep them within expected states (i.e. states with a local free energy — or prediction error — minima). Highly skilled athletes are a case in point. Yet under pressure to produce their best performance, i.e. when very precise predictions are needed to generate context-specific and perfectly fitted actions relative to the task domain, some athletes choke. Active inference provides a way to approach choke, suggesting initially that choke results from an inability to properly fit a model (prediction) to sensory (proprioceptive, exteroceptive) data.

In recent years, active inference has been used to model sensorimotor disruption in terms of precision-modulated imbalance between sensory input and predictions (Adams, Shipp and Friston, 2013; Brown *et al.*, 2013; Clark, 2016, p. 218; Limanowski, 2017). This

work offers the foundational neurocognitive notions necessary to explain why choke arises when athletes are prevented from selectively disattending to their bodies during action execution and preparation. The notion we shall explore here is that choke arises given less precision assigned to sensory predictions conditioned on action, which, in turn, results in excessive self-attention (i.e. in more and more predictions required to settle on an action). In other words, in choke the agent (or better, its computational dynamics) fails to properly attune its predictions to the causes of sensations, leading to less and less precision assigned to the selection of action policies (i.e. predictions about ensuring movement). It is this apparent failure to properly calibrate proprioceptive input during movement with predictions about sensations given movement that leads to choke. In addition to addressing choke, our key proposal builds on and expands this idea, showing that an active inference approach to choke fills a theoretical gap in the sport psychology literature by offering an information-theoretical account that flexibly explains how self-monitoring, in combination with other contingencies frequently encountered in pressure-filled environments, can damage the performance of expert athletes.

This is not the only explanatory advantage offered by active inference. It also makes it possible to adjudicate between rivaling theories of choke when contextualized in the broader psychological debate on performance disruption, of which choke is just a particular case. The debate on performance disruption is dominated by two theories: (a) the self-monitoring/execution focus theory (EFT); and (b) the distraction/cognitive overload theory (DOT). EFT asserts that expert performance is systematically damaged by explicit attention to one's own body while movement is executed or prepared (Masters, 1992; Beilock and Carr, 2001; Wulf, McNevin and Shea, 2001). DOT, on the other hand, states that high performance is primarily damaged by a *lack* of attention to action execution, not the *excess* of it (Wine, 1971; Sarason, 1988; Mullen, Hardy and Oldham, 2007).

Although EFT and DOT are usually taken to offer competing accounts of choke, in principle they are not mutually exclusive explanations of performance disruption. In fact, while they cannot both be right about the causes of choke, each theory seems able to correctly describe only a specific type of performance disruption, and choke is one of the two. EFT correctly identifies the cases of performance disruption that have to do with the *misassignment* of well-honed sensorimotor routines to deliberate control (despite habitual

action control not requiring explicit decision). DOT is better positioned to explain other forms of performance disruption, as it highlights that performance may be damaged via *depletion* of available cognitive resources, a problem that can affect both experts and novices, and not only in skilled sport activities. We will argue that the specific kind of performance disruption experienced by expert athletes during pressure-filled scenarios like sport tournaments is the one satisfactorily accounted for by EFT, not by DOT.

In this paper, after introducing the EFT/DOT distinction (Section 2) and illustrating the general features of the active inference approach to choke (Section 3), we aim to move the field forward by showing how active inference can unify EFT and DOT within an overarching explanatory framework of performance disruption (Section 4). This framework allows us to: account both for diverse manifestations of performance disruption in sports and for the specificity of choke among them (subsection 4.1); recognize that the cognitive mechanism that distinctively underpins choke in expert athletes is misassignment, as per EFT, not depletion, as assumed by DOT (subsection 4.2); explain how the disruptive effect of self-monitoring is modulated by individual and context specificities (subsection 4.3). Also, we discuss how active inference can capture the complex phenomenology of choke, accounting for some key features well-known to practitioners and experimenters in sports (Section 5). We conclude in Section 6.

2. Choke in Sport Psychology

Choke is the paradoxical phenomenon where expert athletes drastically underperform in situations that pressure them to produce their best performance, e.g. during tournaments (Baumeister, 1984). In the paradigmatic choke scenario, an athlete struggles to execute tasks that in normal circumstances they would have considered routine and effortless, e.g. a professional golf player that repeatedly fails a short-distance, uncomplicated putt that they normally have very high chances to complete successfully.

The disruption of the skilled sensorimotor performance of an expert can have numerous causes and manifest itself in different ways. Sport psychology is specifically concerned with the forms of performance disruption whose causes are not merely organic or physiological, but involve emotional, motivational, attentional, and cognitive factors such as distraction, fatigue, extreme arousal/drive, and panic, *inter alia*. Choke is just another variety of performance disruption

imputable to psychological factors, but it is a very peculiar one in that it does not have a specific symptomatic expression or an obvious exogenous cause.⁵ Peculiarly, choke seems to hit harder when the athlete is completely focused on the task at hand and is doing all that is (supposedly) required to achieve optimal results, including filtering out extraneous thoughts and worries (Gray and Allsop, 2013). This is why the substandard results due to choke are unexpected and typically experienced with a mix of disbelief and frustration (Hill and Hemmings, 2015). These negative emotions contribute to aggravate pre-existing tension and insecurity, which, in turn, accentuate the risk of choke.

To explain choke, we need to adjudicate between two leading theories of performance disruption in experimental and theoretical research: the *self-monitoring/execution focus theory* (EFT), and the *distraction/cognitive overload theory* (DOT). Both of these theories link performance disruption to the inappropriate usage of cognitive and attentional resources, but they differ in that they ground performance disruption in either an excess or a lack of attention to the task at hand, respectively (Cappuccio *et al.*, 2019).

EFT claims that skilful performance drops when the fluid motor sequences generated by well-practised, and to a large extent automated, sensorimotor routines are disrupted by obsessive self-monitoring. The assumption is that well-practised motor behaviours are faster, more fluid, and consequently more precise when executed without explicit self-monitoring, as deliberately focusing on the action execution reportedly interferes with the dynamic, smooth integration of fine-tuned and well-coordinated movements that characterize expert, automated action routines (Beilock and Gray, 2007; 2012).

DOT, in turn, states that performance disruption is caused by recurrent thoughts and distracting worries, which fill up the mind of the athlete, preventing them from correctly focusing on the task at hand: in this case, the assumption is that high performance requires high concentration, because the total amount of working memory available to an expert is limited, and has to be allocated among all the tasks simultaneously carried out (Sarason, 1988; Christensen, Sutton and McIlwain, 2015).

⁵ This is one of the reasons to neatly distinguish between choke and ‘the yips’, a neurological condition characterized by patently debilitating symptoms like spasms and tremors.

These two approaches suggest opposite remedies for preventing performance disruption: EFT recommends distraction from overt execution focus; DOT recommends focus to prevent distraction (Jackson and Beilock, 2008). Nonetheless, the debate in sport psychology is polarized around EFT and DOT, and often stresses their points of contrast and conflict. These two theories are often taken as mutually exclusive models of the same specific form of performance disruption, i.e. for the cognitive mechanism responsible for the phenomenon usually referred as choke. Namely, the proponents of DOT assume that choke arises from a *depletion* problem (excessive cognitive task demands on limited cognitive resources), while the proponents of EFT maintain — against DOT — that choke is caused by a *misassignment* problem (explicit control is used for a sensorimotor routine that is better controlled by implicit processes than explicit ones). Hence, DOT assumes that skilful expertise primarily depends on the *quantity* of cognitive processing (working memory usage) done by the system to carry out the task at hand, while for EFT the problem is the *type* (deliberate or automatic processing, *cf.* Schneider and Chein, 2003).

The explanations based on depletion and misassignment support conflicting predictions about performance in dual-task conditions — e.g. when a golf player must put a ball in a golf hole while conducting a secondary cognitive task that engages working memory (Beilock *et al.*, 2002). Depletion assumes that even a perfectly well-trained action can fail if the cognitive burden of all the concurrent tasks combined is excessive, unless the secondary task is precisely about monitoring the primary one to improve its execution. Misassignment, in turn, assumes that the cognitive task can damage the concurrent expert sensorimotor performance only if the former is specifically *about* the latter, i.e. if deliberate, cognitively effortful attention is used for controlling a well-trained embodied task that, in itself, being automated, does not require attentive analysis or deliberate control to be executed properly.

A vast amount of empirical data collected by independent research groups using different experimental settings indicate that misassignment explains much better than depletion why many experts unexpectedly fail in well-trained, self-paced, closed, and typically ballistic sensorimotor tasks that involve a significant habitual component. This evidence includes the following:

1. performance of experts is damaged more by skill-focus dual-task conditions (attention to movement instructions) than split-

- attention dual-task conditions (attention to external stimuli instructions), while the reverse pattern of disruption applies to novices (Beilock and Carr, 2001);
2. performance of experts does not experience the typical time/accuracy trade-off that characterizes the performance of novices (Beilock *et al.*, 2008);
 3. experts experience a trade-off between the quality of their sport performance and the accuracy of their skill-focused judgments about that performance because they tend to pay more attention to their own actions when they are under pressure (Gray, 2014);
 4. experts display characteristic ‘expertise-induced amnesia’: despite showcasing better self-judgment and retrospective task analysis than novices, they have less accurate episodic memory (Beilock and Carr, 2001);
 5. experts often experience an attenuated sense of agency during flow states, suggesting an inverse correlation between peak performance and perceived self-awareness (Swann *et al.*, 2012);
 6. experts perceive the target of their actions with greater detail and accuracy during successful performance (for example, expert golfers report a bigger-than-real size of golf holes) which suggests a correlation between peak performance and an increased focus to the action’s target (external focus) as opposed to oneself (internal/execution focus) (Gray and Cañal-Bruland, 2015);
 7. expert performance is damaged by ‘internal’ (i.e. execution) focus (body part, component movements), not external focus (action effects, strategic intentions) (Wulf, McNevin and Shea, 2001);
 8. experts can generally filter out negative thoughts better than novices, using concentration and visualization techniques, which speaks against DOT’s claim that overload is the primary cause of choke in experts (MacIntyre and Moran, 2007);
 9. experts who tend to choke often find relief in therapeutic approaches based on distraction (Cappuccio *et al.*, 2019); and, finally;
 10. there is no obvious correlation between working memory capacity and skill level, and greater capacity correlates to better performances only in sport tasks that heavily rely on working memory (Beilock and Carr, 2005). This makes it difficult for DOT to explain why choke affects only experts, not novices.

So, while both EFT and DOT can account for at least one form of performance disruption each, experimental data on attention misassignment in sport performance clearly suggest that choke can be accounted for by EFT much better than DOT. If DOT is still very popular it is mainly because it relies on a very familiar neurocognitive model based on the traditional notion of depletion, while many psychologists are still not acquainted with the story used by EFT to explain the neurocognitive underpinnings of the misassignment problem.

EFT claims that what causes choke is explicit attention to the control of expert movements, and active inference casts this in terms of an excessive perceptual sampling of the environment (i.e. attending to bodily movements) combined with a low precision assigned to the active policies that normally elicit the proprioceptive sensations that successfully drive expert action. This idea is *prima facie* in tension with the intuition that more proprioceptive information can only help in making action control more accurate by increasing bodily self-awareness. Why would bodily self-awareness disrupt action control? Dispositional reinvestment theory answers this question by claiming that explicitly monitoring one's movements damages action execution because it defers action control to explicit rules and instructions that the expert had internalized during learning, involuntarily triggering their regression to a novice-like semi-skilled condition (Masters and Maxwell, 2008). Other qualitative accounts appeal to an alleged interference between thought and action (Dreyfus, 2005), subconscious self-sabotage (Jordet, 2010), or 'estrangement from oneself as a lived body' caused by self-consciousness (Limanowski, 2017). All of these accounts capture some important implication of the misassignment mechanism, but they do not exhaustively account for the neurocognitive and information processing mechanisms that explain EFT at the subpersonal level.

We aim to develop such an account via the active inference scheme in this paper. In the next section we will examine this scheme in detail before discussing how it situates itself in the sport psychological debate in Section 4.

3. The Active Inference Model of Choke

In this section we provide an overview of active inference under the *free energy principle* (FEP), before turning to consider choke through the lens of active inference. The FEP states that for organisms to

maintain their integrity they must minimize an information-theoretic quantity known as *free energy*, given that free energy can be shown to be an upper bound on surprise. Free energy minimization is therefore also an expression of Bayesian model optimization, given that reducing surprise is equivalent to garnering evidence for a model or prediction about the bodily or extra-bodily environment. Crucially, active inference captures the idea that Bayesian model optimization need not only occur by updating internal states in light of new evidence (aka perceptual inference). Bayesian model optimization can also take place during action. Active inference provides the FEP with an embodied and enactive perspective (Ramstead, Kirchhoff and Friston, 2019), where action is cast as the process of selectively sampling expected sensory observation conditioned on embodied activity.

Active inference rests upon a generative model of sensory outcomes. A generative model is a probabilistic mapping of how some sensory outcomes might have been generated given prior beliefs about external causes, the probabilistic dependencies between external causes and sensory outcomes, and a likelihood function between external states and sensory outcomes (Parr and Friston, 2018). Generative models serve the function of enabling a system to predict sensory outcomes by inferring the causes that could have elicited its sensory observations.

In the context of generative models, active inference allows an agent to seek out the action policies or routines that yield its expected (future-oriented) sensory observations, which suggests that sensory observations depend on action. This yields active inference with a hierarchical and counterfactual dimension, as it implies inferring over counterfactual sensory outcomes given different action policies. Crucially, given ‘that the generative model enables inference about hidden states based on observations, agents can also form expectations about future states [movements, states in the world, etc.]’ (Schwartenbeck *et al.*, 2018, p. 5).

This is especially important for our purposes, for active inference emphasizes the crucial role of action in the minimization of surprise or prediction error, where the motor system is non-trivially involved in free energy or prediction error minimization.⁶ Unlike in perception,

⁶ Note that we are helping ourselves to a simplification here, treating the notions of free energy minimization and prediction error minimization as (more or less) equivalent.

where prediction error signals are about exteroceptive states, action selection targets predictions and error signals in relation to *proprioception* (i.e. they are about posture and position of the body's joints and forces applied to them) (Adams, Shipp and Friston, 2013; Friston *et al.*, 2012). This captures the following distinction between *perceptual inference* and *active inference*. In perceptual inference, prior beliefs are updated in light of error signals. In active inference, the motor system suppresses error signals by performing the predicted movement (Friston, 2011b; Adams, Shipp and Friston, 2013; Brown *et al.*, 2013).⁷

Within active inference, the exploration–exploitation distinction will be especially relevant as we turn to the phenomenon of choke. This distinction speaks to goal-directed behaviour in the setting of motor control, given that motor control can be associated with finding the right balance between *exploitation* (choosing the most valuable or optimal movements given prior beliefs about the world) and *exploration* (choosing movements that promote foraging and learning about the world). The distinction between exploitation and exploration has been shown to map onto the distinction between *pragmatic* and *epistemic* actions, respectively, enabling agents to exploit specific actions to elicit certain transformative outcomes (e.g. reaching for the golf club), and to explore alternative strategies enabling better decision making (e.g. when a golf player is in the rough having to make decisions about the direction of wind, ground conditions, club selection, and so on). Hence, in exploitative (pragmatic) action, the agent adjusts their movements in accord with their current beliefs; whereas in explorative (epistemic) action, the agent is engaged in sampling sensory outcomes with high surprise or uncertainty in order to update their beliefs, and only subsequently find an optimal or close to optimal sequence of actions given the task at hand (Kirsh and Maglio, 1994). This distinction is relevant to sport psychology because pragmatic actions can be associated with *specialized* action implementation (informing task-specific skilful habitual action), while epistemic actions can be cast as *generalized* action selection strategies (in deliberate and reflective tasks), allowing for sampling of

Nothing in our account hangs on the precise details of the relationship between free energy minimization and prediction error minimization, and we will therefore not labour on this issue any further in this paper.

⁷ Later in this paper we will elaborate on the notion that suppressing prediction error through performing the predicted movement involves a form of *sensory attenuation*.

alternative outcomes. Depending on the task at hand, agents will make use of either pragmatic or epistemic actions, or some mix of both action selection strategies (Chen *et al.*, 2019).

This prompts the question of how an active inference system is able to turn inference into action. The simple answer is that action is achieved by the system predicting or inferring its own motor trajectories (Clark, 2016). In action generation, active inference takes the form of making inferences about counterfactual outcomes (the trajectories of one's limbs and bodies), specified in terms of proprioceptive consequences. Hence, under active inference, movement comes about by predicting the proprioceptive states that bring such movement about. As Clark puts it: 'Such predictions are self-fulfilling prophecies. Expecting the flow of sensation that *would* result *were* you to move your body so as to keep the surfboard in that rolling sweet spot results (if you happen to be an expert surfer) in that very flow, locating the surfboard right where you want it' (*ibid.*, p. 111). Crucially, generating a high volume of precise counterfactual predictions about possible proprioceptive outcomes while reducing the flow of error signals relative to such predictions is paramount to produce accurate skilled actions and fluid bodily movement. This means that movement results from settling on the right balance between weighting sensory predictions and their error signals. As Clark puts it: 'At the limit, errors associated with the higher-level proprioceptive predictions (specifying the desired trajectory) would be accorded a very high weighting, while those associated with current proprioceptive input (specifying the current position of the limb or effector) would be low-weighted' (*ibid.*, p. 216). Differences in how error signals are weighted implies that some signals come to play a functional role on the generation of actions, while others can be confidently ignored by the system.

This is important for our purposes, since to reduce the precision of ascending information and hence initiate skilful action the system needs to reduce proprioceptive and somatosensory prediction error via sensory attenuation. Sensory attenuation is normally associated with self-generated movement, and it is necessary to produce and control intentional behaviour under active (Bayesian) inference. Sensory attenuation can be understood as the attenuation of sensory precision. A failure of sensory attenuation may lead, among other things, to false inference and beliefs about agency (e.g. illusion in schizophrenia, *cf.* Brown *et al.*, 2013).

To allow sensory attenuation in the context of high sport performance (i.e. very accurate self-paced sensorimotor tasks), the agent may need to withdraw attention from their body's occurrent state and rather focus on the predicted sensory effects of the intended end state. They need to attend the distal target or expected outcome of one's action, rather than the movements that compose that action, using an external focus rather than internal (i.e. execution) focus, as per EFT. Choke occurs, when — in the face of having to produce and select very precise action policies — the system cannot reduce prediction error given a lowering of the precision assigned to the relevant action policies. This happens when, compulsively acting as this was the best method to produce very accurate and fast movements, the agent uses execution focus to guide their action. Experienced agents often expect that attending their own movements will increase the precision of their proprioceptive predictions, as they are used to believing that deliberately supervising their body calibration could help maximize action accuracy and efficacy.⁸ This expectation proves wrong in a significant number of cases.

Choke occurs in such cases because, instead of providing the system with additional information useful to correctly calibrate action, over-scrutinizing (i.e. generating more sensory samples than usually required to perform the action) is effectively the same as not being able to trust one's predictions about movement induced sensations. This brings out an interesting way to approach what happens at the experiential and computational scales. Experientially the agent is attending to her movements too much. Computationally, this kind of over-attending rests on prediction errors being precision-weighted higher than predictions about movement induced sensations, which leads to the agent having to over-sample her movements.

Due to such imbalance, the predictions generated at the higher levels, which encode the general expertise necessary to accurately shape skilful actions, are not precise enough to cancel out the errors ascending from sensory input, and moreover the sensory stream is not precise enough to update the high-level beliefs. 'Consequently, no sufficiently precise proprioceptive prediction errors are generated, and no (or abnormal) movement results' (Adams, Shipp and Friston,

⁸ Later (subsection 4.3) we will clarify that self-monitoring does not necessarily lead to performance disruption, but it does so when it occurs in combination with other circumstances, for example when it is performed exploratively rather than exploitatively.

2013). In other words, paying attention to one's own movement, or to their own body during the delicate process of action preparation (in which predictions about the effects of different possible action courses are generated and compared), is likely to hinder — rather than facilitate — the calibration of skilful action because explicitly attending to the sensory input generated by one's own body increases the weight of the ascending prediction errors, which contradict 'the descending predictions of the sensory consequences generated by movement. These errors have now more influence on higher-level beliefs, which are therefore adjusted to accommodate the fact that no movement is sensed' (*ibid.*).

A different way of putting this would be in terms of the distinction between exploitative and explorative actions. Expert performers are able to exploit specific action policies to elicit expected proprioceptive consequences. Through learning, the distribution of possible action policies is optimized by pruning away suboptimal policies. This suggests that, for experts, certain action policies when calculated against expected proprioceptive outcomes have a sharp probability distribution, raising the probability of very specific actions being selected given the task environment. From this perspective, choke is effectively the result of a flattening of the probability distribution over which action policies are inferred, forcing the agent to entertain multiple policies with high surprise — in other words, forcing the agent to sample alternative outcomes conflicts with their expectations given a history of performance. Choke, under active inference, can therefore be cast as suboptimal precision-modulation between predictions and error signals.

There are aspects of the active inference model of performance disruption that even its proponents have not explicitly discussed: firstly, previous accounts have described only choke, without comparing it with other (only apparently similar) types of skilful performance disruption; secondly, they appeal only to EFT, not DOT, to explain performance disruption; thirdly, the account of choke based on EFT assumes that self-monitoring and misassignment always appear together. The next section aims to examine in detail these three aspects.

4. Situating Active Inference in the Sport Psychology Debate

To link the active inference model to the sport psychology debate, we need to answer the following questions. Does the model account for the cases of performance disruption illustrated by EFT and DOT? (4.1); does it rely on the depletion or the misassignment mechanism, and how does it differentiate among them? (4.2); and does choke systematically result from self-monitoring? (4.3).

4.1. EFT vs. DOT

Active inference allows us to distinguish between choke and other forms of performance disruption, including the forms that DOT attributes to overload and distraction. Also, the model ultimately confirms that EFT can account for choke much better than DOT.

Active inference unpacks the EFT explanation of choke, i.e. the claim that over-attending one's actions as they are executed is detrimental to performance, in terms of a generative model for action that is overfitted because of the disproportionate weight attributed to the proprioceptive sensory input. An overfitted, or over-predictive, model is a model that captures 'not only the variance due to the variables of interest but also that from random error, which organisms are likely to encounter in an uncertain world' (Marewski, Gaissmaier and Gigerenzer, 2010, p. 106). Under active inference, the model is overfitted when it extracts from the proprioceptive sensory signal more parameters than could be justified by the data because of its inability to properly filter out all the irrelevant information (noise). This increases the discrepancies between the predicted and the perceived trend in proprioceptive sensory data to the point that the system becomes unable to plan the action that would reduce such discrepancies, resulting in no or suboptimal action (or in inflexible, stereotyped behaviour, as in autistic subjects, *cf.* Idei *et al.*, 2018).

Conversely, a statistical model is underfitted if it cannot accurately capture the underlying trend of the data because it ignores some of the relevant input, essentially treating it as noise. Underfitting can account for some of the cases of performance disruption considered by DOT, namely when poor self-attention reduces the ability to extract sufficient parameters from the proprioceptive sensory signal, which results in no or suboptimal action. In so far as poor attention is often caused by distraction and cognitive overload, also DOT can be accommodated in the active inference scheme, but in a role that has

nothing to do with choke. Thus, cognitive overload and distraction may prevent the system from making proper use of the sensory input signal, whose precision is insufficient to provide a guide for action. This can happen when the circumstances are unexpectedly complex, challenging, and unfamiliar: for example, in scenarios characterized by high uncertainty due to the insufficiency or irrelevance of the available prior knowledge. Navigating such scenarios require computationally demanding decisional processes that are not necessarily eased by sensorimotor expertise. In the tasks in which analysis of proprioceptive sensory data is crucial to be successful, thus significant demands on working memory may lead to troubled or substandard performance.

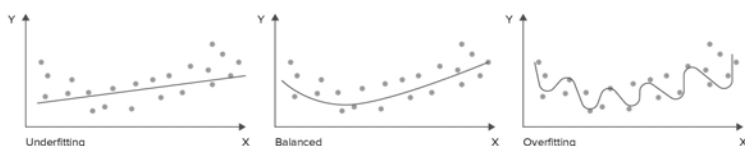


Figure 1. In the context of linear regression, a model generated by machine learning is considered underfitted, robust, or overfitted, depending on whether the number of parameters extracted from data is too small, adequate, or too big.

However, as argued in Section 2, this is not what normally happens to athletes suffering from choke during competitions. The frustration experienced by a choking athlete is caused by their temporary inability to produce routine, habitualized actions in familiar sensorimotor contexts that are not taxing for working memory. The kind of under-performance based on EFT can only be explained by unattenuated proprioceptive input and overfitting, which is what typically happens to expert athletes when task pressure and performance anxiety compel them to obsessively focus on the details of the motor components of the action they are about to execute.

EFT and DOT identify different cases of performance disruption; they also offer different cognitive mechanisms underpinning action control. This is important to answer the second question; namely, does choke arise given depletion or misassignment mechanisms, and can active inference help us differentiate among them?

4.2. *Misassignment vs. depletion*

Active inference supports an explanation of choke in terms of misassignment, rather than resulting from issues around depletion. Performance disruption characterized as choke is not caused by an overflow of the available cognitive capacity caused by too demanding computational tasks, but by the improper precision modulation of sensory signals. Interestingly, when the cause of substandard performance is self-monitoring, distraction can be an effective intervention to improve performance, while reducing the cognitive load in order to free up working memory is not necessarily beneficial (Buszard, Masters and Farrow, 2017). Thus, if proprioception plays a role in paralysing the system this is not because ascending signals carry too much information *per se*, but because top-down predictions do not capture all the counterfactual trajectories of bodily movement implied by such information. Choke is not caused by the depletion of working memory, as claimed by DOT, because the problem is not the *quantity* of information *per se*, but the *type* of sensory information (proprioceptive) utilized to calibrate action control. Crucially, and as we have been arguing, active inference not only casts choke in terms of first-order predictions about sensorimotor contingencies; in addition, it adds the notion that choke occurs given the failure to lower the precision assigned to prediction error.

The active inference model of choke challenges a centralized view on expert performance and skilful action. This view presupposes the notion — borrowed from traditional computer science — that cognitive performance directly correlates to working memory availability: performance decreases when the system is burdened with heavy computational tasks and increases when action execution is attentively monitored, as self-observation provides proprioceptive information relevant to online control. Also, this view assumes that skilful embodied action control, not differently from logico-symbolic and propositional tasks, depends on the system's computational power (i.e. the data processing rate). The active inference model challenges these assumptions because it offers an alternative explanation of how performance disruption can be caused by mechanisms different from computational overload. This suggests that expert skilful control of sensorimotor action has less to do with brute computational power and much more with online sensorimotor attunement, coupling agents, and environment dynamics.

The aforementioned experimental results on performance in dual-task conditions suggest that choke occurs when misassignment (not depletion) ensues. This confirms the hypothesis of an at least partial functional segregation of skilful action control from the processes consuming working memory, which corroborates EFT and challenges DOT, proving that the problem behind choke is not the quantity of cognitive resources used for a sport task, but the type.

4.3. Choke's specificity

The active inference model imputes choke to self-monitoring because self-monitoring (action control guided by internal or execution focus) prevents the required attenuation of proprioceptive and somatosensory input. We need to clarify how deep and systematic the relation between self-monitoring and performance disruption really is. We have already remarked that EFT is better equipped to explain choke than rival theories, although not all forms of performance disruption are caused by self-monitoring. Now we must remark that EFT is on the right track even though not all instances of self-monitoring result in performance disruption. In fact, the cases of athletes chronically prone to choke are arguably less common than the athletes occasionally indulging in self-monitoring to guide their actions. In laboratorial settings, the negative impact of self-monitoring on expert performance is statistically significant but not as debilitating as choke in real-life professional competitions (Christensen, Sutton and McIlwain, 2015). On the sporting field, athletes are often tempted to monitor their own posture and movement to guide their performance, and this does not always provoke catastrophic task disruption. Some sport tasks (batting in cricket, swimming) and artistic performance (classical ballet) rely explicitly on self-monitoring to be carried out successfully (Sutton, 2007; Montero, 2016).

This does not mean that execution focus cannot produce chronic (long-term and severe) instances of choke, but that its disruptive potential is unleashed only in particular circumstances and in specific individuals. The disruptive impact of self-monitoring on sensorimotor control is modulated by multiple person-level variables that, combined with self-monitoring or triggered by it, cause severe performance disruption. These variables impact on expert athletes as they prepare to act with the intent to produce the best possible sensorimotor outcomes. What are these variables, and how should the active inference model be expanded to accommodate them?

To answer this question we need to distinguish between the constitutive elements of active inference leading to performance disruption and the intervening variables that aggravate it. The most detrimental effects of choke are generated by the compulsive iteration of a cognitive-behavioural loop in which imprecise predictions and excessive proprioceptive input reinforce one another. The *constitutive elements* of this loop are subpersonal mechanisms involved in free energy (or prediction error) minimization. The *intervening factors* are personal-level dynamics that may or may not intervene to aggravate the subpersonal dynamics: they are contingent upon individual idiosyncrasies related to psychological background, training methods, context, and so on. Without these factors, the consequences generated by self-monitoring/execution focus may not be too severe for the athlete's performance. The combination of constitutive elements and intervening factors explains why choke is determined by both cognitive mechanisms underlying human action control and peculiar circumstances that depend on the psychological or attitudinal specificities of the performer or the context of the performance.

First, consider the five constitutive elements (subpersonal mechanisms) of choke:

- i) The *heavy pressure* that producing the best possible outcomes characterizes. This is not an emotional characterization, but an initial task-constraint that identifies specific policies to regulate sensorimotor control in accord with the cognitive demands implied by commitment to excellence.
- ii) *Self-monitoring/execution focus* is applied during the movement preparation/calibration stage. It results in an increased bandwidth and sampling rate of the proprioceptive signal and an even higher weighting of this input in the precision-calibration system.
- iii) The system, requested to process an increasing amount of *non-attenuated irrelevant proprioceptive prediction error*, is overfitted as it extracts from data more parameters than would ever be needed or justified from data.
- iv) The overall level of systemic *surprise (disorder, noise)* increases due to the high number of sensory prediction errors. As a consequence, the attempts to calibrate action fail, and spurred movements or no movements ensue.
- v) The system recalibrates to account for the lack of success of the latest inferences produced by the system. The *weight assigned to prior beliefs* is weakened to reflect the system's decreasing trust

in its own predictions in consideration of its failed attempts to skilfully regulate action. Precision weighting of the following descending prediction will decrease accordingly.

Then, consider the following four intervening factors (person-level individual and contextual specificities):

- a) *Reinvestment tendencies* intervene, reinforcing the causal link that connect i and ii. This is the individual predisposition to analytically control action execution, decomposing movement into its basic steps on the basis of rules and instructions, like a novice. Reinvestment has to do with the athlete's background, as it is more likely to occur to the athletes who were trained with explicit instructions and rules.
- b) *Explorative approach* applied to proprioception, intervening between ii and iii. This factor describes how the sensory information gathered through self-monitoring/execution focus is used to fulfil an epistemic function rather than a pragmatic one (see later).
- c) *Low self-efficacy* intervenes primarily between iv and v, modulating how the system's priors are updated on the basis of failure. This is determined by the lack of confidence in one's abilities or chances of success, which may depend on emotional or motivational fragility, or inaccurate self-evaluation. It undermines the validity of the predictions about one's success.
- d) *Performance anxiety* intervenes primarily between v and i to modulate the pressure related to the situation. This emotional/motivational factor accounts for recurrent worries of failure and the importance individually attributed to the competition. Like low self-efficacy, it is modulated by negative emotions and depends on personality.

When combined, these dynamics generate an iterative process that, reinforcing itself through time, eventually leads to the meltdown of the prediction-action generation system. Schematically:

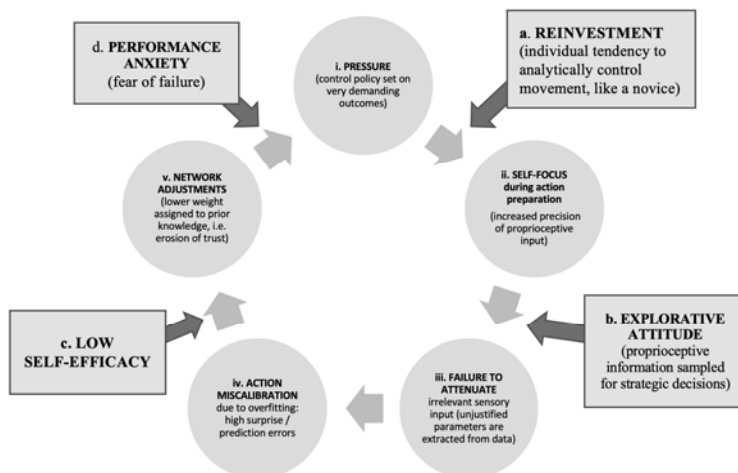


Figure 2. The circular process that iteratively leads to increasing precision-weighting miscalibration involves five structural components as part of the active inference model and four intervening factors related to psychological or contextual circumstances.

The exploitation/exploration distinction, previously introduced, needs to be further discussed to characterize factor b (explorative approach). Normally, well-trained skilled actions involve an automated component, and allow for better performances when action is based on habitual, pre-reflective embodied expertise. On the contrary, inexpert action, or skilled action in problematic and uncertain conditions, produces better performance when it is controlled by deliberate attention. The difference is that expert action in familiar scenarios has fewer degrees of freedom, as it tends to approximate an already familiar optimal model, while inexpert or uncertain action requires more degrees of freedom and more complex decisions to infer the best model. That is why expert action is successful only when accompanied by an exploitation policy, while the exploration policy facilitates inexpert or uncertain action. Sampling the perceptual environment to collect proprioceptive sensory information relevant to action execution is not detrimental *per se*, but it is detrimental when expert actions that should rely on an exploitative engagement with the environment are instead guided by an explorative approach. Exploration is key when action requires explicit decisions and strategic planning, but it is counterproductive when the policies governing an habitual behaviour are already confidently defined by expertise.

The misassignment problem associated with EFT occurs when the system samples the perceptual environment for the wrong purpose, as this causes an unnecessary inflation of the ascending proprioceptive prediction errors that, in turn, hinders the efficacious usage of the embodied expertise embedded in automatic sensorimotor routines. Choke occurs when an action that is supposed to be controlled by the exploitative policy (as it is well-trained and habitual) is rather assigned to an explorative policy, with the agent exploring its own body (sampling proprioceptive information) as if this could help improve movement calibration.

Choke is governed by an incremental, compulsive dynamic. Its disruptive trajectory is determined by the fact that, despite the high uncertainty scenario, the system is still coerced to produce optimal performances by the commitment to excellence that characterizes competitive sports. The context itself instils perfectionism in the athlete while preventing them from withdrawing from the task, despite the increasing levels of perceived risk and frustration. Prediction error peaks when the error tolerance is so low that fulfilling one's expectations is beyond their control capability. Increasing levels of surprise reinforce the urge to scrutinize even more closely their own body posture (joint angles, arm and leg positions, etc.) with higher-rate sampling. Self-observation, in turn, generates greater expectations to produce even more precise and detailed predictions about one's own actions. Action control meltdown occurs when the system eventually loses any capability to fulfil expectations that were hugely inflated despite trust in its own predictive ability dramatically decreasing.

Can this spiral be broken? Two theoretical options are in principle available to reduce the effects of prediction error and block the choke loop: (a) forcefully inducing proprioceptive sensory attenuation; or (b) artificially inflating confidence in the descending predictions (manipulating how precision is perceived/assessed). Both come with risks. Sensory attenuation can be forced by external focus or distractors (e.g. dual-task conditions) that compete with self-monitoring and decrease the bandwidth of proprioceptive information processed online. This has the effect of reducing the sensorial grain and sampling rate of haptic and visual inspections that target joint position, muscular tension, nociceptive signals, etc. Confidence inflation can be induced artificially (neuro-stimulation techniques, drugs) but it can be accompanied by behavioural rigidity (reduced adaptivity to changing contexts) or perceptual illusions (Brown *et al.*, 2013; Clark, 2016, p. 219).

To sum up, choke is control paralysis generated by frustrated perfectionism coupled with incremental erosion of trust in one's expert sensorimotor predictions. The causal dynamics underlying performance disruption are neither simply emotional (anxiety) nor merely attentional (execution focus), but a combination of the two, in addition to a wrong approach to the task (explorative rather than exploitative): as the gap between anticipated and ideal circumstances increases, the expert athlete forces themselves to generate prophecies that cannot be fulfilled just because they are too accurate and demanding.

5. Accounting for Choke's Phenomenology

The active inference model of choke must be further articulated and enriched to account for the complex phenomenology of performance disruption as it is experienced by athletes. To update the active inference scheme on the basis of the experiential reports on choke, we will address the following three aspects: choke is a skill-level specific phenomenon (5.1); choke is alleviated by self-regulatory behaviours and exacerbated by negative beliefs about oneself (5.2); and choke is characterized by the suspension of two phenomena that typically accompany peak performance in experts: expertise-induced amnesia and sense of agency suppression (5.3).

5.1. Experts vs. novices

We have already mentioned that EFT, unlike DOT, construes choke as a phenomenon that affects only experts, not novices. Active inference allows us to link this notion with the idea that skill level in sport correlates to anticipatory abilities. The predictive abilities of expert athletes are more far-reaching and fine-grained than those of novices because their learning history enables them to identify, anticipate, and respond to familiar scenarios. Skilled rugby players, for example, can anticipate seemingly random ball-bounce more accurately than less-skilled counterparts, and skilled performers are able to utilize postural cues from the kicker to predict bounce outcome (Runswick, Green and North, 2019). Expert basketball players are capable of predicting the outcome of a free-throw, of which they observed only the initial movements, faster and more accurately than novices (Aglioti *et al.*, 2008). Motor circuits of expert athletes are selectively recruited to simulate the goal-directed actions observed in others, with activation patterns congruent to the performative expertise of the observer,

testifying that motoric familiarity facilitates the prediction of sensori-motor outcomes (Calvo-Merino *et al.*, 2006).

Anticipatory abilities are crucial to dynamically calibrate skilful action in fast changing environments. Eye tracking studies have documented that the attentional focus of expert table tennis players anticipates the trajectory of the ball during rapid exchanges, rather than fixating on the ball, and that the reach of such anticipation correlates to the player's skill level (Koedijker and Mann, 2015). The predictive expertise of skilled athletes shows in what Christensen and Bicknell (2019) call 'anticipatory control': expert players efficiently scan the environment to identify opportunities for action and viable adaptive strategies and, moreover, they do so dynamically and predictively, anticipating imminent changes in the landscape of opportunities for action, estimating the effects of their own actions in such landscapes, and predicting counterfactual sensory outcomes based on dependencies between sensory and motor dynamics. This allows them to regulate their conduct in a dynamic landscape of affordances, adapting not only to immediate fluctuations but also future-oriented ones.

Active inference conceptualizes the anticipatory control of expert athletes as the ability to competently navigate their niche to reduce surprise. This ability comes with a potential risk: with a heightened sense of anticipatory control inevitably come stronger expectations about the effects of one's own actions, which, in turn, require more effort to correct the selection of policies in adverse and unfamiliar circumstances.

Expertise requires trust in one's own well-trained motor habits, as expert performance largely depends on automated skilful routines that do not involve attentive supervision or explicit decision. However, confidence breakdown is both possible and likely to happen when pressure-filled environments are specifically designed to test one's ability while implicitly demanding the maximum cognitive and physical effort to achieve optimal results. When the tension reaches its peak, the dynamic integration of affordance perception and outcome prediction that is the root of anticipatory control risks collapsing.

Performance disruption due to tool change is an interesting example of how anticipatory control can collapse. It is well documented that several top golf players have suffered serious difficulties in their professional careers after having changed their equipment due to contractual obligations with a new technical sponsor (Dudurich, 2014): e.g. Curtis Strange in 1989 (switching to Maruman), Nick Price in

1995 (Atrigon), Lee Janzen in 1993 (Ben Hogan), Corey Pavin in 1995 (PRGR), and in David Duval in 2001 (Nike). As the same equipment was used successfully by other top players, we can infer that performance disruption depends not only on objective changes in the environment, but also on the subjective unreadiness to adjust to those changes. Flexibility to tool change may vary significantly and arguably correlates to skill-level, as novices do not seem to experience the same decline in performance when shifting golf equipment.

Professional players tend to develop an intimate familiarity with their favourite pieces of equipment, consequently may feel more the tension between expectations, based on previous experience, and actual tool use. Expert golf players produce much more detailed reports about the qualities of their familiar tools than novices, which indicates that experts perceive very precisely whether their tool-use experience is comfortable and confident or not (Roberts *et al.*, 2001).⁹ Behavioural rigidity reflects the expert athlete's difficulty to update the expectations formed during skill development. Also, as highlighted by reinvestment theory, it largely depends on the training methods used during the earlier stages of their career (Masters and Maxwell, 2008). The fact that sensitivity to equipment-change correlates to skill level confirms the intuition that sport expertise correlates both to better predictive/anticipatory abilities and to a greater fragility due to possible miscalibration and unbalances.

5.2. *Pre-performance rituals*

Another interesting aspect of the phenomenology of choke is the ritual-like routines that professional athletes often repeat before performance to avert performance disruption. Apotropaic beliefs are often associated to these rituals. Such superstitious beliefs can affect performance, as evidenced by a study showing that participants performed better on a golf putting task when using a ball that was said to be 'lucky' (Schippers and Van Lange, 2006). Pre-performance routines generally are more useful to the athletes who believe that the ritual is to bring luck (Damisch, Stoberock and Mussweiler, 2010). In turn, it is well-known that recurrent thoughts of failure and feelings of

⁹ These qualitative reports were assessed along 10 dimensions: feel from impact; impact sound; shaft feel; club weight; club control; feel of club position during swing; grip; ball flight; club appearance; golfer's psychology, which includes associations to experiences unrelated to sport performance, for example superstitious associations.

persecution, like those entertained by superstitious people who believe themselves to be victim of a bad luck, have a negative effect on performance. The relationship between negative thoughts and performance disruption is easily explained by DOT in terms of a self-fulfilling prophecy: fear of performing poorly generates recurrent thoughts and worries that deplete the athlete's cognitive resources, causing them to actually perform poorly. But why would the belief in one's own *good* luck *increase* the actual chances of success, given that higher self-confidence cannot increase the maximum amount of available working memory? Without appealing to magic or DOT, active inference can explain why pre-performance rituals improve performance, and why this effect can be even stronger when superstitious beliefs are attached to them.

Pre-performance routines that professional athletes repeat before important challenges represent a method to reduce the perceived randomness of these events, internalizing the feeling that the world is well ordered and predictable. Pre-performance routines cannot bring about the desired outcomes, but can generate very familiar, easily replicable action patterns, whose sensory effects are well-known and expected by the expert athlete. Routine behaviour has a tranquilizing effect, as it defuses the spiral of anxiety and disappointment that typically troubles the performance of people who lack confidence in the own ability. How does it work?

As illustrated by our choke schema, loss of confidence (c) plays a role in the loop that leads to catastrophic performance disruption. Such loss erodes the system's capability to trust its own high-level predictions. Pre-performance rituals help the athletes prevent such loss of confidence by reducing surprise (iv).

By performing ritualized actions whose sensorimotor consequences are very easily predicted, athletes reduce the mismatch between predicted and perceived experience, reinforcing the system's trust in its own anticipatory abilities. A similar effect is achieved by fidgeting, an unintentional, self-regulatory behaviour that, while generated with no apparent reason, contributes to reduce the surprise associated with perceived randomness and uncertainty (Perrykkad and Hohwy, 2020).

Pre-performance rituals and fidgeting behaviour use only sensorimotor dynamics and work independently of any superstition belief — for example the belief that two causally unrelated events can affect one another at a distance. Beliefs in luck imply the perception that desirable events occur more often than non-desirable ones. Such beliefs reinforce trust in one's own predictive abilities by supporting

the conjecture that hidden regularities are behind the occurrence of apparently random events. Bleak and Frederick (1998) have shown that those who believe less in chance are more likely to engage in superstitious behaviours because they have more trust in their own abilities to predict the transformative effects of such behaviours. Like pre-performance routines, also superstitious beliefs can help the system prevent the deterioration of trust in its priors in the face of disruptive events that threaten self-confidence.

5.3. Attenuated sense of agency during 'flow'

Expert athletes often report the feeling of being passively conducted when they perform skilfully. Various modalities of attenuation of self-generated sensory signals can be observed during movement (Blakemore, Wolpert and Frith, 1998; Brown, Friston and Bestmann, 2011) and movement preparation. Skilful sport performance based on automated motor routines has its own pre-reflective 'auto-pilot' modality, often associated with a state of deep absorption called 'flow'. Swann *et al.* (2012) have documented the tendency of expert athletes to describe their role in flow as a passive one, their action control being accounted for in terms of 'letting it happen' rather than 'making it happen'. In a flow state the friction between agent and world seemingly disappears (Cappuccio, 2017). Similar statements, emphasizing passivity over agency in the generation of quick, precise, skilful action are common among expert athletes (e.g. Bruce Lee's famous claim: 'I do no hit. It [my fist] hits by itself'). The athlete witnesses flow condition as a smooth and uninterrupted flux of continuously changing sensorimotor contingencies to which their movements automatically attune.

According to forward model theories, the attenuated sense of agency is an effect of optimal execution, which comes with the complete fulfilment of sensorimotor expectations as action flawlessly approximates its expected end state (Wolpert, Ghahramani and Jordan, 1995; Kawato, 1999). The experience of performing a voluntary action is made possible by the mismatch between predicted and perceived sensory effects as action is generated, therefore a nearly perfect cancellation of prediction error diminishes the subjective feeling of performing voluntarily (Tani, 2017, pp. 235–7). However, according to active inference, the extreme attenuation of proprioceptive sensory input does not only accompany skilled performance, it also enables it: sensory attenuation is instrumental to sustain action prediction

minimizing the overwhelming weight of the errors generated by the robust proprioceptive feedback that result from expert movements. As Limanowski (2017) clarifies: ‘the “experiential absence” of the body is necessary for action in the world... attention directed towards the objective body is detrimental to skillful action, i.e. automatized movement controlled by strongly predicted sensory consequences’.

Therefore, sensory attenuation is not only an epiphenomenon of the forward models that ‘predict and thus cancel out the sensory consequences of one’s movements based on the body’s current state and corollary discharge’ (Blakemore, Wolpert and Frith, 1998; *cf.* Friston, Thornton and Clark, 2012, for a more detailed comparison of these accounts): sensory attenuation is also a necessary condition to bring about the needed ‘dis-attention away from sensory input, which would otherwise bias perceptual inference and potentially preclude movement’ (Limanowski, 2017). Arguably, choke disrupts skilful performance by making this condition impossible to achieve.

6. Conclusion

In reviewing the active inference model of choke we have discussed its position within the current psychological debate on performance disruption. In active inference terms, choke arises from imprecise precision-modulation relative to predictions and error signals. At the computational level, the cause of this imprecision can be described as an insufficient attenuation of proprioceptive sensory signals which prevents the system from inferring the policies required to predict proprioceptive input to optimize performance. The problem arises because the execution of automated action routines is inappropriately assigned to deliberate, attentive control, not because of the depletion of available cognitive resources.

At the behavioural level, the phenomenon of choke experienced by expert athletes is caused by execution focus; that is, excessive attention to the component movements of well-trained sensorimotor action routines during their execution. The explanation of choke based on self-monitoring/execution focus is best captured by EFT. Other kinds of performance disruption are efficaciously accounted for by DOT. Here we have argued that EFT and DOT can both be subsumed under active inference. Moreover, in our expanded model of the causal underpinnings of choke we have identified five structural components (pressure; volume of proprioceptive input; overfitting; sensory prediction error; trust) that are iteratively reinforced as the processes

underlying choke worsen. We distinguish these components from four intervening variables (reinvestment; explorative approach; self-efficacy; anxiety) that may make the effects of choke more or less severe, modulating the underlying cognitive mechanisms. Execution focus becomes particularly detrimental to automated action when self-monitoring is conducted exploratively (to make a decision about the movement to be performed), rather than exploitatively (to calibrate a familiar movement).

We have also examined how active inference accounts for certain experiential aspects of choke that are familiar to scholars and practitioners. Choke is a skill-level specific phenomenon because the risk of misassigning action control arises only when sensorimotor routines are already automated through expertise. This is when the predictive abilities become more sophisticated but also potentially fragile and more prone to overfitting, as demonstrated by expert athletes negatively affected by tool change. Also, choke is attenuated by ritual-like pre-performance routines that have a self-regulatory function and preserve trust in one's own predictive abilities, shielding skilful action control from the disturbing effects of wrong precision modulation. Last but not least, we have emphasized that the attenuation of the sense of agency is an effect associated with peak performance, which explains why selectively disattending one's own body is key to prevent choke in expert sport action.

References

- Aglioti, S.M., Cesari, P., Romani, M. & Urgesi, C. (2008) Action anticipation and motor resonance in elite basketball players, *Nature Neuroscience*, **11** (9), pp. 1109–1116.
- Adams, R.A., Shipp, S. & Friston, K.J. (2013) Predictions not commands: Active inference in the motor system, *Brain Structure and Function*, **218** (3), pp. 611–643.
- Baumeister, R.F. (1984) Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance, *Journal of Personality and Social Psychology*, **46** (3), pp. 610–620.
- Beilock, S.L. (2011) *Choke*, Boston, MA: Little, Brown Book Group.
- Beilock, S.L. & Carr, T.H. (2001) On the fragility of skilled performance: What governs choking under pressure?, *Journal of Experimental Psychology: General*, **130** (4), pp. 701–725.
- Beilock, S.L., Carr, T.H., MacMahon, C. & Starkes, J.L. (2002) When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills, *Journal of Experimental Psychology: Applied*, **8** (1), pp. 6–16.

- Beilock, S.L. & Carr, T.H. (2005) When high-powered people fail: Working memory and 'choking under pressure' in math, *Psychological Science*, **16** (2), pp. 101–105.
- Beilock, S.L. & Gray, R. (2007) Why do athletes choke under pressure?, in Tenenbaum, G. & Eklund, R.C. (eds.) *Handbook of Sport Psychology*, pp. 425–444, Hoboken, NJ: John Wiley & Sons.
- Beilock, S.L., Bertenthal, I.B., Hoerger, M. & Carr, T.H. (2008) When does haste make waste? Speed-accuracy tradeoff, skill level, and the tools of the trade, *Journal of Experimental Psychology: Applied*, **14** (4), pp. 340–352.
- Beilock, S.L. & Gray, R. (2012) From attentional control to attentional spillover: A skill-level investigation of attention, movement, and performance outcomes, *Human Movement Science*, **31** (6), pp. 1473–1499.
- Blakemore, S.J., Wolpert, D.M. & Frith, C.D. (1998) Central cancellation of self-produced tickle sensation, *Nature Neuroscience*, **1** (7), pp. 635–640.
- Bleak, J.L. & Frederick, C.M. (1998) Superstitious behavior in sport: Levels of effectiveness and determinants of use in three collegiate sports, *Journal of Sport Behavior*, **1998**, pp. 1–15.
- Brown, H., Friston, K. & Bestmann, S. (2011) Active inference, attention, and motor preparation, *Frontiers in Psychology*, **2**, art. 218.
- Brown, H., Adams, R.A., Parees, I., Edwards, M. & Friston, K. (2013) Active inference, sensory attenuation and illusions, *Cognitive Processing*, **14** (4), pp. 411–427.
- Buszard, T., Masters, R.S. & Farrow, D. (2017) The generalizability of working-memory capacity in the sport domain, *Current Opinion in Psychology*, **16**, pp. 54–57.
- Calvo-Merino, B., Grèzes, J., Glaser, D.E., Passingham, R.E. & Haggard, P. (2006) Seeing or doing? Influence of visual and motor familiarity in action observation, *Current Biology: CB*, **16** (19), pp. 1905–1910.
- Cappuccio, M. (2017) Flow, choke, skill: The role of the non-conscious in sport performance, in Radman, Z. (ed.) *Before Consciousness: In Search of the Fundamentals of Mind*, pp. 246–284, Exeter: Imprint Academic.
- Cappuccio, M.L., Gray, R., Hill, D.M., Mesagno, C. & Carr, T.H. (2019) The many threats of self-consciousness: Embodied approaches to choking under pressure in sensorimotor skills, in Cappuccio, M.L. (ed.) *Handbook of Embodied Cognition and Sport Psychology*, p. 101, Cambridge, MA: MIT Press.
- Chen, A.G., Benrimoh, D., Parr, T. & Friston, K. (2019) A Bayesian account of generalist and specialist formation under the active inference framework, *BioRxiv*, [Online], doi: 10.1101/644807.
- Clark, A. (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York: Oxford University Press.
- Christensen, W., Sutton, J. & Mcllwain, D. (2015) Putting pressure on theories of choking: Towards an expanded perspective on breakdown in skilled performance, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 253–293.
- Christensen, W. & Bicknell, K. (2019) Affordances and the anticipatory control of action, in Cappuccio, M.L. (ed.) *Handbook of Embodied Cognition and Sport Psychology*, pp. 601–621, Cambridge, MA: MIT Press.
- Damisch, L., Stoberock, B. & Mussweiler, T. (2010) Keep your fingers crossed! How superstition improves performance, *Psychological Science*, **21** (7), pp. 1014–1020.

- Dreyfus, H.L. (2005) Overcoming the myth of the mental: How philosophers can profit from the phenomenology of everyday expertise, *Proceedings and Addresses of the American Philosophical Association*, **79** (2), pp. 47–65.
- Dudurich, M. (2014) Ranking the 5 worst equipment changes by a PGA tour player, *Bleacher Report*, [Online], <https://bleacherreport.com/articles/2221591-ranking-the-5-worst-equipment-changes-by-a-pga-tour-player#slide4> [16 December 2019].
- Friston, K. (2010) The free-energy principle: A unified brain theory?, *Nature Reviews Neuroscience*, **11** (2), pp. 127–138.
- Friston, K. (2011a) What is optimal about motor control?, *Neuron*, **72** (3), pp. 488–498.
- Friston, K. (2011b) Embodied inference: Or ‘I think therefore i am, if i am what i think’, in Tschacher, W. & Bergomi, C. (eds.) *The Implications of Embodiment (Cognition and Communication)*, pp. 89–125, Exeter: Imprint Academic.
- Friston, K., Adams, R.A., Perrinet, L. & Breakspear, M. (2012) Perception as hypotheses: Saccades as experiments, *Frontiers in Psychology*, **28** (3), art. 151.
- Friston, K., Thornton, C. & Clark, A. (2012) Free-energy minimization and the dark-room problem, *Frontiers in Psychology*, **3**, art. 130.
- Gray, R. (2014) Embodied perception in sport, *International Review of Sport and Exercise Psychology*, **7** (1), pp. 72–86.
- Gray, R. & Allsop, J. (2013) Interactions between performance pressure, performance streaks, and attentional focus, *Journal of Sport & Exercise Psychology*, **35** (4), pp. 368–386.
- Gray, R. & Cañal-Bruland, R. (2015) Attentional focus, perceived target size, and movement kinematics under performance pressure, *Psychonomic Bulletin & Review*, **22** (6), pp. 1692–1700.
- Hohwy, J. (2016) The self-evidencing brain, *Noûs*, **50** (2), pp. 259–285.
- Hill, D.M. & Hemmings, B. (2015) A phenomenological exploration of coping responses associated with choking in sport, *Qualitative Research in Sport, Exercise and Health*, **7** (4), pp. 521–538.
- Idei, H., Murata, S., Chen, Y., Yamashita, Y., Tani, J. & Ogata, T. (2018) A neuro-robotics simulation of autistic behavior induced by unusual sensory precision, *Computational Psychiatry (Cambridge, Mass.)*, **2**, pp. 164–182.
- Jackson, R.C. & Beilock, S.L. (2008) Performance pressure and paralysis by analysis: Research and implications, in Farrow, D., Baker, J. & MacMahon, C. (eds.) *Developing Sports Expertise: Researchers and Coaches Put Theory into Practice*, pp. 104–118, London: Routledge.
- Jordet, G. (2010) Choking under pressure as self-destructive behavior, in Nicholls, A.R. (ed.) *Coping in Sport: Theory, Methods, and Related Constructs*, pp. 239–259, New York: Nova Science Publishers.
- Kawato, M. (1999) Internal models for motor control and trajectory planning, *Current Opinion in Neurobiology*, **9** (6), pp. 718–727.
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K. & Kiverstein, J. (2018) The Markov blankets of life: Autonomy, active inference and the free energy principle, *Journal of the Royal Society Interface*, **15** (138), 20170792.
- Kirsh, D. & Maglio, P. (1994) On distinguishing epistemic from pragmatic action, *Cognitive Science*, **18** (4), pp. 513–549.
- Koedijkker, J.M. & Mann, D.L. (2015) Consciousness and choking in visually-guided actions, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 333–348.

- Limanowski, J. (2017) (Dis-)attending to the body, in Metzinger, T. (ed.) *Philosophy and Predictive Processing*, pp. 283–295, Frankfurt am Main: MIND Group.
- MacIntyre, T. & Moran, A.P. (2007) A qualitative investigation of meta-imagery processes and imagery direction among elite athletes, *Journal of Imagery Research in Sport and Physical Activity*, **2** (1).
- Marewski, J.N., Gaissmaier, W. & Gigerenzer, G. (2010) Good judgments do not require complex cognition, *Cognitive Processing*, **11**, pp. 103–121.
- Masters, R.S.W. (1992) Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure, *British Journal of Psychology*, **83** (3), pp. 343–358.
- Masters, R. & Maxwell, J. (2008) The theory of reinvestment, *International Review of Sport and Exercise Psychology*, **1** (2), pp. 160–183.
- Montero, B. (2016) *Thought in Action*, New York: Oxford University Press.
- Mullen, R., Hardy, L. & Oldham, A. (2007) Implicit and explicit control of motor actions: Revisiting some early evidence, *British Journal of Psychology*, **98** (Pt 1), pp. 141–156.
- Parr, T. & Friston, K. (2017) Working memory, attention, and salience in active inference, *Scientific Reports*, **7**, 14678, pp. 1–21.
- Parr, T. & Friston, K. (2018) The anatomy of inference: Generative models and brain structures, *Frontiers in Computational Neuroscience*, **12**, art. 90.
- Perrykkad, K. & Hohwy, J. (2020) Fidgeting as self-evidencing: A predictive processing account of non-goal-directed action, *New Ideas in Psychology*, **56**, 100750.
- Ramstead, M.J.D., Kirchhoff, M.D. & Friston, K.J. (2019) A tale of two densities: Active inference is enactive inference, *Adaptive Behavior*, 1059712319862774.
- Roberts, J., Jones, R., Harwood, C., Mitchell, S. & Rothberg, S. (2001) Human perceptions of sports equipment under playing conditions, *Journal of Sports Sciences*, **19** (7), pp. 485–497.
- Robertson, I. & Kirchhoff, M. (this issue) Anticipatory action: Active inference in embodied cognitive activity, *Journal of Consciousness Studies*, **27** (3–4).
- Runswick, O.R., Green, R. & North, J.S. (2019) The effects of skill-level and playing-position on the anticipation of ball-bounce in rugby union, *Human Movement Science*, **69**, 102544.
- Sarason, I.G. (1988) Anxiety, self-preoccupation and attention, *Anxiety Research*, **1** (1), pp. 3–7.
- Schippers, M.C. & Van Lange, P.A. (2006) The psychological benefits of superstitious rituals in top sport: A study among top sportspersons, *Journal of Applied Social Psychology*, **2006**, pp. 2532–2553.
- Schneider, W. & Chein, J.M. (2003) Controlled & automatic processing: Behavior, theory, and biological mechanisms, *Cognitive Science*, **27** (3), pp. 525–559.
- Schwartenbeck, P., Passecker, J., Hauser, T., FitzGerald, T., Kronbichler, M. & Friston, K. (2018) Computational mechanisms of curiosity and goal-directed exploration, *BioRxiv*, [Online], doi: 10.1101/411272.
- Sutton, J. (2007) Batting, habit and memory: The embodied mind and the nature of skill, *Sport in Society*, **10** (5), pp. 763–786.
- Swann, C., Keegan, R.J., Piggott, D. & Crust, L. (2012) A systematic review of the experience, occurrence, and controllability of flow states in elite sport, *Psychology of Sport and Exercise*, **13** (6), pp. 807–819.
- Tani, J. (2017) *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, Oxford: Oxford University Press.

- Wine, J.D. (1971) Test anxiety and direction of attention, *Psychological Bulletin*, **76** (2), pp. 92–104.
- Wolpert, D.M., Ghahramani, Z. & Jordan, M.I. (1995) An internal model for sensorimotor integration, *Science*, **269** (5232), pp. 1880–1882.
- Wulf, G., McNevin, N. & Shea, C.H. (2001) The automaticity of complex motor skill learning as a function of attentional focus, *The Quarterly Journal of Experimental Psychology: Section A*, **54** (4), pp. 1143–1154.

Chapter 2 – The Embodiment of Social Skills.

Social Attention for Interactive Engagement

*This chapter investigates how embodied robots establish and maintain social attention with humans, presenting some original applications of humanoid robotics for the diagnosis and the treatment of autism. The first part investigates why the impairment of social attention is one of the core symptoms of autism. The second part, dedicated to **CP3**, discusses why smart technological devices in general, and socially interactive humanoid robots in particular, are suited to support social attention in autistic people. The third part, dedicated to **CP4**, presents the results of an empirical study involving a semi-autonomous adaptive artificial agent developed by our team to measure, solicit, and maintain social attention during playful interaction with autistic children.*

Chapter 2, Part 0: Preamble

2.0.1 Themes and problems

Chapter 2 investigates (i) the robots' capability to produce and read social attention, (ii) the way humans rely on social attention to engage socially with robots, and (iii) the function of robot's embodiment in establishing social attention with humans. Social attention deficits are a defining component of autism, so much so that the scientific knowledge of social attention was importantly influenced by the last 40 years of autism studies. From a thematic point of view, I was attracted by the scientific study of the autistic condition because of its relevance to embodiment: the growing clinical knowledge about the Autistic Spectrum Disorder (ASD) allows us to put to test some key theses of ECT. The intersection between ECT and ASD is represented the Embodied Account of Autism (EAA), which informs the background of the work with robots done by me and my collaborators in this space. I have investigated social attention and its embodiment using Nao, an embodied social robot specifically personalized to effectively engage with autistic children through its social attention capabilities: with its distinctive embodied features and its design expressly inspired by the principles of ECT, our customized Nao allowed my research team to create an efficacious HRI application for the diagnosis and treatment of autistic symptoms. Also, Nao allowed to us to test, at once, our theoretical hypotheses on the embodied underpinnings of autism and the effectiveness of the ECT-aware design in social robotics.

2.0.2 Research outcomes covered in this section

The outcome described in Chapter 2 is an interactive automated/semi-autonomous system designed for maximising social engagement in autistic children (Alnajjar et al 2019 [SP12], 2020 [CP4]). Importantly, the system has the appearance of a programmable robot (Aldebaran's Nao) equipped with humanoid embodied features (sensors and actuators) that allow it to both interact socially with a human and detect their social engagement. Thus, both the embodiment of the robots and the embodiment of the human users are crucial for the system to operate correctly. Importantly, the details of the physical embodiment of the robot play two different roles: first to attract and measure the attention of the user and, second, also to actively modulate such attention for prolonging the user's engagement. This project is based on some earlier preparatory studies, including a systematic review of the

technological literature about autistic children and an empirical study to validate the proof-of-concept (Alahbabi et al 2017 [SP13]; Alnajjar et al 2021b [CP3]).

2.0.3 Author's publications used in this chapter

In Chapter 2, I will review and discuss the contents of 7 of my publications (2 “Core Publications” and 5 “Support Publications.”)

Core Publications

[CP3] Alnajjar F.S., Cappuccio M.L., Mubin O., Arshad R., Shahid S. (2021b)

[CP4] Alnajjar F.S., Cappuccio M.L., Renawi A., Mubin O., Kiong Loo C. (2020)

Support Publications

[SP10] Cappuccio M.L., Stephen V.S. (2013)

[SP11] Shepherd V.S., Cappuccio M.L. (2012)

[SP12] Alnajjar F., Renawi A.M., Cappuccio M.L., & Mubin O. (2019)

[SP13] Alahbabi M., Almazroei F., Almarzoqi M., Almeheri A., Alkabi M., Nuaimi A.A., Cappuccio M.L., Alnajjar F. (2017)

[SP14] Cappuccio M.L. (2015)

Chapter 2, Part 1: Social Attention Deficits in Autism

2.1.1 Towards an embodied understanding of social attention

The HRI applications of ECT reviewed in this chapter aim to facilitate the diagnosis and the treatment of the Autism Spectrum Disorder (ASD) by means of interactive humanoid robots capable to manipulate social attention. The chapter begins with an introduction to social attention deficits in the ASD and continues with an inquiry into the embodied features that determined the enormous success of humanoid robots in the clinical treatment of autism, before reviewing the experimental and modelling work conducted by me and my research team and assessing its original contribution to these areas of investigation. In the first section, I will characterize the impairment of social cognition in general as a distinctive problem associated with ASD (Ghiglino et al 2023). This characterization will highlight social attention as a primary area of interest for the study of social cognition in ASD people (Gallese 2007; Gallagher 2008). The nature of attention is embodied, social robots must be designed very carefully to produce and manipulate social attention. As social attention deficits are a defining component of autism, we need to contextualize social attention within ASD studies and the impairments of social cognition associated with autism.

Social attention (Shepherd & Cappuccio 2012 [SP11]) is an important topic at the intersection of social cognition, attention, and embodiment. Social attention is the attention that social agents pay to each another with the purpose to establish and maintain social engagement (Ciardo & Wykowska 2022). It represents a very important form of *interaction* and *communication* at once. It is possible to recognize the “social” (or pseudo-social in the case of robots) valence of attention in interaction, regardless that the interactants are human adults, children, animals, or robots: on the one hand, attention serves to gather information about the other’s status and activity; on the other hand, attention serves to communicate that one is actively gathering such information. Thus, the information collected through social attention is about the others and about the activity that one is conducting, or intends to conduct, jointly with them (Franco 2013). Ultimately, social attention is embodied because it leverages a system of embodied practices aiming to request, validate, and reciprocate an ongoing, two-way exchange of information that serves both recreational engagement and goal-oriented interaction (Gallagher 2008). That is why the role played by embodiment in both expressing and reading social attention is inherently, not accidentally, cognitive: it

contributes to disambiguate the information expressed via interaction, providing it with an operational meaning.¹²

Social attention plays a major role in HRI and Social Robotics and any effective social robot must be designed to both accurately capture the social connotations of the user's attention and to produce attentional cues that correctly articulate and develop the social interaction with them (De Jaegher Di Paolo & Gallagher 2010). The embodied cues produced by a robot can guide the user's attention in intuitive ways, redistributing the cognitive load between multiple concurrent tasks and producing spontaneous interactions with minimal efforts. For social robotics, this attentional management is very important, as it supports relationship building, empathy, and trust by aligning the robot's behaviour with human social expectations (Gaudiello et al 2015). Effective social robots leverage embodiment not just to capture attention but to maintain an attentional balance that enhances the user's experience without overwhelming or distracting them. As social attention deficits are a defining component of autism, we need to contextualize social attention within ASD studies and the impairments of social cognition associated with autism (Schultz 2005).

2.1.2 Social cognition deficits in the Autistic Spectrum Disorder

ASD is a neurodevelopmental condition characterized by pervasive challenges in social communication and interaction, alongside restricted and repetitive patterns of behaviour, interests, or activities (American Psychiatric Association 2013).¹³ Despite decades of investigation, ASD's etiology remains multifaceted, involving complex interactions between genetic, epigenetic, and environmental factors (Sandin et al., 2017). ASD is a spectrum disorder, reflecting wide variability in symptom severity and functional impact across individuals (Lord et al., 2020). According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), diagnostic criteria for ASD include persistent deficits in social communication and interaction across multiple contexts, as well as restricted,

¹² This information derives from the movements of eyes and head, or whole-body behaviours that indicate the direction of attention, its target, and a sustained focus. Depending on how it is bodily expressed (and the amount and extension of the bodily features involved), social attention can be explicit (public, as performed in front of everybody) or rather covert (seemingly visible only to a few and concealed to the others); direct (staring, that is fixating the other's eyes) or rather indirect (intermittent and focussing on the other's body rather than their eyes); perceptually based (rooted in online sensory input) or rather symbolic (mediated by non-sensory and representational processes).

¹³ With an estimated prevalence of 1 in 44 children in the United States (Centers for Disease Control and Prevention, 2021), autism represents a significant area of research in cognitive and developmental neuroscience.

repetitive behaviours (APA, 2013). These manifestations can range from subtle social awkwardness and sensory sensitivities to profound difficulties in language development and self-care.

A hallmark feature of autism lies in the domain of social cognition, the subset of cognitive processes that enables individuals to perceive, interpret, and respond to social cues (Cappuccio 2015 [SP14]; Gallagher 2007, 2008). Social cognition encompasses a range of abilities, including emotion recognition, perspective-taking, and understanding social norms. Deficits in these domains are central to the behavioural profile of ASD and are critical to understanding the condition's impact on daily functioning. Social cognition refers to the cognitive processes that underlie the perception, interpretation, and generation of responses to social stimuli. Core components of social cognition include: (i) *Theory of Mind* (ToM), the ability to infer the mental states, such as beliefs, desires, and intentions, of oneself and others relying on folk psychological principles, i.e. a set of rational laws that specify how mental states relate to or produce other mental states, and, in some cases, on the internal simulation of the other's reasons, motivations, and evaluative perspective (Baron-Cohen, Leslie & Frith 1985); (ii) *Emotion recognition*, the capacity to identify and interpret emotional expressions in facial, vocal, or postural cues (Dawson 2012); (iii) *Social perception and interaction*, the ability to decode and produce social signals that convey information about one's own and others' intentions or emotions (Schultz 2005); (iv) *Social reasoning*, the higher-order processes involved in understanding social norms, making moral judgments, and predicting others' behaviour based on contextual cues (Adolphs 2009). These abilities are essential for navigating social interactions and maintaining meaningful relationships. Deficits in social cognition are well-documented in ASD, which is why they are often iconically identified with the four major social impairments typically faced by individuals with the autistic condition: difficulty with false-belief test; impaired joint attention (absent or delayed development of declarative pointing); emotion recognition impairment (e.g., distinguish embarrassment from sarcasm); reduced social normal understanding (implicit rules and conventions)¹⁴.

¹⁴ Future research must address critical gaps, such as the heterogeneity of social cognition deficits across the autism spectrum and the role of cultural factors in shaping social cognitive processes. Early interventions targeting joint attention and emotion recognition have shown promise in improving social outcomes (Kasari et al., 2012). Additionally, technological advancements, such as virtual reality-based interactive tutors and conversational agents for social skills training, provide novel avenues for enhancing social understanding in interactional contexts (Milne et al 2011a, 2011b, 2013, 2018). Integrating neuroimaging, computational modelling, and longitudinal studies will be essential for unravelling the complexities of ASD and its impact on social cognition. ASD presents not only serious challenges in the clinical domain but also theoretical challenges: autism solicits a deep revision of many intellectualist and meta-representationalist concepts with which clinicians and cognitive scientists are used to categorize fundamental abilities such as ToM, emotion recognition, and social reasoning. We know that these impairments stem from a confluence of genetic, neurological, and developmental

2.1.3 Neurocognitive Theories of ASD

To understand social attention deficit as a distinctive feature of ASD, we need to map the main theoretical models proposed to explain the neurocognitive atypicalities of autism, situating the Embodied Account of Autism (EAA) among them. These theories have been central in the debate, though recent trends favour multi-factor models rather than competing hypotheses. The mechanisms driving social cognition deficits in ASD remain under investigation, but five major hypotheses target neuro-functional, developmental, and behavioural-perceptual aspects, providing insight into diagnostic and therapeutic strategies.

(i) *Impaired Theory of Mind (iTOM)* – This account imputes autism to an impairment of Theory of Mind, supposedly preventing autistic individuals from inferring others' beliefs and intentions, as indicated by underperformance in false-belief tests. However, the idea that ASD people lack ToM has been questioned: such deficits occur only in a minority of cases and can be mitigated through interventions. (ii) *Weak Central Coherence Theory (WCC)* – “Weak Central Coherence” claims that the systems responsible for integrating information globally are impaired in ASD. Evidence suggests an overload of local sensory information, possibly linked to cortical network imbalance. Though focused on cognitive style rather than social engagement, it is compatible with ECT and useful for developing EAA. Detail-oriented processing may hinder holistic integration of sensorimotor experiences and social contexts, producing imperfect enactive coupling. (iii) *Executive Dysfunction Hypothesis (EDH)* – This theory attributes autism to impairments in executive functions, the skills for regulating thoughts, actions, and emotions in goal-oriented ways. Such deficits explain many ASD symptoms but do not specify a single causal mechanism, as executive functions cover a wide range of processes. (iv) *Broken Mirror Hypothesis (BMH)* – The Broken Mirror account (Ramachandran & Oberman 2007) links social deficits to dysfunctional mirror neuron systems (Cattaneo et al. 2007). Gallese and colleagues argue mirror neurons underlie “embodied simulation”: in the premotor cortex, they support both motor competence (action execution, planning) and recognition/understanding of observed or imagined actions (Gallese 2007; Gallese & Sinigaglia 2011; Iacoboni & Dapretto 2006). Mirror neurons have been proposed as neural correlates of empathy, with neuroimaging showing mirror properties in the amygdala and cingulate cortex. Embodied Simulation Theory holds that the brain re-uses motor and somatosensory competences to understand others' actions (Aglioti et al. 2008), drawing on one's own repertoire of actions to attribute purposes to others. Artificial models of mirror neurons have been developed in developmental robotics and

factors, but we still need rigorous methods to highlight the intricate interplay between biology and behaviour or, in a different way, between mental processes and physical embodiment.

active inference (Metta et al. 2006; Cappuccio 2009 [SP1]; Asada 2016). (v) *Sensory Overload Hypothesis* (SOH) – This account attributes ASD social attention deficits to heightened sensory sensitivity, which makes everyday stimuli overwhelming. ASD brains may process input in ways that are more intense, fragmented, or disorganized than neurotypical ones. The most common consequence is overload, which may appear as hypersensitivity (e.g. loud noises, bright lights), hyposensitivity (seeking strong input such as spinning or pressing), or modulation difficulties (rapid shifts between the two). This leads to behavioural manifestations like stimming (rocking, hand-flapping as self-regulation), withdrawal from overstimulation (e.g. noisy classrooms), and meltdowns or shutdowns when input becomes unmanageable.

Fig. 6 - Theories of Autism: Comparing the Key Features

Theory	Core Idea	Strengths / Evidence	Limitations / Critiques	Relation to ECT and EAA)
Impaired Theory of Mind (ToM)	ASD arises from difficulty inferring beliefs, intentions, and mental states of others (e.g. false-belief tests).	Explains social reasoning deficits in some individuals.	Only a minority of ASD people show consistent ToM deficits; interventions can improve performance. Overly narrow, ignores sensorimotor aspects.	Weak alignment. Focuses on representational reasoning rather than embodied, interactive processes.
Weak Central Coherence Theory (WCC)	ASD marked by bias toward local detail over global integration due to imbalance in cortical networks.	Supported by sensory overload symptoms and detail-oriented processing. Compatible with ECT and enactivism.	Focuses on cognitive style, not causal mechanisms; less direct about social interaction.	Moderate alignment. WCC may explain difficulties integrating sensorimotor experiences holistically, consistent with EAA.
Executive Dysfunction Hypothesis (EDH)	ASD linked to impairments in executive functions (planning, inhibition, self-regulation, goal-directed action).	Explains many behavioural symptoms across attention, flexibility, and control.	Too broad; does not identify precise mechanisms or explain all ASD expressions.	Partial alignment. Refers to motor cognition (planning/inhibition) but lacks explicit embodied focus.
Broken Mirror Hypothesis (BMH)	Social deficits stem from dysfunction in mirror neuron systems (Ramachandran & Oberman 2007; Gallese 2007). Mirror neurons enable embodied simulation of others' actions/emotions.	Neuroimaging supports mirror neuron involvement in action understanding and empath. Influential in linking motor & perceptual mechanisms to social cognition.	Controversial: evidence for mirror neuron deficits in ASD is mixed. Not sufficient alone to explain the spectrum's diversity.	Strong alignment. Embodied simulation via mirror neurons exemplifies enactive, sensorimotor grounding of social cognition; close to EAA.
Sensory Overload Hypothesis (SOH)	ASD results from atypical sensory processing (hyper/hyposensitivity, modulation difficulties), causing overwhelm, stimming, withdrawal, or meltdowns.	Explains a wide range of everyday experiences and coping behaviours; clinically relevant for therapies.	Broad and heterogeneous; mechanisms of overload not fully understood.	Strong alignment. Emphasizes online bodily/sensory activity shaping cognition; complements BMH in an embodied account.

2.1.4 An embodied approach to autism

Some theories align more closely with ECT. The BMH, and partly WCC, stress that social cognition arises from sensorimotor experience and bodily interaction. They suggest impairments in motor cognition and sensory integration can impede embodied social understanding (Klin et al. 2003) or render interaction cognitively demanding. The SOH complements these views, emphasizing the role of online bodily and sensory activity in shaping cognition: difficulties integrating sensory inputs could disrupt the dynamic body–brain–environment interaction described by ECT. SOH and BMH reinforce each another: dysfunctional mirror neurons may explain why ASD individuals are overwhelmed by social stimuli, while unfiltered input may in turn prevent mirror systems from performing embodied simulation. In my work with social robots, I have treated these hypotheses as parts of an integrated embodied perspective, highlighting their consistency and setting aside differences. This allows us to outline an articulated theoretical account of ASD informed by embodied cognition, the EAA.

Chapter 2, Part 2: Robotic Approaches to the Diagnosis and Treatment of Autism

This section is centred on Core Publication #3 [CP3]

2.2.1 Technological approaches to the diagnosis and treatment of Autism: a survey

This Part is dedicated to the systematic review of the scientific literature on the technological approaches to diagnose and treat autism that my team and I published in 2020 in the *International Journal of Humanoid Robotics* (Alnajjar et al 2021b [CP3]). The review focusses on “sensor-based assessment systems for automated and unbiased quantitative assessments of children’s engagement and attention fluctuations during interaction with robots”. This review shows that advances in digital, smart, and autonomous technologies have revolutionized the diagnosis and treatment of ASD, providing tools that are more precise, scalable, and adaptable than traditional methods. These approaches leverage AI, machine learning, virtual and augmented reality, robotics, and neuroimaging to address challenges in both early detection and intervention. For example, machine learning algorithms are used to analyse behavioural and physiological data to identify ASD markers, such as atypical gaze patterns, facial expressions, and voice modulation. AI tools like the Autism Detection in Toddlers (ADIT) system offer scalable, non-invasive methods for early diagnosis. Also, wearable devices and cameras collect data on social and motor behaviours in real-world settings, providing a richer picture of ASD traits than clinical assessments (Ploetz et al., 2012). Virtual Reality and Augmented Reality can be used to create simulated social environments, i.e. controlled environments where individuals with ASD can practice social interactions without the unpredictability of real-world settings (Parsons & Mitchell, 2002). Immersive technologies can also be used for sensory training, helping children with ASD adjust to overwhelming sensory stimuli by gradually exposing them to controlled virtual scenarios. This training can produce good results in conjunction with techniques like EEG-based neurofeedback to train ASD individuals to regulate their brain activity, targeting regions implicated in social cognition deficits (Coben et al., 2014). All these approaches are promising but their application often involves complicated laboratorial procedures and sophisticated settings, while their therapeutic efficacy remains, in many cases, not completely demonstrated, if not largely speculative.

On the contrary, as my colleagues and I argue in the same review [CP3], robotics represent a very concrete and promising approach to ASD intervention, with noticeable

effects on social engagement starting from the very first sessions. In many countries, robots became a methodological standard and an indispensable piece of equipment in many clinical settings (Pennisi et al 2016). The advantages of using robots for the treatment of ASD are well-known (Coeckelbergh et al 2017). First of all, unlike humans, robots can repeat actions, gestures, and responses with high precision and predictability. The consistency of mechanically produced movements and routines is of great help to children with ASD, who often find human interactions overwhelming due to their variability (Scassellati 2005). Also, robots, especially humanoid ones, can appear less intimidating than humans. Their simplified facial expressions, non-threatening presence, and limited range of motion reduce the sensory and social complexity that often overwhelms autistic individuals. This is particularly important because many children with ASD exhibit strong interest in technology: robots can harness this interest to create engaging interventions that encourage interaction and motivate skill development (Diehl et al., 2012). In addition to this, robots' functionalities are highly personalizable and their physical appearance can be customized at will (possibly dynamically, and in real time). Also, their software (the social engagement protocols) can be programmed to adapt their behaviour to the individual needs of each child, providing tailored support for specific deficits, such as joint attention, emotion recognition, or turn-taking in conversations.

2.2.2 ASD intervention with social robots: five areas

There are five areas of ASD intervention in which social robots have proven particularly effective. (i) *Social Skills Training* – Social robots like Nao and Kaspar are used in numerous clinical institutions all over the world to teach children basic social skills, such as maintaining eye contact, recognizing emotions, and taking turns in conversation. Because of their non-threatening, predictable and schematic social features, these robots induce and maintain social behaviours in a way that is both accessible and engaging. (ii) *Joint Attention and Imitation* – Social robots are often used to improve joint attention, a foundational skill often impaired in ASD. For example, robots can direct the child's gaze to an object using ostensibly their hand or finger, in a way that scaffolds the development of shared attention. (iii) *Emotion Recognition* – Social robots with simplified facial expressions can help children with ASD practice recognizing and interpreting emotions in a controlled, step-by-step manner. (iv) *Behavioural Monitoring* – Advanced social robots equipped with sensors and AI can monitor a child's behavioural responses during interactions, tracking important behavioural variables (e.g., production of verbal responses or attempts to interact) and providing real-time feedback to clinicians and caregivers. (v) *Physical Therapy and Motor Coordination* – Social robots like Milo and Keepon are used in motor skill interventions,

helping children with ASD improve coordination through guided physical activities, exercises involving imitation of body posture and gesture, and timed responses coordinated with rhythm and music.

The use of robotics in ASD interventions has generated considerable enthusiasm due to its noticeable effects and due to its inherent potential for scalability, personalization, and integration with technologically innovative learning environments (Pennisi et al 2016; Esteban et al 2017). However, open-questions and doubts remain, including the *uncertain long-term efficacy* (robots can successfully engage children in the short term, but evidence of long-term improvements in social and cognitive outcomes is still limited, also due to the well-known difficulty of measuring social effects in a timeframe of months or years), the *limited generalizability* of the social skills that they train (it is still unclear to what extent the skills learned through robot interactions transfer to real-world human interactions), and – concerning the requirements for technology diffusion– *the comparatively high cost and the accessibility* (humanoid robots can be quite expensive, limiting their availability to many families and schools).

Our line of research made a quite specific but very concrete step towards the resolution of these three problems, proposing a robotic intervention that is applicable over different temporal scales and that, using relatively accessible equipment, leads to easily exportable social skills.¹⁵ Our experiments helped us realize that robots are uniquely suited for working with children with ASD by leveraging not only their physical appearance but also their capability to detect and measure embodied attention.

2.2.3 Main applications of robotics in ASD studies

Social robotics is perfectly situated to test the EAA account for at least four major reasons, which concern the embodied features of robots and the role played by these features in interaction with ASD children. (i) *Robots as embodied agents* – First, as inherently embodied systems, robots capitalize on their physical presence to interact dynamically with the child's environment. These interactions are central to embodied social cognition, which argues that social understanding is not abstract but arises from engaging with others through gestures, movement, and shared activities (Gallagher, 2005). (ii) *Scaffolding Sensorimotor Integration*

¹⁵ The research done with my team in this space builds on the work of many important pioneers, like Kerstin Dautenhahn and Ben Robins (Dautenhahn 2007, Nehaniv et al 2013). While recognizing this debt, we have been originally experimenting with new robotics interventions when this methodology was not a universally adopted standard yet.

– Robots can produce controlled actions and gestures as scaffolds for sensorimotor development by providing structured, physical interactions. For example, a robot might guide a child through repetitive movements that promote motor coordination or initiate joint attention by pointing or turning its head. (iii) *Simplified Interaction Contexts* – Robots offer simplified yet embodied social interactions that can be gradually made more complex. This aligns with embodied cognition’s emphasis on the stepwise development of social skills through direct experience and with the hypothesis that ASD derives from an information overload, i.e. an unbearable excess of social interaction. (iv) *Dynamic Reciprocity* – Finally, robots can engage in reciprocal, mutually coordinated sensorimotor interactions, such as mirroring the child’s movements or imitating the patterns of their touch. These embodied exchanges mimic the natural dynamics of human interaction, as conceptualized by enactive cognition theory, while providing a lower-stress environment for practice.

Leveraging these features, robotic interventions inspired by ECT are designed to address diagnostic, therapeutic, and epistemic-theoretical goals. (i) *Assessment* – First, robots can be used in diagnostic settings to evaluate and measure embodied social behaviours, such as a child's ability to follow gaze, imitate movements, or synchronize actions with the robot. These behaviours are difficult to assess using traditional methods but are critical to understanding autism from an embodied perspective. (ii) *Treatment* – Second, embodied robot-based interventions emphasize sensorimotor coordination and real-time social interactions, aiming to strengthen the foundational skills that underlie more complex social cognition. For example, robots might guide children through physical exercises that simultaneously promote motor control and joint attention. (iii) *Explanation, understanding* – Third, robotics also provides an experimental platform for assessing the behaviour of autistic patient (Milne et al 2018) or for formulating hypothesis about the causes of that behaviour, via artificial simulations of a cognitive model (Idei et al 2018). For example, robotics allow experimenting with sensorimotor dynamics: robots can be programmed to simulate sensorimotor interactions that can be either simple or complex (based on the intensity and richness of the sensorimotor cues involved) and either typical and atypical (based on how common they are), allowing researchers to investigate how such interactions influence social cognition (Ghiglino et al 2023). Also, it allows studying action-perception loops: robot-child interactions provide a controlled setting to study how children with ASD respond to dynamic feedback (with behaviour, emotion, or verbal expression), shedding light on how atypical sensorimotor experiences might cause autistic responses by either *overstimulating* or *understimulating* the children.

In the third and final Part of this Chapter, our experimental study with ASD children (Alnajjar et al 2020 [CP4]) allows us to achieve these three goals at once, efficaciously demonstrating the advantages of the technological approaches to autism based on robotics. EAA not only opens new avenues for practical intervention but also deepens our theoretical understanding of autism as a condition grounded in sensorimotor experience. Future research should focus on integrating robotics within embodied/enactive frameworks that are both therapeutically efficacious and grounded in a holistic view of human cognition.

2.2.4 Outcomes, limitations, and future research

This section is dedicated to my team's comprehensive survey of the literature about the diagnosis and therapy of ASD children based on interactive, smart and digital technologies (Alnajjar et al 2021b [CP3]). Our review highlighted two crucial notions: first, that direct social attention typically is a form of explicit engagement that autistic children tend to avoid systematically, in a continuous effort to evade the pressure of the other's gaze; second, that autistic children typically display greater interest for the synthetic interactants, compared to the human ones, and engage for longer periods and more comfortably with robots than with caregivers. However, the review also highlighted two problems that negatively affected the earlier ASD interventions based on humanoid robots (and, more broadly, automated technologies): the repetitiveness and inflexibility of pre-programmed routines and the inadequacy of the assessment tools. What are the best strategies to address these problems? Part 3 of this Chapter will answer this question, examining in detail the HRI solutions envisioned by our research team.

Core Publication #3 [CP3]

Alnajjar F.S., Cappuccio M.L., Mubin O., Arshad R., Shahid S. (2021b),
“Humanoid robots and autistic children:
a review on technological tools to assess social attention and engagement”
International Journal of Humanoid Robotics, World Scientific Publishing Co.

Redacted.

Chapter 2, Part 3: A Semi-Autonomous System to Measure and Train Social Attention

This section is centred on Core Publication #4 [CP4]

2.3.1 A project conceived and developed in the UAE

Part 1 and Part 2 of this Chapter served to introduce the line of studies conducted by me and my group of collaborators based in the United Arab Emirates (UAE), at the Cog Sci Lab and at the AI & Robotics Lab of United Arab Emirates University (UAEU). The group was primarily formed by me, Dr Alnajjar, and our research assistants and collaborators, including post-doctoral researchers and post-graduate students (one of them, Moza Alahbabi, was our doctoral candidate). The early conceptualization of these studies started in 2016, and I and my colleagues continued working on the experiment design and proof of concept during the following two years. After collecting the required data, in 2017 and 2018, we were able to produce 4 original papers in which I was directly involved: Alahababi et al (2017) [SP13]; Alnajjar et al (2019) [SP12]; Alnajjar et al (2020) [CP4]; Alnajjar et al (2021b) [CP3]. The last two of them [CP3 and CP4], in which I played a primary role, were published in 2020 and 2021, and represent the main and final research outcome of this project¹⁶.

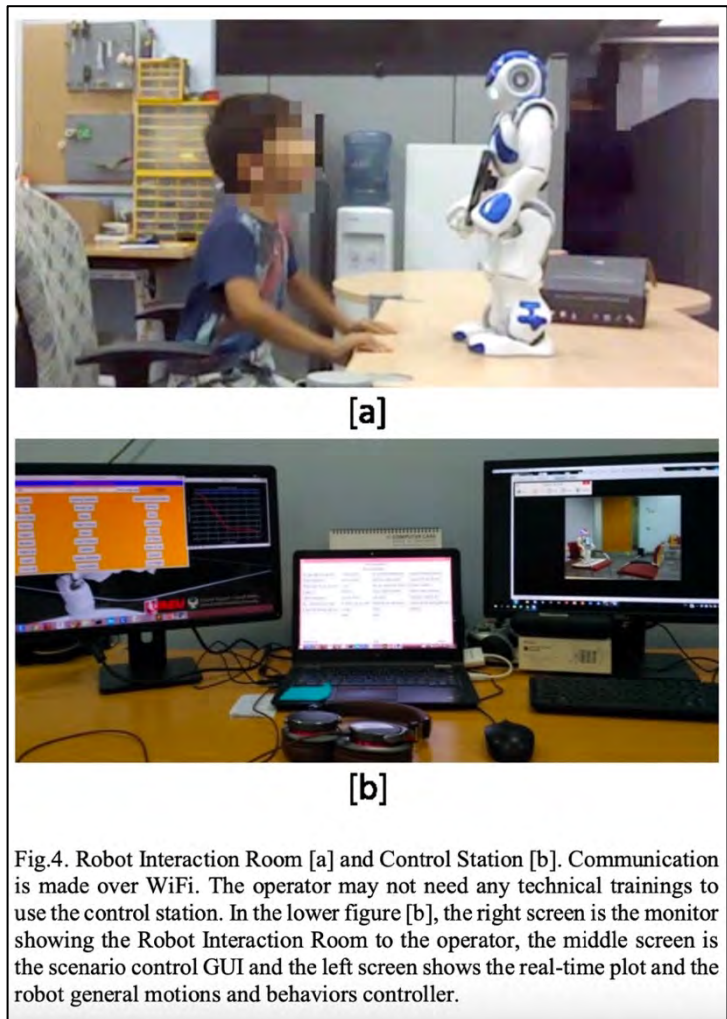


Fig.4. Robot Interaction Room [a] and Control Station [b]. Communication is made over WiFi. The operator may not need any technical trainings to use the control station. In the lower figure [b], the right screen is the monitor showing the Robot Interaction Room to the operator, the middle screen is the scenario control GUI and the left screen shows the real-time plot and the robot general motions and behaviors controller.

¹⁶ These activities were funded by a mix of different sources, including a small research budget from the Provost Office, a grant from Zayed Health Centre and a UPAR grant for basic research from the UAE government.

Our most important empirical study (Alnajjar et al 2020 [CP4]) was possible thanks to an agreement between UAEU and Al Ain Hospital (Al Ain City, UAE), which provided all the young ASD patients after we obtained the standard ethical approval from the ethical committees of both institutions and the voluntary endorsement of the patients' parents, who were very happy their children would be given a chance to experiment a modern, perfectly safe, and apparently very promising therapeutic process. During the early stages of our project (conceptualization and proof-of-concept), we also abundantly relied on the collaboration with a school for special needs students in Abu Dhabi, where we tested the preliminary versions of our methodology with the invaluable help of the clinical personnel and the teachers. Let us review the four publications that we produced as an outcome of our investigation.

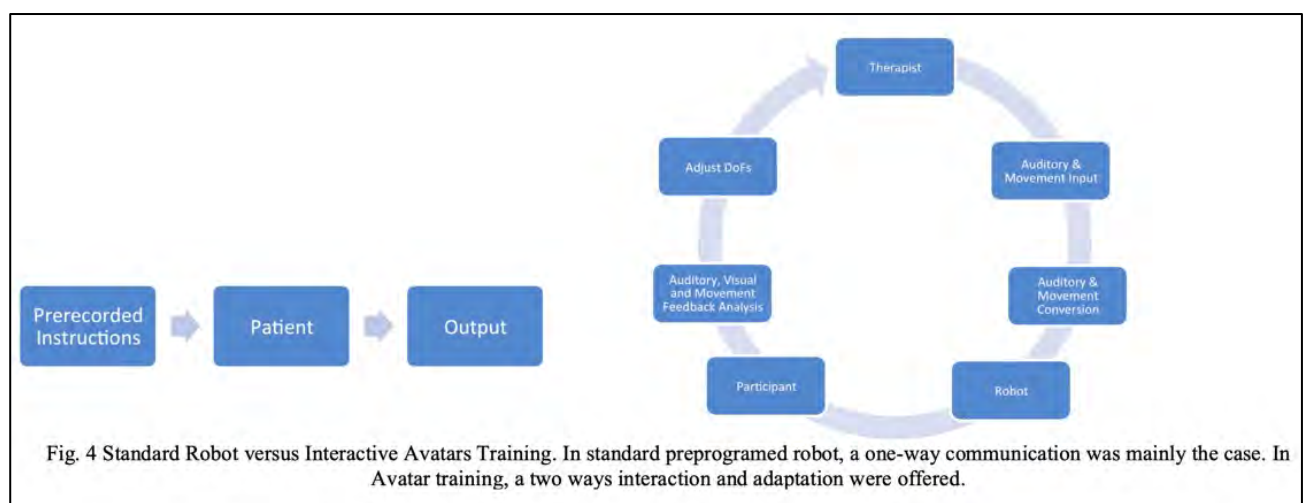
As I mentioned earlier, the first paper we start working on (which was also, incidentally, the last one to be accepted for publication, in 2021), was a review paper: "Humanoid robots and autistic children: a review on technological tools to assess social attention and engagement", *International Journal of Humanoid Robotics*, World Scientific Publishing Co. [CP3]. In Part 2 of this Chapter, I have presented the main outcomes of our literature review and how they informed our work. I have highlighted two important problems that affect the robotics interventions with autistic children: repetitiveness and lack of an accurate assessment system validated by healthcare professionals. Our empirical study [CP4] build on these notions to create a social robot that adjusts in real time to the needs of the child so to prevent both distraction (lack of attention) and debilitating social pressure (excess of attention), as both conditions tend to disrupt social engagement.

2.3.2 Engaging with children through a robotic avatar

I will start reviewing the results of our preparatory study (Alahbabi et al 2017 [SP13]). The paper was presented at an IEEE conference and then published in 2017: "Avatar Based Interaction Therapy: A Potential Therapeutic Approach for Children with Autism", *Proceedings of 2017 IEEE International Conference on Mechatronics and Automation*. This study, which was empirical and involved the interaction of a remotely-controlled social robot with children, represented a preliminary version of the automated system for the assessment of social attention that we describe in CP4. Designed to provide a proof of concept, it does not use the semi-autonomous and adaptive interaction protocol that we developed later: the robot is remotely controlled by a therapist, in a typical "Wizard of Oz" kind of scenario. Also, to simplify data collection and reduce the complications during the interactions with the robots, the study was conducted with neurotypical children, instead of autistic children.

The study tested the functionalities of a robotic-avatar therapeutic system specially designed for children with autism. As an artificial “avatar”, i.e. a robotic twin of the therapist, the avatar robot was remotely controlled by the therapist and was able to fully replicate with its body the therapist’s presence. Therefore, the robot was able, on the one hand, to mimic the therapist’s motion and project his/her voice, and, on the other, to deliver multi-modal sensory feedback to the therapist, including visuals and audio from the child environment. By operating the robot remotely, concealed behind a room divider, the therapist could engage the children in interaction without revealing his/her presence, so to secure the children’s trust and facilitate a more spontaneous interaction with them. This Wizard of Oz setting, by replacing the human presence with a cute robotic character, presents two advantages: first, it lowers the distress generated by direct interaction with the therapist, making the therapeutic sessions more appealing and fun for the children and also more steady and controllable for the therapist; second, it allows the therapist to articulate the session adaptively, in always changing ways, based on the needs and the intuitions of the moment.

The ultimate goal of the study was to compare the performance of our robotic-avatar system with a pre-programmed robot designed to interact with autistic children following a conventional interactive therapy routine supervised by a therapist. The preliminary results, obtained measuring the response levels of 5 healthy children, showed that our proposed avatar robot was better at increasing the response and the interaction time of the children, thus increasing their attention span. Additionally, our system was able to automate and facilitate the measurement of the interaction and its progress over time.



The technical solution we devised was represented by a “robotic avatar-based therapeutic system specially designed for autism patients: the system increases the immersion of both the patient and therapist and facilitates evaluating the progress of the therapy. Our avatar

robot was able to replace the therapist in the room [provided that the therapist was in the room nearby]" (idem). The therapist's body would be fully replaced by the robot's body, as the latter would substitute the former's physical presence and appearance. That is why the therapist was asked to wear a portable motion sensors system that would track his/her movements while immediately replicating the same movements in the robot, thus giving the therapist a partial control over the robot motion while allowing him/her to receive a visual and auditory stream from the robot. "The therapist would also be able to talk to the patient through the robot, by using a robotic voice. We believe that this work can fill the gap between therapists and robot programmers, in addition to being more fun for the patient, in comparison to pre-programmed alternatives" (idem).

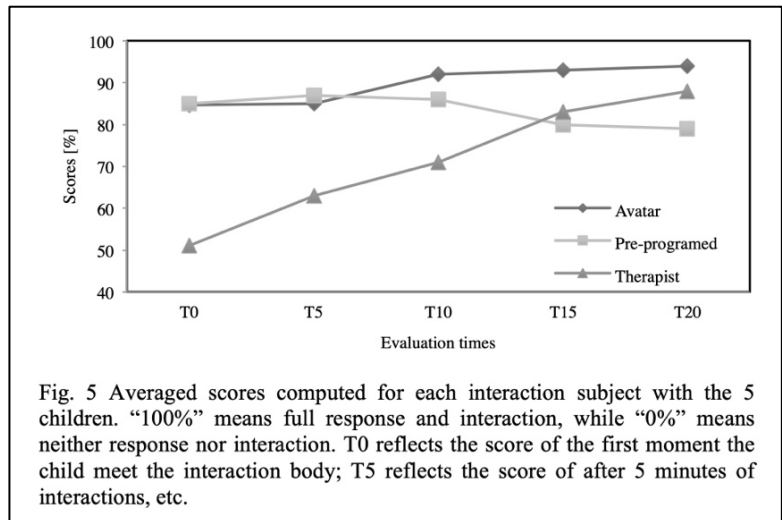
In terms of technological equipment, the system primarily relied on a Perception Neuron set of 32 sensors, which was used for real-time capturing of the therapist's body movements, a humanoid Nao Robot by Aldebaran, to interact with the child and to act and talk on behalf the therapist, and a Microsoft Kinect System to map in real-time the child's movements and verify that they matched the movements performed by the therapist, as the main playful activity proposed by the therapist was a game of body imitation. As the children were primarily asked by the avatar to engage in an imitative task (with movements produced in real time by the therapist and replicated by the robot), Kinect was used to analyse the motion of the participants and verify that they be matched to the movements done by the therapist. Additionally, a Bluetooth heart-rate measuring band was worn by the participant and remotely connected to the therapist's computer: the children's heart was continuously monitored by the therapist to determine their level of anxiety when dealing with the robot.

The children's motion was captured by the Kinect system and compared in real-time to the adjusted movements of the therapist. A score was calculated for each group of movements and stored in a database within the participant's record. These data were used during the subsequent visits to analyse the individual progress of each user and to advise the therapist when their movements had to be adjusted.¹⁷

¹⁷ The evaluation of the mimicking accuracy was calculated using RULA (Rapid Upper Limb Assessment) readings, a system that combines multiple scores: Posture Score A (calculating the deviation of the angles of the limbs from the target posture), Muscle Score (assigned for static poses held for more than 10 seconds or for repetitive actions done more than 4 times per minute); Force Score (given to the weight being held, if the posture includes grasping and carrying of objects); Attention Score (based on whether the subject is looking at the robot or not); Response factors score (measuring whether the patient is talking/answering the avatar-robot or not, using a sound recognition algorithm). We have added the last two scores as a customization of the standard RULA system, to provide a broader characterization of the children's level of engagement during the imitative task.

The study proved the potential of our system and its greater efficacy and flexibility, compared to existing systems that rely only on the therapist's intervention or on automated but non-adaptive technologies: "although the therapist, as an interaction body, was having bad response from the child at the initial period of the meeting, he overcame the pre-

programmed robot by around 14 minutes of interaction (after the therapist could win the trust of the child). Yet, the avatar robot kept its lead until the end of the session. The parents who attended the sessions showed their admiration to their child interaction with the avatar."



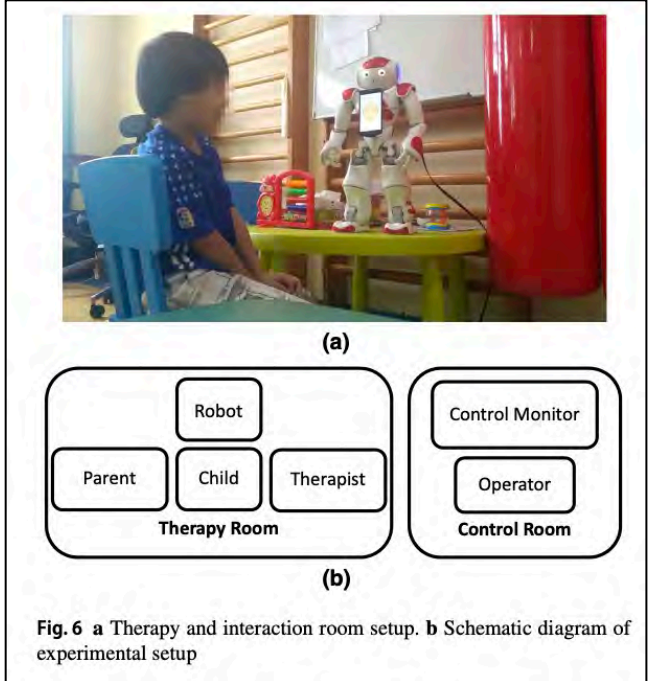
However, this study still does not say anything about the potential to engage autistic children and provide a valid support to treat social attention deficits: "although our robot avatar was better than pre-programmed robot in engaging in interaction with normal children, but we cannot predict yet how differently autistic children would react. Our future plan, therefore, is to introduce the avatar system to autistic children and compare it with conventional therapy/robot systems. We are also planning extra tuning of the scoring system to be more robust and accurate."

2.3.3 A semi-autonomous adaptive system to solicit and measure social attention

These plans were fulfilled by the study documented in the last two publications produced in collaboration with the UAE research team (**SP12** and **CP4**). Both publications refer to the same empirical project: the first paper, "A low-cost autonomous attention assessment system for robot intervention with autistic children" (Alnajjar et al 2018 [**SP12**], was published in the *Proceedings Of The Tenth IEEE Global Engineering Education Conference (Educon 2019)*, held at the American University in Dubai on 9-11 April 2019. This short paper offers only a succinct overview of our project and was presented when our study was still at an intermediate stage of development. The fully developed project and its final outcomes were systematically expounded in the last publication, "Personalized Robot Interventions for

Autistic Children: an Automated Methodology for Attention Assessment”, *International Journal of Social Robotics* (2020) [CP4].¹⁸

2.3.3.1 Semi-autonomous and socially interactive – While our first empirical study with the robot-avatar [SP13] was only meant to validate a proof-of-concept, the following study aimed to make fully operational our system for the assessment and the treatment of ASD symptoms, helping autistic children to cultivate their social attention and social engagement skills. So, while the earlier study used a remotely controlled robot and did not involve ASD patients, the last study used a semi-autonomous robot equipped with an adaptive interaction protocol modulated in real-time by an advanced system for tracking the social responses and the varying level of attention of autistic children.



2.3.3.2 Adaptive and Professionally Calibrated – As we explain in the introductory paragraphs of CP4, we developed and tested a robot specifically designed to train the social attention abilities of children diagnosed with autism spectrum disorder (ASD). Our research was motivated by the awareness that social interaction and communication skills significantly impact human development, learning, and well-being. In the paper, we also emphasize the importance of robot embodiment and the anthropomorphic attitudes that are solicited by the humanoid form of embodiment.

In this context, the adaptive and semi-autonomous qualities of a robot trained via machine learning are presented as a solution to the problem of repetitiveness that afflicts interactive

¹⁸ This arguably is the most important publication of our team and a major achievement for the Robotics & AI Lab at UAEU. I was personally very invested in this study and I provided a major contribution to all the stages of its development, from the conceptualization and the interaction design to data collection, analysis and writing of the paper, although when the paper was accepted by the journal and eventually went to press, I had already moved to Australia to take on a position at the School of Engineering & Information Technology of University of New South Wales Canberra.

robots based on a pre-written, invariant script.¹⁹ Not only does our social robot solve the first problem highlighted in our literature review (repetitiveness), incorporating a semi-autonomous system with an adaptive pattern of tasks and contents; also, it addresses the other problem mentioned by the review (lack of an objective, yet clinically validated, assessment methodology) by incorporating a robust system for the quantitative measurement of social attention and engagement. We designed, perfect, and finely calibrated this system in accord with the indications received by the ASD professional therapists that worked with us. Our robot is specifically programmed to detect the level of social attention in the child directly facing it, measuring their varying level of engagement in the short and in the medium timeframe by using a combination of different sensors and algorithms. Accordingly, the robot modulates its interactions with the children based on their attentional feedback to maximise the duration and the quality of their social engagement. The robot's behaviour (the tasks and activities performed to playfully engage with the patient) adjusts to the fluctuating attentional conditions by promoting the kinds of social behaviours and activities that maintain the child's attention for the longest time. This way, the robot can concretely help autistic children to autonomously develop their capacity for social engagement by training their capability to independently produce social attention. This form of adaptive intervention is beneficial in developing the child's social cognition skills. It makes them more robust and resilient to the stress associated with social engagement, under the assumption (subsequently verified by the children's parents and therapists) that a growing familiarity with social robots will also facilitate the interaction with the human adults, involving them into social activities that are familiar and comfortable although they require mutual social attention.

2.3.4 Components and methods

The following four tables summarise key information about participants, procedures, technology systems, and research outcomes.

¹⁹ "Most preprogrammed systems have fixed behaviour (i.e., they are unable to autonomously perform adaptable closed-loop interactions), are not tailored to the individual needs of patients, and cannot keep track of their recovery progress. For these reasons, semi-autonomous and adaptive robots are greatly needed to recognize the behavioural cues of children and respond accordingly. On the other hand, semi-autonomous adaptive systems, especially complex systems, need high-performance hardware, such as GPUs chips, to process real-time data and update interactions. Moreover, such fully autonomous and complex robots and systems are not yet reliable outside controlled research setups" (Idem).

Summary Table I - Participants

Characteristic	Description
Population	Children diagnosed with Autism Spectrum Disorder (ASD)
Sample Size	N = 11 children
Age Range	Maximum age: 16 (mean age: 9)
Gender	Male
Recruitment	Participants recruited through clinical/educational institutions (Al Ain Hospital)
Inclusion Criteria	ASD diagnosis, ability to engage in short interactive tasks with a robot
Exclusion Criteria	Severe ASD, inability to complete robot-assisted sessions, no verbal capability

Summary Table II - Procedures

Aspect	Description
Robot Used	NAO humanoid robot with additional mobile display
Intervention Style	Semi-autonomous robot interaction with minimal human intervention
Task Focus	Enhancing attention span and engagement of children with ASD
Interaction Setup	Robot presented emotional cues (via speech and gestures) and assessed child responses
Sessions	Multiple interactive sessions conducted individually with each child
Personalization	Adaptive system adjusted to child's attentional cues and preferences
Data Collection	Automated methodology for assessing attention: gaze tracking, response time, behavioural logging

Summary Table III - Robot Features and Technical Setup

Component	Description
Platform	NAO humanoid robot (SoftBank Robotics)
Embodiment	Humanoid design with speech, gestures, head movements, and visual display
Role	Delivered interventions, maintained engagement, monitored children's responses
Sensors	Vision-based system (for gaze/attention tracking), microphone (for vocal cues), tactile sensors
Attention Detection	Automated attention assessment based on gaze direction, response latency, interaction patterns
Personalization Algorithms	Adjusted difficulty, timing, and content of interventions based on individual child's performance
Intervention Style	Emotional cues (e.g., "that hurts" or "that feels good" feedback), playful gestures, social behaviours
Autonomy Level	Semi-autonomous – robot behaviour guided by pre-defined algorithms with minimal therapist intervention
Data Logging	Automated recording of attentional measures, behavioural responses, and session performance
Integration	Designed to complement therapist-led interventions, not replace them

Summary Table IV - Outcomes

Category	Findings
Attention	Robot-based methodology successfully measured and improved children's attention duration and quality
Engagement	Participants showed higher levels of engagement with robot than with baseline conditions
Social Behaviours	Children responded appropriately to robot's emotional cues (e.g., touch, gaze, speech)
Personalization Benefits	Adaptive interventions based on each child's responses were more effective than non-personalized interactions
Clinical Implications	Automated assessment reduces therapist workload, provides objective data, and complements traditional therapies
Limitations	Small sample size, short-term assessment only, results need replication in larger cohorts

2.3.4.1 Attention assessment factors – The study is based on a “novel numerical diagnostic assessment method that does not require any external camera or monitoring equipment”. Instead, it utilizes the image processing capabilities of the Nao robot and the capabilities of the mobile phone to detect patient attention cues. The attention score is continuously changing (incrementally) based on attention cues detected in real time. Attention assessment in the context of robotic interventions on autism can be modelled to include one or more of the following factors: (i) *Eye/face attention* – represents how much the individual is paying eye/face contact to the robot. This can be measured using a variety of measurement techniques, ranging from eye-pupil diameter measurement to image processing utilizing the robot's onboard camera; (ii) *Joint attention* – represents the ability of the individual to react to robot's motoric actions, like looking at the same object that the robot is pointing to. This can be measured using the same hardware used to measure attention with different data processing techniques or it can be measured separately using a motion tracking method; (iii) *Facial emotions* – detection of the facial reaction of the individual interacting with the robot. Detecting facial features can be used to detect emotions and other aversion and attention cues. A camera could be enough to detect facial features, and a depth camera can be used for better accuracy. The output of facial features detection helps in understanding the emotional state of the individual to relate it to other attention measures; (iv) *Sound/Linguistic response* – the ability to answer questions during within a certain timeframe; (v) *Imitation response* – the ability to imitate the robot's movements (using a matching similar not different from the one used in the avatar study).

The system we devised for this experiment used four out of the five sources of information, while the previous proof-of-concept study used all of the five. This is much more information compared to other studies of the same kind.

TABLE I. USED INCREMENT AND DECREMENT VALUES AFFECTING ATTENTION SCORE.

	Increment Value (points)	Decrement Value (points)
Attention	5	1
Joint Attention	15	1
Sound Response	Eq. (1)	5

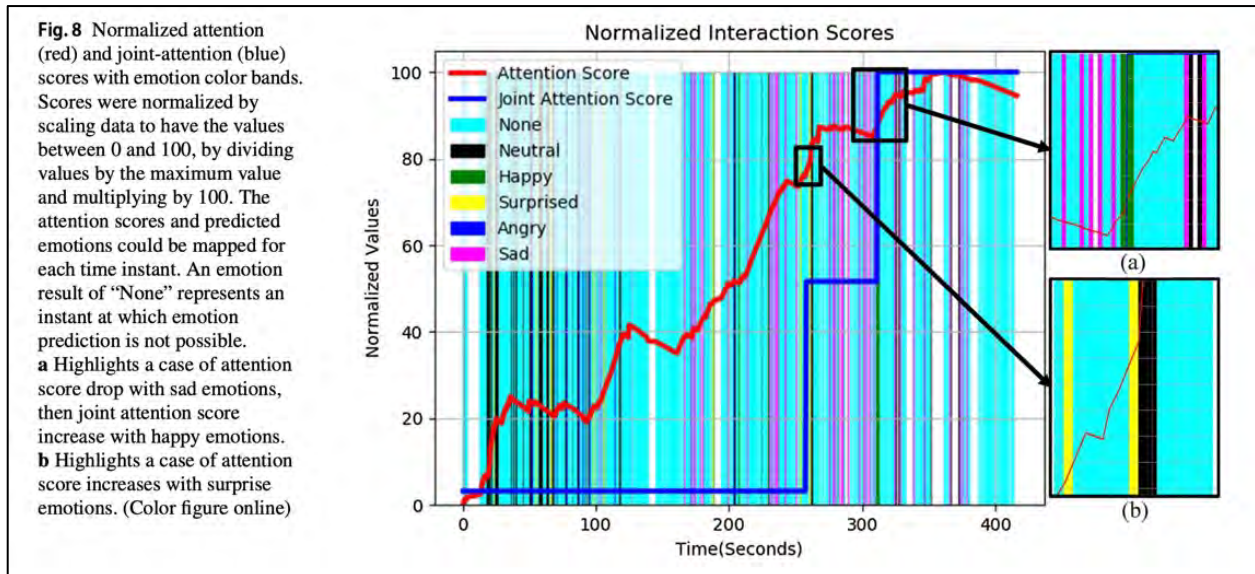
Also, most of the existing assessment techniques in robot intervention studies either use external hardware components not mounted on the robot, which limits portability, or a single attentional cue due to robot's onboard hardware limitations. Our study, on the contrary, integrates information from all these domains, therefore it provides a very robust, comprehensive, and accurate real-time indication of the subject's state of attention. At the same, we used very simple solutions, to maximize the amount of the information collected and its quality, making good use of the sensors available on Nao with the additional sensors embedded in a smartphone mounted on the robot's chest. Also, the system we devised uses an upgraded robot's onboard/mounted hardware to implement the automatic assessment of the relevant type of attention cues. Thus, our study introduces a simple and efficient numerical assessment system for robot intervention without the use of external monitoring equipment.

2.3.5 Measuring Attention via a Cumulative Score System

The scenario includes dialogues to interact with the participant to grab his/her attention and test his/her interaction capabilities. Sound responses, joint attention requests and robot movements are automatically called in the preset scenario parts.

2.3.5.1 Attention Score – This score part is designed to measure how the participant reacts to robot motoric actions and requests. For instance, if the robot talks about a box on the side and points towards it, the participant is expected to look towards the same box. The participant's head motion detection is simplified to detect the participant's left and right ears. If the participant looks to his/her right side, his/her left ear should appear to the robot. And the same applies for his/her left side and right ear. This detection is made using the same image frames retrieved from the mobile camera. If the detected participant's head direction

matches the current robot's head direction, the joint attention score is incremented by a set increment value; else, it is decremented by a set decrement value.²⁰

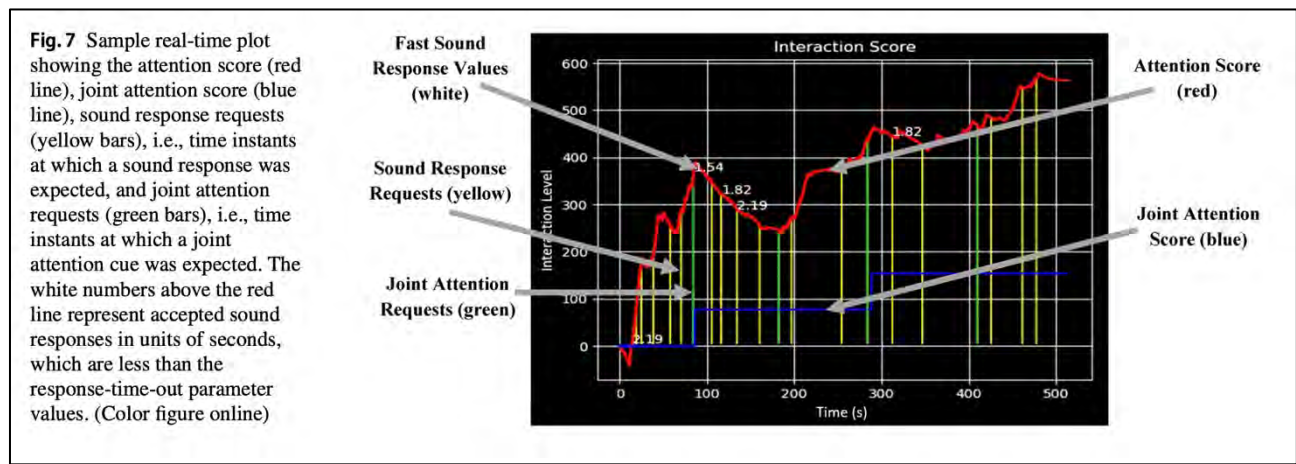


2.3.5.2 Joint Attention Score – The joint attention score measures the extent to which the patient's responses are synchronized with the robot's motoric actions and requests. For example, it detects whether the patient is looking toward the box that the robot is talking about and pointing towards. For simplicity, the head direction of the patient is detected by identifying his or her right and left ears. The camera detects the left or right ear of the patient depending on which way the face of the patient is turned during the interaction²¹.

²⁰ The camera of the mobile phone was configured as an IP camera using the IP camera mobile application utilized for face detection to produce and accumulate an attention score. Haar classifiers help detect the faces of patients from the image frames received from the mobile camera. If the patient is facing the front of the robot body, where the mobile phone is attached, the face of the patient is detected, the assessment system increments the attention score by 5 points. If the face of the patient is not facing the robot, then his or her face is not detected, and the assessment system decrements the attention score by 1 point. This score value is updated continuously in each iteration of the algorithm. This attention score facilitates comparison of the results from multiple interaction sessions for the same patient.

²¹ The detection algorithm uses the same video frames that are employed for face detection, which are retrieved from IP mobile camera. The joint attention score is incremented by 15 points if the head direction of the patient matches the expected direction in response to actions of the robot. Otherwise, if the patient does not respond as expected, the joint attention score can be decremented. The decrement value was set to zero in the following experiments. Joint attention score is updated only when a joint attention request is triggered in a scenario part. Since the attention score is designed to report on the general attention state of the patient, it is incremented by 10 points and decremented by zero points at joint attention events.

The interaction score part is designed to measure how the participant reacts to robot motoric actions and requests. For instance, *if the robot talks about a box on the side and points towards it, the participant is expected to look towards the same box*. The participant's head motion detection is simplified to detect the participant's left and right ears. If the participant looks to his/her right side, his/her left ear should appear to the robot. And the same applies for his/her left side and right ear. This detection is made using the same image frames retrieved from the mobile camera. If the detected participant's head direction matches the current robot's head direction, the joint attention score is incremented by a set increment value; else, it is decremented by a set decrement value.



2.3.5.3 Emotion detection score – The assessment system estimates the emotional state of the patient by detecting facial features. The results of the emotion detection module onboard the Nao robot were used to avoid an additional processing load on the system computer. Retrieving these results was timed in the attention assessment system loop to sync logged emotion estimation results with attention assessment score. This sync method avoids any effect on results' analysis. Emotion detection uses the onboard Nao head camera, and its results were accessed using Naoqi Python module. The system supports the detection of five possible emotions (happy, sad, angry, surprised, and neutral) by estimating the probability of each emotion at each instant. The system chooses the emotion with the highest probability, logs it, and represents it in the final results plot. Emotions are represented as colour bands, with a different colour for each emotion.

2.3.5.4 Sound response score – A mobile application called WO Mic turns the mobile phone microphone into an IP microphone, and client software on the computer connects the WO Mic as a sound input for a sound response module algorithm. This technique allows the operator to have a microphone monitoring the interaction room without affecting the sound response module functionality. The sound response algorithm starts calculating the delay time at the instant when the Nao robot finishes asking the question and stops at the instant

a sound peak is heard, which occurs when the patient starts speaking²². The faster the sound response, the bigger the increment in the attention score. The lowest response time, ideally zero, generates the maximum increment value, which is a parameter that can be preset. Sound response is updated only when sound response request is triggered in a scenario part. Sessions with patients would ideally be conducted in a quiet room without interruptions or noise. Five sound response parameters were considered.

2.3.6 Therapist–System Agreement

An important step to validate the methodology consisted in assessing how closely the automated system’s measurements of attention matched those of trained human therapists. The study reports an overall 82.4% therapist–system agreement, derived from the observation that, across 34 intervention sessions, the system’s attention assessments diverged from the therapist’s ratings in only six cases. Thus, 28 out of 34 comparisons yielded consistent outcomes, leading to an agreement rate of $28/34 \times 100 \approx 82.4\%$. Agreement was determined by whether the system and the therapist identified similar trends in attentional shifts (such as increases, decreases, or maintenance of attention) rather than requiring identical numerical values. Discrepancies were highlighted in Fig. 14 of the published paper, where mismatches are visually marked for ease of inspection.

The statistical approach here is descriptive in nature, as it relies on a *proportional accuracy measure* rather than advanced inferential statistics. The comparison method rests on two complementary strategies: (i) session-by-session trend analysis, whereby therapist and system ratings are compared over the same time windows, and (ii) visual validation of mismatches, using red markers to indicate divergences between the two sources of assessment²³. The validation method, while straightforward and transparent, neither estimates the likelihood that agreement exceeded chance levels nor quantifies the strength of the correlation between human and system scores.

²² The recorded delay time is saved as the sound response time and used to calculate the increment (*Incr*) in the attention score instances as a function of the detected response time (*RespTime*), preset maximum score for the response time parameter (*MaxRespScore*), and Response- time-out value:

$$Incr = MaxRespScore - \frac{MaxRespScore}{Response-time-out} \times RespTime. \quad (1)$$

²³ More sophisticated statistical measures—such as Cohen’s kappa for inter-rater reliability, correlation coefficients for scale alignment, or confidence intervals for generalizability— could have been used in the study, but they were not employed as deemed superfluous.

This approach carries several limitations. First, the sample size is relatively small (34 sessions across six participants), which restricts the robustness and generalizability of the findings. Second, the binary definition of agreement (match vs. mismatch) overlooks gradations of partial overlap or the possibility that therapist and system ratings may align in direction but differ in magnitude. Third, the reliance on therapist judgment as ground truth introduces potential subjectivity and variability, since clinical assessments are context-dependent and may not be perfectly reliable across raters. Finally, the absence of inferential testing means that the reported agreement remains a descriptive metric without statistical confirmation. Despite these limitations, the reported 82.4% agreement indicates a promising level of convergence between automated and human assessments: thus, it supports the feasibility of using robotic systems for attention monitoring in autism interventions while highlighting the need for more rigorous statistical validation in future studies.

Despite these limitations, the study is methodologically rigorous, and its findings are sufficiently significant to support the conclusions. First, the agreement measure, although descriptive, was applied consistently across all 34 sessions, ensuring internal validity through uniform comparison criteria. Second, the high level of convergence (82.4%) between therapist judgments and system outputs indicates more than incidental alignment, particularly given the diversity of participants and session contexts. The discrepancies observed were not random but interpretable, often arising in borderline cases where even human raters might reasonably differ, which suggests that the automated system captured attention trends in a clinically meaningful manner. Moreover, therapists acted as domain experts, providing an authoritative benchmark that—while subjective—remains the gold standard in autism intervention practice. Finally, the study combined numerical assessment with visual trend analyses (see Fig. 14 in the published paper), allowing for qualitative corroboration of the quantitative findings. Taken together, these elements provide a triangulated validation process that, while not exhaustive, is robust enough to establish the practical usefulness of the system in real-world therapeutic contexts.

2.3.7 An adaptive and interactive system

The interactive scenario is divided into sections, where each section contains dialogues, motions, and plays for a certain topic, such as greetings and getting to know each other, entertainment games and songs, and conversational questions and requests²⁴.

²⁴ A typical scenario starts with a greeting phrase such as “Hi” or “Hello.” The greeting is followed by interaction responses and questions, such as the following (Table 1):

(S1) General questions such as “How are you?”, “Do you know my name?”, “How old are you?”

Each section in the scenario has a numerical weight that is updated whenever the attention score is updated. This value increases or decreases by 0.2 points depending on whether the attention score increases ($attention\_trend + 1$) or decreases ($attention\_trend - 1$) at the time of using the corresponding section, as defined in Eq. (2):

$$W_{new} = W_{old} + (0.2 \times attention\_trend) \quad (2)$$

The section weights are stored individually and used in the following session to determine the number of phrases to use from each section, by multiplying the weight times the default (first) number of phrases to use. The weights in the first session are the same for all patients (default values). The minimum weight is zero, sections are not used, and the maximum value is the size of the section (number of phrases defined in the section). If the weight reaches the section size value, it means that the patient is interested in the section, and the operator is prompted to add more phrases to the section for the subsequent interactions. The scenario is defined as an Excel file with each section as a separate sheet. Scenarios are easy to edit, and they are updated as soon as a session starts. The scenario sections phrases definition includes the choice to call for sound responses, joint attention requests, and robot movements, in addition to text to be said by the robot.

Table 1 Scenario section titles

Label	S1	S2	S3	S4	S5	S6	S7
Title	General questions	Playing dialogue	Math questions	Colors learning	Emotions recognition	Food dialogue	Birthday singing
Label	S8	S9	S10	S11	S12	S13	
Title	Driving questions	Repeat words	Who is this person?	See that box	Football playing	Physical exercises	

(S2) Playing responses such as “Can we play together?”.

(S3) Math related questions such as: “Can you count?”, “How much is 1 plus 1?”.

(S4) Color questions such as: “What color do you like?”, “What is my color?”? “What is your shirt color?”

(S5) Emotions teaching such as: “Are you happy?”, “I feel sad”, “The face on my screen is excited”.

(S6) Food related responses such as: “Do you like chocolate?”, “I am hungry”, “What did you eat today?”.

(S7) Birthday singing activity such as: “When is your birthday?”, “Lets sing it together!”.

(S8) Driving mimicking such as: “Do you drive?”, “I drive very good”, “See how I drive, lets race!”.

(S9) Learning and repeating words such as: “I want you to repeat after me”, “Say small”.

(S10) Who is this questions for joint attention initiation such as: “Do you know who this on the right is?”.

(S11) See that box questions for joint attention initiation such as: “Do you see that box?”, “What is inside it?”, “Can you bring it here?”.

(S12) Football playing mimicking such as: “Do you play football?”, “I know how to shoot!”.

(S13) Stimulating physical exercises such as: “Stand up”, “Raise your right arm, like this”, “Where is your right ear?”, “Can you jump?”, “Raise your left hand”.

The scenario provides encouragement phrases as well such as “Great!”, “You are awesome”, “Wow!”, “No, try again”, and “I can’t hear you, talk louder!”. The mobile screen shows an image of an emotion (happy, surprised, excited, sad, or angry) that is relevant to each scenario phrase, presented as the robot’s emotional state.

2.3.8 Challenges and solutions

Assembling such a complex interactive and adaptive system inevitably involves facing complications of various nature. Some of the challenges I faced were greater than others but eventually all of them were effectively addressed thanks to the unique blend of interdisciplinary competences of our research team.

2.3.8.1 Adaptive calibration – The main challenge, which coincides with the main goal of the study itself, was to devise a system that gradually but steadily introduces ASD children to social interaction in a way that is robust (not casual or discontinuous), enduring (not short-lived), and pleasant for the child, without overwhelming, distressing, or confusing them. This requires carefully controlling both the quantity and the quality of the interaction. It could be difficult for a human to read the child's responses and adjust their behaviour with sufficient rapidity and precision, but the system we created was able to achieve this result using the adaptive features of machine learning and, most important, of the robot's embodiment.

Autistic children (probably even more than neurotypical children) love robots, attracted by curiosity, but tend to recoil from intense social interactions, repulsed by the overwhelming stimulation. The purpose of this study was to devise an interaction protocol capable to correctly balance the two opposite forces of attraction and repulsion, building on a dynamic system that modulates in real time (a) the stimulation aiming to produce curiosity for the robots and (b) the stimulation concerning the social nature of the interaction with the robot. The child's playful engagement with the robot is the result of the calibration of these two concurrent stimulations, supported by the specific embodied and interactive features of the robot. The dynamic character of the system depends on the robot's capability to adjust its behaviour in real-time to increase or decrease the social stimulation on the child based on the feedback provided by the child, i.e. whether the child shows a growing or diminishing level of social attention over time.

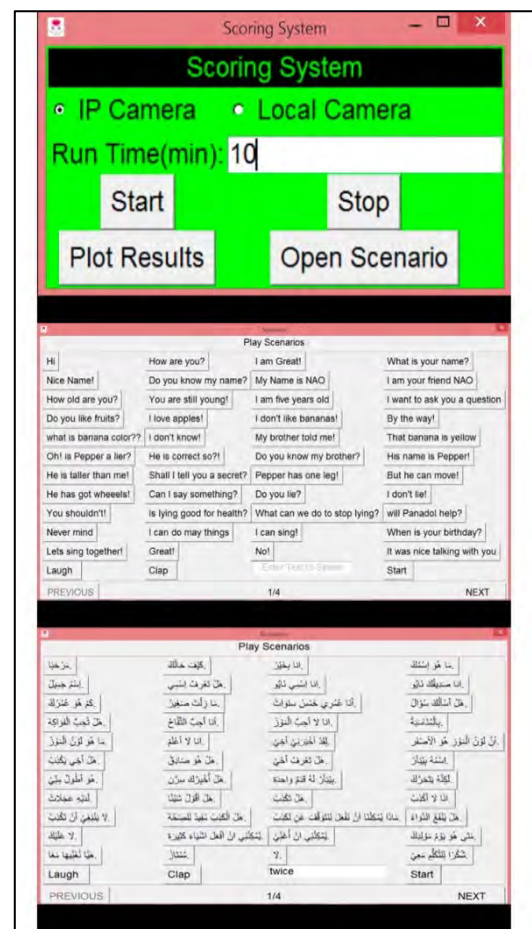


Fig. 3. Assessment System and Scenario GUIs. The first page in the GUI [a] is the assessment system control interface. It supports local and IP cameras. Plot Results will plot the latest session results. The second [b] and third [c] pages are the English and Arabic versions of the designed scenario, respectively. They can be accessed from Open Scenario button.

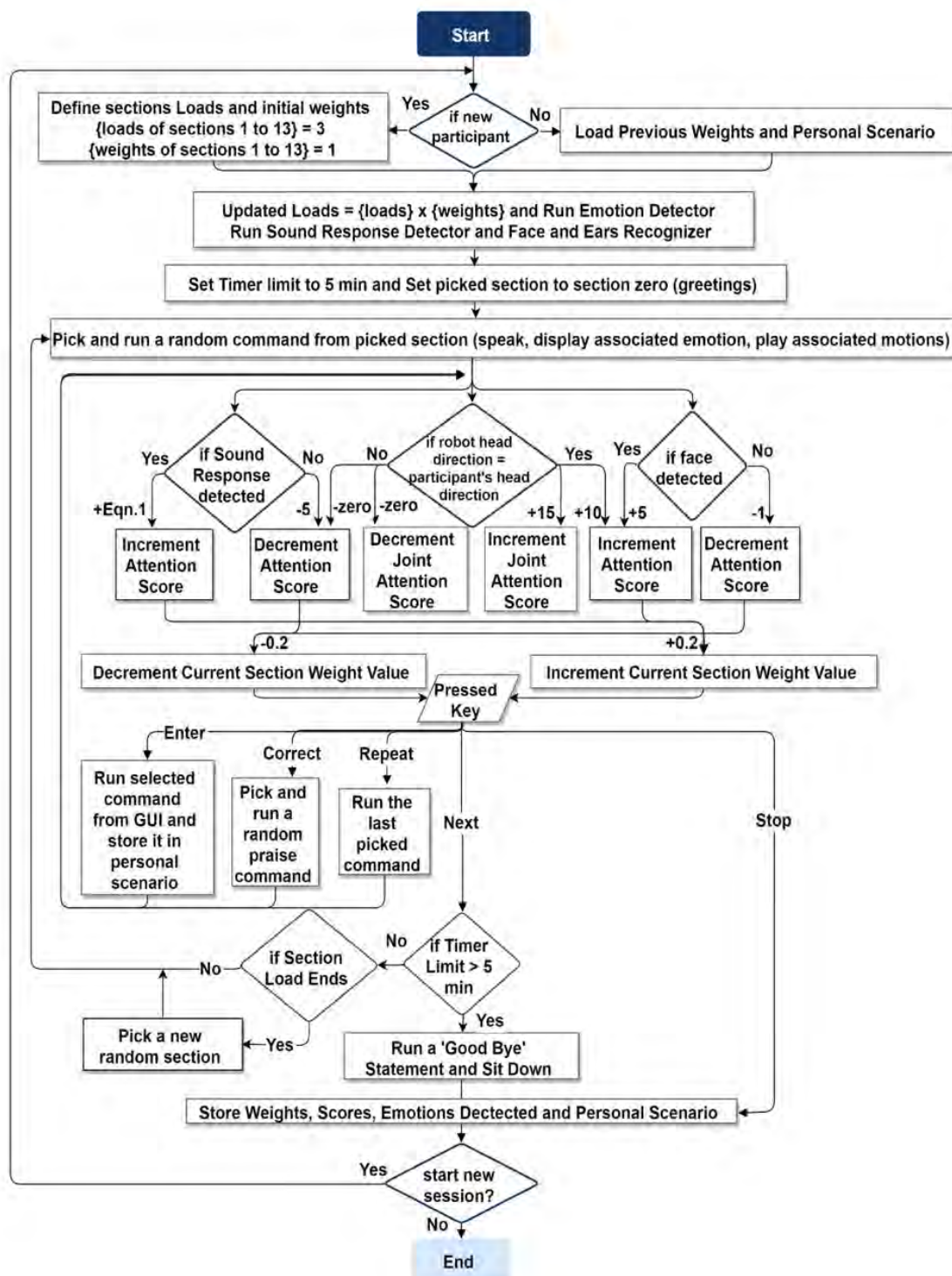


Fig. 5 Interaction and assessment system flow chart

Therefore, the kind of feedback we have tracked by means of the robot is the engagement level, represented as a vector that varies in real time and that, in our model, is based on the combination of all these three factors: social attention (head direction), emotional temperature (surprised and amazed facial expressions), verbal commentary (speech production). We selected these factors, and we chose to work with them, because they

manifest the state of engagement in a rather clear, overt, and unambiguous manner and also because they can be tracked relatively easily using sensors and machine learning technologies that already mature, sufficiently reliable, and quite easily accessible. The reading of the engagement level emphasizes the cumulative result of these three factors, without distinguishing between their individual contribution to the total. We were asked by referees why we decided to go this way, which seems to neglect situations in which attention, emotion and verbal response have different weights, underusing this information for making differential analyses.

Our answer is that, while the system we devised certainly allows different analyses of that kind and can, in principle, distinguish the varying pre-eminence of one engagement factors over the other two in changing conditions and scenarios, we deliberately decided to not use this feature in our study for two reasons. The first reason is that the three information channels were not used in our study to derive dependent variables, as the purpose of our study was not building an explanatory or predictive model of ASD children attention in varying conditions; rather, the three information channels were deliberately used as concurrent components of an input signal that at once affects and is affected by engagement level. This input serves to modulate a feedback activity in real-time, hence retroacting on the signal itself: like in other feedback systems of this kind (for example EEG-based neuro-feedback or heartbeat rate-based feedback), which can effectively use the available information to modulate the system's behaviour even when such information is "dirty" (incomplete or degraded), also in our system the quality of the signal and the identification of its distinct components are far less important than the behavioural outcome of the system as an adaptive whole. In other words, the system is outcome-driven, not aiming at producing accurate representations of the independent variables.

2.3.8.2 Variety – Our goal was to provide a sufficient number of interactive scenarios while ensuring that they were sufficiently different from each other to meet the varying tastes and preferences of the participants. The scenarios had to be designed not only to attract their attention but also to maintain it, with activities (questions & answers) that are sufficiently original to motivate the continuation of the interaction but also sufficiently predictable to allow children to adjust and feel comfortable. We got inspiration from similar protocols in the literature and the very important suggestions by the therapists we worked with. We also used our creativity and our intuition, as we become progressively more familiar with the task at hand and with the specific challenges involved in the relevant kind of HRI design.

2.3.8.3 Safety – Having verified that our interaction protocol did not present any physical or psychological risk for the children involved, the major concern was for the integrity of the Nao robot, which was occasionally overwhelmed by the excited responses of the children, who often wanted to touch it, hold it and interact with it in a number of playful yet extremely vigorous ways. We dealt with this risk by ensuring that the interaction was always supervised by qualified medical personnel and that proper distance from the robot was maintained.

2.3.8.4 Screen interface – Human faces represent a source of overwhelming stimuli from the point of view of a young ASD person, but stylized faces and cartoons are manageable and don't bring too much distress. The small display added to NAO provided a supplementary video interface: it is a mobile phone that shows emotions running a custom designed mobile application named Emotions Selector. A mobile magnetic holder was customized to fit on the chest-mount to carry the mobile screen. The mobile application was designed using the MIT App Inventor [12], it receives the emotion control data from the computer via socket communication over Wi-Fi. The application is made easy to use without complications other than specifying the computer's IP address to link it with the mobile automatically²⁵. Adding this form of facial expression was very important to communicate emotion to the children and to encourage them to interact because, contrary to some old theories, ASD children have emotions and feelings and are sensitive to the emotional expressions of others. Due to the inherent limitations of the robot Nao, which is designed move head and torso and illuminate its eyes, but has no way to express facial emotions, we had to use an additional interface to provide additional visual feedback to the children, presenting stylized faces with emotions. This interface was realized by a smartphone attached to the robot's chest by means of a plastic prosthesis designed by third parties and 3-D printed in our laboratory. We were initially worried that the double source of social information (the gaze of the robot, through Nao's own "eyes") in addition to the faces on the screen could generate ecologically invalid situations, but we soon realized that the children did not have the smallest problem interacting with the robot this way.²⁶

²⁵ The display is assembled with the aid of a custom chest-mount designed for the Nao robot, shown in figure, to minimize the changes in the moment of inertia due to the added weight. The used chest holder (chest-mount) is designed in our lab using Fusion 360 CAD Software to be a multipurpose tool for holding any object using up to four screws for rigidity. Velcro straps are used to fasten the holder from the back. It can be used by other researchers to carry a mobile, a camera, a laser sensor or any other item on the chest of Nao robot. Moreover, the USB socket on the head of Nao robot is used to charge the mobile phone. This technique maintains the robot's equilibrium ability; thus, robot's built-in features can be used during the diagnosis and therapy processes.

²⁶ On the other hand, we were also worried that the screen could be a potential source of distraction as, the children "are used to playing games with such devices, which may lead to drops in attention score. However, the issue occurred only at the beginning of the early sessions where the children were exploring the robot features and it never

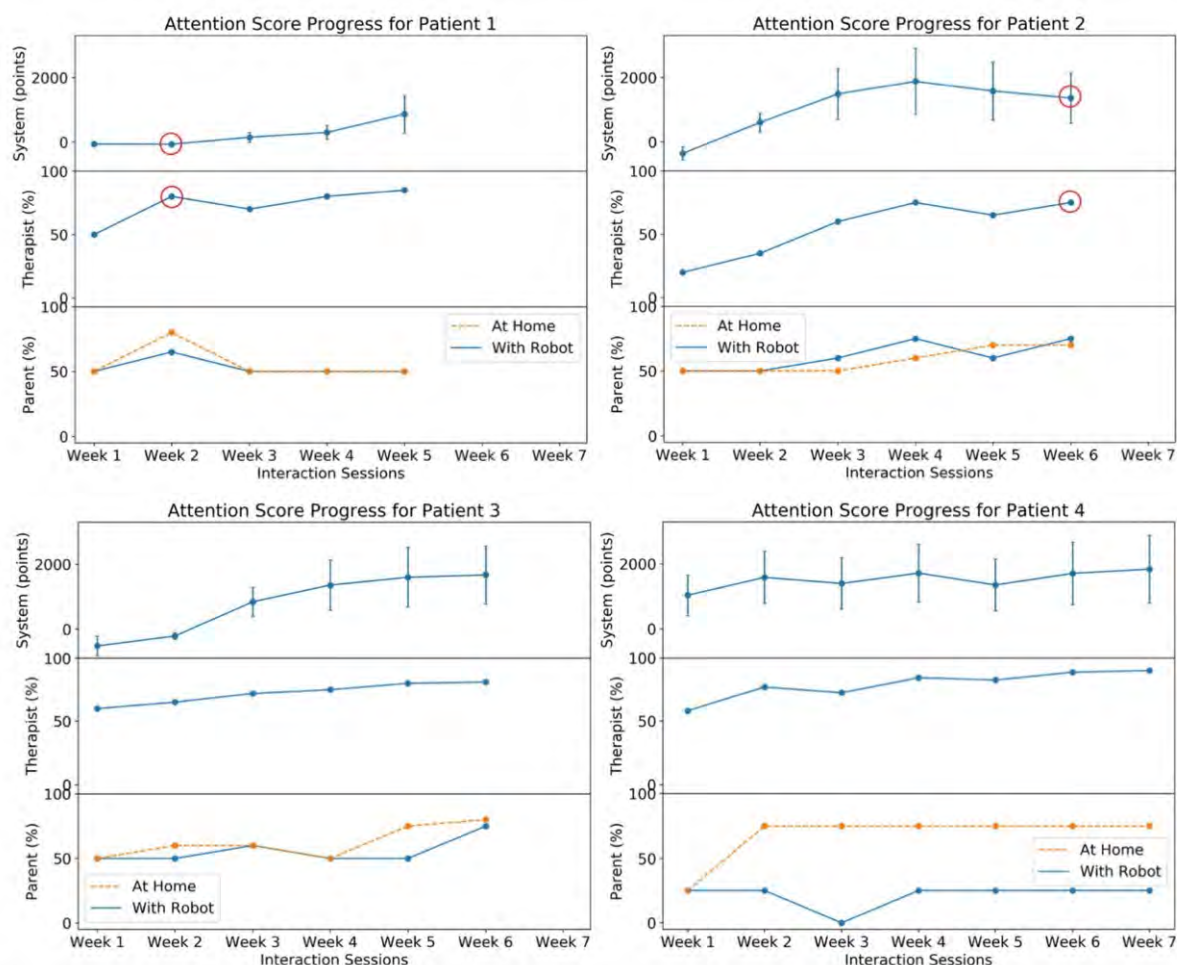
2.3.9 Outcomes, limitations, and future research

All in all, our study was successful on all fronts and produced some very encouraging results, providing an enduring contribution to the development of social robotics in the field of ASD studies.

2.3.9.1 Long-term interaction progress – Six out the 11 patients continued in a long-term study. Further experiments involving these six patients were performed during the following 7 weeks to examine their interaction progress and the variations in the patterns of preferred scenarios. Individual preferences for different scenarios emerged quite clearly from week 4. The results of the empirical sessions for all patients are shown in the figure below, which compares the progress in terms of attention scores, therapist assessment, and parent feedback. All six participants portrayed a trend of increasing attention scores. The therapist and system assessment trends are similar for most of the patients.

“The long-term interaction progress results show that robot intervention in autism therapy is highly beneficial for autistic patients. They enjoy the treatment sessions and give enough attention to learn or enhance a skill in every session. Noticeable changes in patient behaviours and skills took a few weeks, until the patients became familiar with the robot voice and moves. Breaking the ice took one or two sessions on average, until the patients interacted with the robot freely and excitement reached a steady level.”

had serious consequences. Moreover, it has been partly addressed by fixating the display in a way that it cannot be moved by the children.



2.3.9.2 Results – The results were extremely positive and rewarding: “The autistic patients built strong bonds with the robot as a friend who encouraged them to withstand the therapy time. One of the patients once brought some friends to meet the robot as a new friend. We observed that the therapists at the hospital went further by encouraging their patients to do energy-draining physical therapy exercises by offering a session with the robot as a reward. The therapists reported that the patients asked about the robot every day and looked forward to the day of the robot session.”

2.3.9.3 Human assessment – Parents reported that their children rehearsed their interactions with the robot at home, repeated some of the same Q&A routines. Five out of six parents believed that their children performed as well as at home or even better in some sessions. The therapist and system assessment trends are similar for most of the patients. The therapist’s evaluation shows a steady improvement, with all patients demonstrating an increased attention level over 7 weeks, matching the assessment generated by our algorithm. Only six mismatches in the attention score were found between the system and therapist assessments out of 34 sessions (82.4% accuracy).

2.3.9.4 Limitations – The current study supports only one-to-one interactions (to help the child focus on the robot and to facilitate data collection); with a small number of patients (more results may be revealed when it is applied to more patients); with mild ASD severity and at least minimal verbal response capabilities. Also, the mobile phone display on the robot was sporadically a source of distraction.

2.3.9.5 Future developments – Future studies building on our paradigm will have to pursue certain improvements and additional goals: increase the assessment accuracy; broaden the set of available types of interaction, possibly with the use of generative AI (which was not available when we designed this study); increase the avatar's degrees of freedom; use multiple external cameras to preserve the robot's mobility; involve a larger set of patients, including patients with more severe forms of autism. Most important, we think that using a virtual avatar rendered in computer graphics, instead of a mobile robot, could offer certain advantages and represent a different way to leverage the robot's "embodiment": the body shape and appearance (involving a varying number of polygons and more or less complex movement patterns) could be updated in real-time, based on the responses of the ASD patients, so to find the optimal sensory load and avoid the risk of overwhelming visual stimulation while maintaining a level of stimulation sufficient to maintain attention. Also, a "virtual avatar displayed on a tablet could reduce the costs involved in the use of embodied mobile robots, simplifying the use of the interactive system in the home setting." On the other hand, despite all these benefits, we suspect that a virtual avatar would not perform as effectively as a robot 'in flesh and bone', first of all because an image on a screen would never be as engaging as a physical one. That is why "we anticipate that the quality of the interaction with a virtual avatar might be inferior in terms of both quality and length to the interaction established by the children with a physically embodied robot" (idem).

2.4 Conclusions and significance for the field of HRI

Our study contributed to advance the field of robotic interventions on ASD children because it has validated the semi-autonomous robotics system that we designed, successfully documenting its remarkable efficacy. This efficacy is particular evident in two of the three dimensions of usefulness that I have highlighted earlier (section 2.2.2): the *assessment* and the *treatment* dimensions. We have collected sufficient qualitative and quantitative data to confirm that our system is very useful and competitive both in terms of diagnostic and therapeutic opportunities offered to the clinical personnel. This is due, first of all, to its capability to facilitate the collection and comparison of data over a long period of time: 8 weeks, which is significantly more than the average data collection period for these kinds of studies on autistic children. As the short-liveness and transitoriness of the beneficial effects produced by social robots is often indicated as a core limitation of robot-based therapeutic interventions, our capability to monitor the progress made by our participants over such a long period of time, and to confirm that this progress was constant and consistent over time, represents a major success. This success is due to the felicitous combination of heterogenous technological systems and interaction protocols that we put in place. The core feature of the AI systems that we used was their remarkable adaptivity: the intelligent capability to switch from one interactive scenario to another enabled us to keep alive the attention of the participants over a period much longer than usual, exploiting the system's capability to track the participants' attentional responses and quickly change HRI protocols based on the patterns of their engagement. Another element not to be underestimated is the complex and very robust attention measuring system that we devised and effectively applied: by combining a multiplicity of sensors and sampling techniques, it allowed us to track the level of social attention with great precision and stability using diverse information channels which allowed for a more holistic and ecologically-sound interpretation of the engagement level.

This explains how our study provided an original contribution to research, advancing the field in terms of both assessment and treatment. But what about the third dimension, concerning the “explanation” and the theoretical “understanding” of ASD? In other words, did our study provide any body of evidence in favour of some particular theories of autism over the others? I believe that the answer to this question should be positive. To begin with, our interactive system strongly corroborates the Sensory Overload Hypothesis (2.1.3.5), as it clearly shows that the social attention and the engagement of ASD children drastically

improve when the type and the amount of social interaction are kept under control, maintaining the optimal amount of stimulation (to solicit curiosity without overloading the child). Now, this is important in itself, but the theoretical power of Sensory Overload Hypothesis is limited, as it is not a fully-fledged explanatory model. Nonetheless, any confirmation of this Hypothesis indirectly builds in favour of the Weak Central Coherence Hypothesis, which explains overstimulation as an effect of the difficulty to integrate perceptual details within a coherent whole. The same result is also compatible with the Executive Disfunction Hypothesis, because the sensory overload can be explained as a difficulty of inhibiting a compulsive stimulus (it is the inhibition that normally allows neurotypical subjects to freely divert their attention away from a magnetic focus of attention).

More specifically, the Broken Mirror Hypothesis is particularly relevant in the context of our studies, and our two experiments corroborate the involvement of the mirror neurons of the pre-motor cortex in different ways. First, our exploratory study with the robotic avatar shows that imitation tasks, usually very challenging for ASD children, become much easier and fun when the context is not overwhelming, and the children are introduced gradually to it. This highlights that the root cause of the problem in imitational tasks is not ideational or conceptual but lies in the basic motor control and how the perception of goal-oriented action is mediated by one's own motor competences. Second, the study with our semi-autonomous interactive robotic system shows that the low level of interest and engagement often experienced by ASD children is not caused by their lack of curiosity, concern or intellectual understanding, but by the impossibility that they experience to attune to the perceptual and sensorimotor context. When this possibility is finally granted, in a way that is opportunely gradual and calibrated on the child's individual predispositions, their curiosity and engagement can be extended for weeks. Taken together, these findings reinforce the EAA for two important reasons: first, they highlight that the robot's embodiment (with its anthropomorphic, social, yet simplified and less overwhelming appearance) plays a crucial role in determining how the ASD children perceive the task and how confidently they can approach it; secondly, they highlight that the child's embodiment is crucial too, as the cognitive processes involved in this child-robot interaction are mediated by the quality and the quantity of sensorimotor experience, not inferential or conceptual processes (as postulated by ToM and other cognitivist accounts of ASD). Ultimately, the outcomes of our study with humanoid robots support ECT, validating the idea that cognition is fundamentally and not only instrumentally embodied, and indicating an important way to apply its tenets in the clinical and educational domains.

Chapter 3 – The Embodiment of Ethical Skills. Moralizing Attention and Virtue Cultivation

*This chapter explores the variety of human attitudes towards robots and their ethical implications, investigating how attention can have a moralizing and disciplining function when effectively embodied. Inspired by our “ATTA-bot model”, and dedicated to **CP5** and **CP6**, the first part of the chapter distinguishes four different human attitudes towards robots, discussing the role played by Trust, Acceptance, and Tolerance in HRI and acknowledging the unavoidable (but not necessarily detrimental) impact of Anthropomorphism on adoption propensity. The second part, dedicated to **CP7**, **CP8**, and **CP9**, expounds our “Virtuous Robotics” approach to HRI, discussing the advantages and risks of anthropomorphic robots from a virtue ethics perspective. The third part of the chapter, dedicated to **CP10** and **CP11**, reviews the virtuous applications of social robotics in the classroom environment, discussing the case of humanoid robots used to monitor and discipline students.*

Core Publication #4 [CP4]

Alnajjar F.S., Cappuccio M.L, Renawi A., Mubin O., Kiong Loo C. (2020),
“Personalized Robot Interventions for Autistic Children:
an Automated Methodology for Attention Assessment”
International Journal of Social Robotics 13(1), Springer Verlag.

Reproduced with permission from Springer Verlag. All rights reserved.



Personalized Robot Interventions for Autistic Children: An Automated Methodology for Attention Assessment

Fady Alnajjar^{1,2} · Massimiliano Cappuccio^{1,3} · Abdulrahman Renawi¹ · Omar Mubin⁴ · Chu Kiong Loo⁵

Accepted: 25 February 2020
© Springer Nature B.V. 2020

Abstract

We propose a robot-mediated therapy and assessment system for children with autism spectrum disorder (ASD) of mild to moderate severity and minimal verbal capabilities. The objectives of the robot interaction sessions is to improve the academic capabilities of ASD patients by increasing the length and the quality of their attention. The system uses a NAO robot and an added mobile display to present emotional cues and solicit appropriate emotional responses. The interaction is semi-autonomous with minimal human intervention. Interaction occurs within an adaptive dynamic scenario composed of 13 sections. The scenario allows adaptive customization based on the attention score history of each patient. The attention score is autonomously generated by the system and depends on face attention and joint attention cues and sound responses. The scoring system allows us to prove that the customized interaction system increases the engagement and attention capabilities of ASD patients. After performing a pilot study, involving 6 ASD children, out of a total of 11 considered in the clinical setup, we conducted a long-term study. This study empirically proves that the proposed assessment system represents the attention state of the patient with 82.4% accuracy.

Keywords Autism · Robotics · Assessment · Attention · Therapy

1 Introduction

Social robotics research is concerned with the development of robotic companions. The existing social robots can assist people in simple everyday tasks and provide entertainment through basic forms of verbal and behavioral communication. In addition to serving and amusing, could robots also be used to improve the social abilities of humans, training

their capacity for attention and interaction? This question is related to a much broader issue: is it possible for an individual to learn from robots to establish richer relationships with other humans, fulfilling their inherent empathic potential? These questions are particularly pressing for the individuals who are prevented by clinical factors from establishing typical social relations with others—for example autistic subjects who experience social attention as effortful and potentially upsetting.

As we believe that these questions can be answered positively, we developed and tested a robot specifically designed to train the social attention abilities of children diagnosed with autism spectrum disorder (ASD). Our research was motivated by the awareness that social interaction and communication skills significantly impact human development, learning, and well-being. The lack of these elements can hinder individuals from being successfully integrated into a complex society. Children with ASD are prevented from developing these social skills and the ability to respond to them [1]. ASD symptoms are usually exhibited at early ages. Neurological accounts often highlight that parts of the social brain systems of ASD children are hyporesponsive to social stimuli and gaze cues. This hyposensitivity is the cause of

✉ Fady Alnajjar
fady.alnajjar@uaeu.ac.ae

¹ College of Information Technology, United Arab Emirates University, 15551 Al Ain, United Arab Emirates

² Intelligent Behavior Control Unit, RIKEN Center for Brain Science (CBS), Nagoya, Japan

³ Values in Defense and Security Technology, School of Engineering and Information Technology, University of New South Wales at Australian Defense Force Academy, Canberra, ACT, Australia

⁴ School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia

⁵ Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

their unresponsiveness to the social signals of other individuals and the related difficulties in perceiving the eyes of other people as socially salient, which is why ASD children rarely establish eye contact [2]. Avoiding eye contact leads to impaired social attention [3] and, hence, difficulties in communication and interaction with others, which may result in severe academic and social problems. Owing to the continuous increase in the number of children diagnosed with ASD worldwide, the growing related social costs, and the absence of universally established diagnostic and therapeutic protocols, effective treatment of ASD is considered a public health emergency and an open research question [4].

Traditional ASD assessment focuses on multiple parameters, such as social communication and interactions, as well as restricted and repetitive, inflexible behaviors [5]. Attention (manifested through bodily cues like eye contact and interest in a voice, etc.) is one of the most commonly considered factors in the assessment of ASD through the identification of its symptoms [6]. Traditionally, assessment can only be performed only by licensed experts. It is demanding in terms of time and effort as it requires a complete history of symptoms to be obtained through systematic interviews with parents and patients. Moreover, although early intervention is a key factor for treatment, the younger the patient, the harder it is to accomplish the assessment [5]. Assessment is typically performed at the time of enrollment and repeated during long-term therapy to measure progress. If assessments could be repeated more frequently to trace the severity and development of symptoms [6], then therapists could implement more efficacious diagnostic and therapeutic options, and patients and their parents would be better motivated to withstand long, repetitive therapy sessions. This approach, of course, requires proportional commitments of time and effort by trained experts.

Currently, interventions for children with ASD primarily involve various methods for teaching patients the basic social skills through repetitive behavioral training [7]. During intervention, a therapist may utilize several tools and stimuli, such as toys and game-like activities, depending on the interests of the child and the skills to be tested [8]. One of the most important preconditions for the success of these interventions is the knowledge and availability of the kinds of stimuli that usually draw the attention of the child and increase his or her engagement over time, creating more favorable conditions for teaching social skills. Due to the heterogeneity of ASD symptoms among children, the intervention strategies and stimuli that are valid for one child cannot necessarily be applied to another. Individual customization may make intervention design challenging for therapists. Professional therapists certainly have the training and scientific knowledge necessary to promote the development of social skills in ASD patients. Therapists can adjust their behaviors in real

time to maximize the benefits to the patients. Two aspects of this process, however, remain problematic.

First, like any other human, a therapist is an inherently complex interactant. Their interaction with the patient inevitably involves the social information conveyed through facial expressions, gestures, posture, personality, attitude, etc. These irreducible complexities are not handled with equal ease by all ASD patients, and in most cases, they represent an inevitable barrier, probably due to the difficulty patients have processing high volumes of social information [9].

Second, the adaptability of the therapist to the cognitive and emotional characteristics of patients is limited by their number; the diversity of their cultural and personal backgrounds; and their motivation, language, and clinical condition. Challenges related to cultural and linguistic diversity are prominent in multi-ethnic countries such as the UAE, where expatriates of Bengali, Malayalam, Somali, Tagalog, Tamil, Hindi, and Urdu background (i.e., expatriates whose mother language is neither English nor Arabic, the two official languages of the country) represent more than 88% of the population [10]. As each human therapist inevitably has his or her own linguistic and cultural background, it might prove difficult at times to preserve adaptivity and intervention customization for all patients, due to the inherent difficulty of implementing the same diagnostic approaches and intervention protocols irrespective of the individual differences.

The developments in artificial intelligence and robotic technologies have proven promising for stimulating interactions with ASD children [11, 12] and performing more frequent assessments [13]. Robots fill the gap between conventional human therapy and child toys [14] and can perform endless repetitions without boredom, eliminating the concerns to training intensification. Anthropomorphic robots are designed to reproduce human features, behaviors, and emotions while simplifying their informational complexity, thereby reducing the cognitive and emotional burden and decreasing the possible stress for the patient. Consequently, it is believed that robots can improve the quality and length of engagement during social interaction and increase the possibility of stimulating the patients' social and cognitive abilities [15]. Crucially, robots can also automatically conduct score-like assessments [13, 16]. Robots continuously collect information that is relevant to diagnosis, which is useful for building retrievable databases of patient assessments and interaction histories. Such databases can be used to guide therapists in the personalization of their interactions and assessments. Compared with traditional assessments, the assessment methods augmented through robotic means allow for more comprehensive and articulate tracking of greater numbers of patients with more heterogeneous individual situations and needs [17]. Such personalization can be applied locally, within the same team or clinic, or globally as inte-

grated clinical dataset sharable by professionals in different places through interoperable systems [18].

Current robotics techniques, however, still suffer from various limitations. Most preprogrammed systems have fixed behavior (i.e., they are unable to autonomously perform adaptable closed-loop interactions), are not tailored to the individual needs of patients, and cannot keep track of their recovery progress [19]. For these reasons, semi-autonomous and adaptive robots are greatly needed to recognize the behavioral cues of children and respond accordingly [20]. On the other hand, semi-autonomous adaptive systems, especially complex systems, need high-performance hardware, such as GPUs chips [18], to process real-time data and update interactions. Moreover, such fully autonomous and complex robots and systems are not yet reliable outside controlled research setups [21–23].

In the present study, a simple autonomous assessment system based on attention cues was created and deployed, combined with an enhanced adaptive semi-autonomous interaction system based on patient interests. Both systems were implemented in an interactive autonomous humanoid robot. The function of the robot was to increase the attention and engagement levels of the patients during interactive sessions. Increasing the attentional and interactive capabilities of ASD patients has the potential to enhance their academic functioning by gradually habituating them to socially interactive sessions of increasing length. The proposed approach utilizes simplified hardware with some upgrades to the onboard hardware of the robot to target multiple interaction and attention cues simultaneously. This technique can serve as a useful form of ASD intervention to facilitate adaptive interactions with patients based on their status while involving minimal subjective biases. In this study, we empirically tested the proposed system on a group of ASD children. As shown in our report, the empirical results are promising.

The paper is organized as follows: Sects. 2 and 3 provide thorough descriptions of the proposed assessment and therapy system, including the experimental setup and procedures. Sections 4 and 5 analytically present the study results and discuss their implications in the context of the research objectives. Section 6 concisely restates the study methods and main findings and addresses the scope for future research.

2 Methodology

2.1 Participants

The participants in the study were 11 male patients diagnosed with ASD of mild to moderate severity (Childhood Autism Rating Scale [24]—which is essentially a diagnostic scale; CARS2 of 30–36.5) who were under the age of 16 years, with a mean age of $9.03 (\pm 2.56)$ years. Only patients with ver-

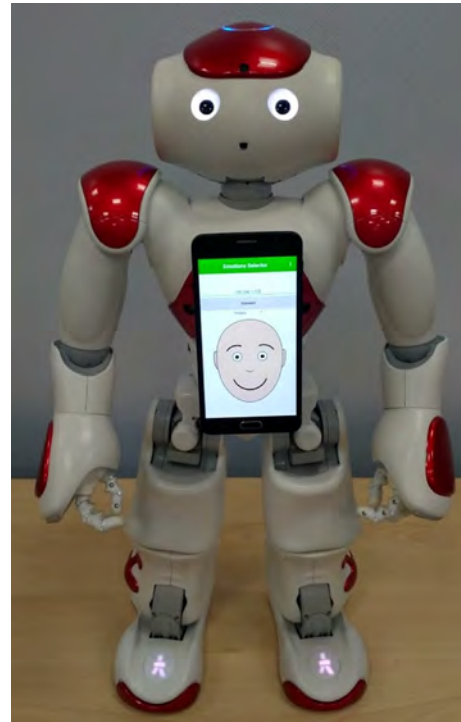


Fig. 1 NAO robot with a mobile phone attached to a chest holder to show the Emotions Selector mobile application

bal response capabilities were considered, as the participants were preliminarily asked to understand verbal communication and respond with yes or no at least. All patients were recruited from Al Ain Hospital, Al Ain city, UAE. This study was approved by the Social Sciences Research Ethical Committee of UAE University (Al Ain, UAE) and Al Ain Hospital (Al Ain UAE). The parents of the children were asked to sign parental consent letters.

2.2 System Design

2.2.1 NAO Robot with a Chest-Mounted Mobile Phone

The proposed approach to robot intervention in autism therapy and diagnosis uses a specifically designed NAO humanoid robot (Aldebaran/Softbank NAO robot) with an additional display on its chest to show facial features. The modified robot is shown in Fig. 1.

We designed a custom NAO robot chest-mount to firmly hold any object at the lower chest region [25], as shown in Fig. 2. The added weight barely affects the center of mass of the robot body, hence the changes in the moment of inertia are negligible. This method helps maintain the equilibrium capability and mobility of the robot, so that built-in interaction features can be employed to draw attention throughout the diagnosis and therapy processes. A mobile phone is used as the added display. To carry the mobile phone, a customized

Fig. 2 **a** Custom-designed NAO chest holder side view, showing the attached standard mobile holder. **b** NAO chest holder rear view, showing the added Velcro strap for fastening



mobile holder was used to fit the phone on the designed chest mount. This model of chest mount was fully designed and 3D printed in our lab. It is a rigid multipurpose holder with four slots for tightening screws. The holder is fastened from the back side using Velcro straps for greater rigidity. This holder can be used to carry a mobile, camera, laser sensor, or any other helpful item on the chest of the NAO robot for navigation or interaction purposes. We have made the chest-holder model files available on the GrabCad website for research use: <https://grabcad.com/library/nao-robot-chest-holder-chest-mount-1>.

The mobile phone displays emotions using a custom designed mobile application called “Emotions Selector”. Emotions Selector was specially designed with the aid of App Inventor [26]. It accepts control messages to switch between emotion photos as a single-character data from the computer via a TCP socket connection over WiFi. The application is simple and easy to use because it employs non-technical operators. It only asks for the IP address of the computer to establish a connection with the mobile device.

2.2.2 Attention Assessment System

Developing accurate and versatile methods for detecting negative child behaviors is important for minimizing distress and facilitating effective human–robot interactions. This paper presents a novel simple numerical diagnostic assessment method that does not require any external camera or monitoring equipment. It utilizes the naoqi image processing capabilities of the NAO robot and the capabilities of the mobile phone to detect patient attention cues and generate numerical measures. All attention cues are detected and updated in every iteration of the system algorithm, running at average speed of 1 Hz, and each score is updated based on certain criteria as discussed below.

Attention Score The camera of the mobile phone was configured as an IP camera using the IP camera mobile appli-

cation utilized for face detection to produce and accumulate an attention score. Haar classifiers [27] help detect the faces of patients from the image frames received from the mobile camera. If the patient is facing the front of the robot body, where the mobile phone is attached, the face of the patient is detected, therefore it is assumed that he or she is paying attention to the robot. Thus, the assessment system increments the attention score by 5 points. If the face of the patient is not facing the robot, then his or her face is not detected and the assessment system decrements the attention score by 1 point. The operator can change the preset increment and decrement values so that each patient has individual settings. This score value is updated in each iteration of the algorithm continuously. This attention score facilitates comparison of the results from multiple interaction sessions for the same patient. Hence, the increment and decrement parameters are not changed between sessions for the same patient.

Joint Attention Score The joint attention score measures the extent to which the patient’s responses are synchronized with the robot’s motoric actions and requests. For example, it detects whether the patient is looking toward the box that the robot is talking about and pointing towards. For simplicity, the head direction of the patient is detected by identifying his or her right and left ears. The camera detects the left or right ear of the patient depending on which way the face of the patient is turned during the interaction. The detection algorithm uses the same video frames that are employed for face detection, which are retrieved from IP mobile camera. The joint attention score is incremented by 15 points if the head direction of the patient matches the expected direction in response to actions of the robot. Otherwise, if the patient does not respond as expected, the joint attention score can be decremented. The decrement value was set to zero in the following experiments. Joint attention score is updated only when a joint attention request is triggered in a scenario part. Since the attention score is designed to report on the general

attention state of the patient, it is incremented by 10 points and decremented by zero points at joint attention events.

Sound Response A mobile application called WO Mic turns the mobile phone microphone into an IP microphone, and client software on the computer connects the WO Mic as a sound input for a sound response module algorithm. This technique allows the operator to have a microphone monitoring the interaction room without affecting the sound response module functionality. The sound response algorithm starts calculating the delay time at the instant when the NAO robot finishes asking the question and stops at the instant a sound peak is heard, which occurs when the patient starts speaking. The recorded delay time is saved as the sound response time and used in Eq. (1) to calculate the increment (Incr) in the attention score instances as a function of the detected response time (RespTime), preset maximum score for the response time parameter (MaxRespScore), and Response-time-out value:

$$Incr = MaxRespScore - \frac{MaxRespScore}{Response-time-out} \times RespTime. \quad (1)$$

From the equation, the faster the sound response, the bigger the increment. The lowest response time, ideally zero, generates the maximum increment value, which is a parameter that can be preset. Sound response is updated only when sound response request is triggered in a scenario part. The value of MaxRespScore used in the subsequent experiments was 30 points.

Some sound response parameters must be preset to account for the nominal speaking volume of the patient and room noise level. However, sessions with patients would ideally be conducted in a quiet room without interruptions or noise. The sound response parameters are as follows:

- **Threshold-peak-level** the sound peak predefined threshold level, if detected by microphone, initiates sound response module and starts counting the response time in seconds. The module assumes that the patient has started speaking at that instant. This parameter was set to 2000 Hz in most cases, and adjustable by the operator based on the experimental environment and patients voice level.
- **Response-time-out** maximum time response value allowed. If the response time exceeds this value and continues counting, the module resets and ignores the rest. Operators should assign high values for patients with some speech deficiencies. This quantity was set to 2.5 s in our case.
- **Text-is-done-time-out** maximum waiting time for the robot to acknowledge that the patient has finished speaking. The module resets if the robot fails to acknowledge that the patient has finished speaking and this timeout value is

reached. This timeout value check catches the robot timeouts, instants when NAO robot hangs and fails to send text-is-done feedback, and it should be set longer than the speech time of the longest expected text. This parameter was set to 5 s in our case.

- **Sync-delay** a small delay time to compensate for the delay time of the employed mobile phone microphone, mainly due to WiFi connection delay, to sync it with sound response module timing. This quantity can be found using an external observer microphone connected to an external computer. First, the employed mobile phone, the system computer speaker and the external microphone are placed very close to each other. Second, the system computer speaker is set to echo the sound received from the mobile phone microphone over WiFi, mobile phone position may be slightly changed to avoid a sound loop. Third, the external microphone starts recording and a peak sound is made, a single table knock for instance. Finally, the recorded sound file on the external computer is used to plot the sound recorded and determine the delay time between the sound peak made and its echo from the speaker of the system computer. This parameter was found as 0.25 s in our case.

Emotion Detection The assessment system estimates the emotional state of the patient by detecting facial features. The results of the emotion detection module onboard the NAO robot were used to avoid an additional processing load on the system computer. Retrieving these results was timed in the attention assessment system loop to sync logged emotion estimation results with attention assessment score. This sync method avoids any effect on results' analysis. Emotion detection uses the onboard NAO head camera, and its results were accessed using naoqi Python module. The system supports the detection of five possible emotions (happy, sad, angry, surprised, and neutral) by estimating the probability of each emotion at each instant. The system chooses the emotion with the highest probability, logs it, and represents it in the final results plot. Emotions are represented as color bands, with a different color for each emotion.

2.2.3 Adaptive Dynamic Scenario and Weights

A dynamic interaction scenario is employed, which depends on the real-time scores of the assessment system to tune the interactions based on the previous interaction results. This method helps maintain interaction and assessment sessions by reducing human intervention in the scenario flow. Moreover, it allows the operator to have control, although limited, by interfering with operator-defined robot responses added to the interaction in real time to ensure the meaningfulness and interactivity of the scenario.

Table 1 Scenario section titles

Label	S1	S2	S3	S4	S5	S6	S7
Title	General questions	Playing dialogue	Math questions	Colors learning	Emotions recognition	Food dialogue	Birthday singing
Label	S8	S9	S10	S11	S12	S13	
Title	Driving questions	Repeat words	Who is this person?	See that box	Football playing	Physical exercises	

The scenario is divided into sections, where each section contains interaction dialogues, motions, and plays for a certain topic, such as greetings and getting to know each other, entertainment games and songs, and conversational questions and requests. A typical scenario starts with a greeting phrase such as “Hi” or “Hello.” The greeting is followed by interaction responses and questions, such as the following (see also Table 1):

- (S1) General questions such as “Where have you been?”, “How are you?”, “Do you know my name?”, “How old are you?”
- (S2) Playing responses such as “Can we play together?”.
- (S3) Math related questions such as: “Can you count?”, “How much is 1 plus 1?”.
- (S4) Color teaching questions such as: “What color do you like?”, “What is my color?”, “What is your shirt color?”
- (S5) Emotions teaching such as: “Are you happy?”, “I feel sad”, “The face on my screen is excited”.
- (S6) Food related responses such as: “Do you like chocolate?”, “I am hungry”, “What did you eat today?”.
- (S7) Birthday singing activity such as: “When is your birthday?”, “Lets sing it together!”.
- (S8) Driving mimicking such as: “Do you drive?”, “I drive very good”, “See how I drive, lets race!”.
- (S9) Learning and repeating words such as: “I want you to repeat after me”, “Say small”.
- (S10) Who is this questions for joint attention initiation such as: “Do you know who this on the right is?”.
- (S11) See that box responses for joint attention initiation such as: “Do you see that box?”, “What is inside it?”, “Can you bring it here?”.
- (S12) Football playing mimicking such as: “Do you play football?”, “I know how to shoot!”.
- (S13) Stimulating physical exercises such as: “Stand up”, “Raise your right arm, like this”, “Where is your right ear?”, “Can you jump?”, “Raise your left hand”.

The scenario provides encouragement phrases as well such as “Great!”, “You are awesome”, “Wow!”, “No, try again”, and “I can’t hear you, talk louder!”. The mobile screen shows an image of an emotion (happy, surprised, excited, sad, or angry) that is relevant to each scenario phrase, presented as the robot’s emotional state.

Each section in the scenario has a numerical weight that is updated whenever the attention score is updated. This value increases or decreases by 0.2 points depending on whether the attention score increases ($attention\_trend = +1$) or decreases ($attention\_trend = -1$) at the time of using the corresponding section, as defined in Eq. (2):

$$W_{new} = W_{old} + (0.2 \times attention\_trend) \quad (2)$$

The section weights are stored individually and used in the following session to determine the number of phrases to use from each section, by multiplying the weight times the default (first) number of phrases to use. The weights in the first session are the same for all patients, and they are the default values. The minimum weight is zero, sections are not used, and the maximum value is the size of the section (number of phrases defined in the section). If the weight reaches the section size value, it means that the patient is interested in the section, and the operator is prompted to add more phrases to the section for the subsequent interactions.

The scenario is defined as an Excel file with each section as a separate sheet. Scenarios are easy to edit, and they are updated as soon as a session starts. The scenario sections phrases definition includes the choice to call for sound responses, joint attention requests, and robot movements, in addition to text to be said by the robot. The section titles are shown in Table 1.

2.2.4 Control Interface

We reduced the need for human intervention to make robot-aided sessions more easily replicable in clinical and domestic environments, as reproducibility is essential to make the system usable by therapists and parents without technical supervision.

The interface allows the operator to control the robot as shown in Fig. 3a. It is designed to be operator friendly and simple for operators with no programming experience. Pre-configuration is performed once by setting the parameters in the “config.ini” file. The operator should select the patient name (from the history of children enrolled) or input a new name to be registered and select the language of interaction (English or Arabic in the current version) and which camera to use (an IP camera or a local camera, if a wired camera is

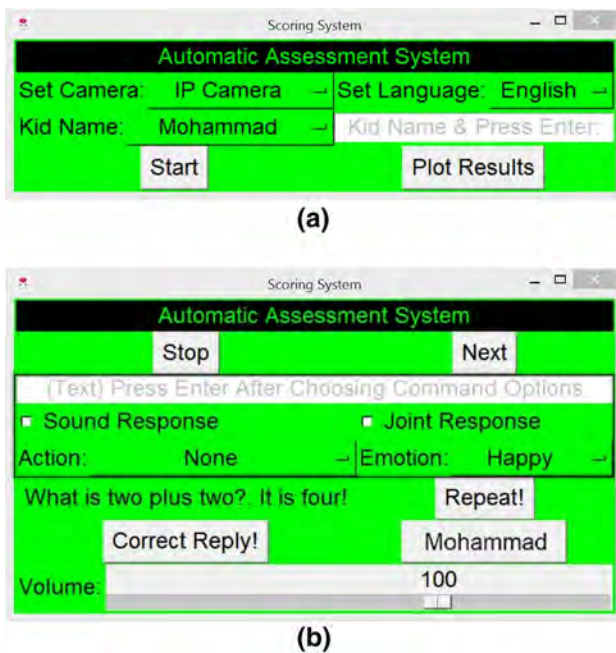


Fig. 3 **a** Scoring system control interface in idle state; no interaction running. The “Plot Results” option plots recent interaction results. **b** Interface during interaction. “Stop” stops the interaction system then saves and plots the results. “Next” goes to the next scenario part. Highlighted box contains operator-defined response definition tools: “(Text)” is to write a text to be said by the robot. All written texts here are saved and can be selected later from a selection lists when clicking into (Text). “Sound Response” is a check box to trigger sound response or not, “Joint Response” is check box to trigger joint response or not. “Action” is to select action to be taken by the robot from a predefined drop-down list of available actions and motions (move hand, point to right, point to left, shoot ball, sing birthday, etc.). “Emotion” is to select emotion photo to be shown by Emotions Selector mobile application from a predefined drop-down list of available emotion photos (Happy, Sad, Neutral, Excited, Crying, Surprised). “Repeat” repeats the last scenario part, which phrase text is shown to the left of “Repeat” button. “Correct Reply!” triggers a random praise phrase. “Mohammad”, or current participant name, triggers calling current participant by his/her name. “Volume:” a slide to simultaneously change the robot speaking volume. (Color figure online)

connected). The system generates a directory with the name of the child in the results folder for the individual data. Then, the operator starts the interaction by pressing “start,” and the interface changes as shown in Fig. 3b.

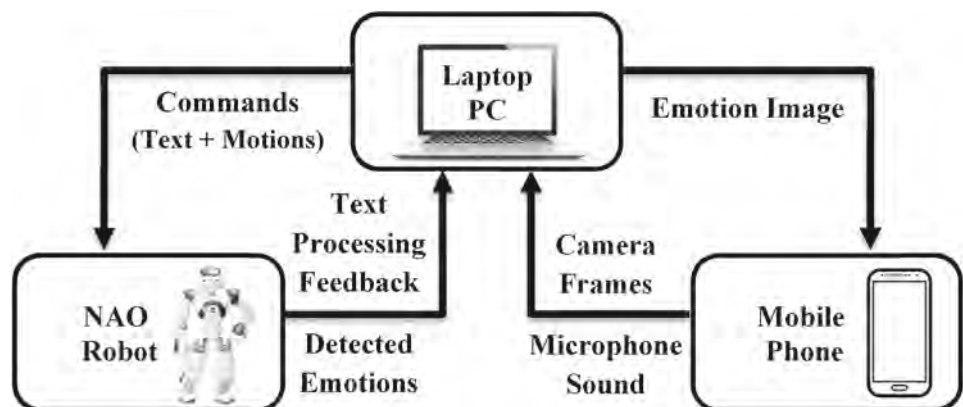
During the interaction, the operator can go to the next scenario part, repeat the last part, call the patient by his or her name, change the speaking volume, trigger a praise phrase, and define and trigger an operator-defined response. The operator-defined responses are saved as a personal individualized scenario.

With the systems described above working together, the control system block diagram can be represented as depicted in Fig. 4. The system parts (laptop PC, NAO robot, and mobile phone) communicate via a WiFi connection. The software is explained in the interaction and assessment system flowchart in Fig. 5. The number of phrases to use from each section is called ‘section load’. Initially, all section loads are set to 3 phrases. Later, after each session, each section load is multiplied by the corresponding updated section weight to update the number of phrases to be used.

3 Experimental Setup

Experiments were conducted in Al Ain Hospital rehabilitation center in collaboration with autism therapists. The operator, who is not a therapist, sat in a separate control room monitoring the robot–child interaction setup remotely and controlling the assessment system. The robot was placed in the therapy and interaction room standing on a table so that it was at the same level as the patient sitting on a chair, as shown in Fig. 6a. The setup was composed of the robot, patient, therapist to intervene in potentially harmful situations, and parent (if available), as shown in Fig. 6b. A camera recorded the interaction sessions, so that the parents could watch the sessions if they were not available at the session time.

Fig. 4 Control system block diagram



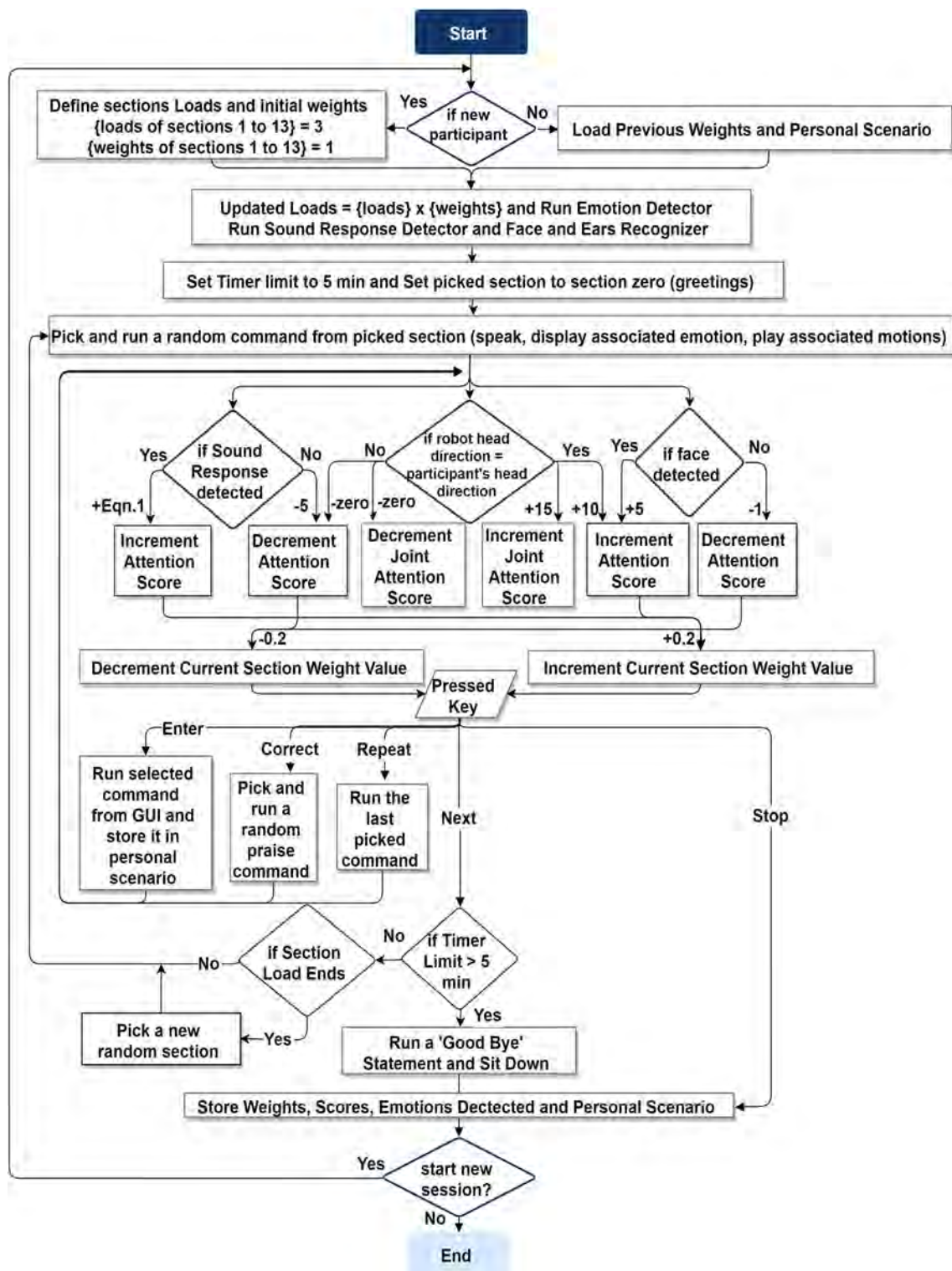


Fig. 5 Interaction and assessment system flow chart

3.1 Experimental Protocols

The robot was introduced to the child as a new friend. The experiment started by asking the child to sit on a chair facing

the NAO robot and talk with it. The therapist was asked to fill out the therapist assessment sheet during the experiment if possible, or later at the time of watching the video. At the end of the session, the parent was asked to fill the parent

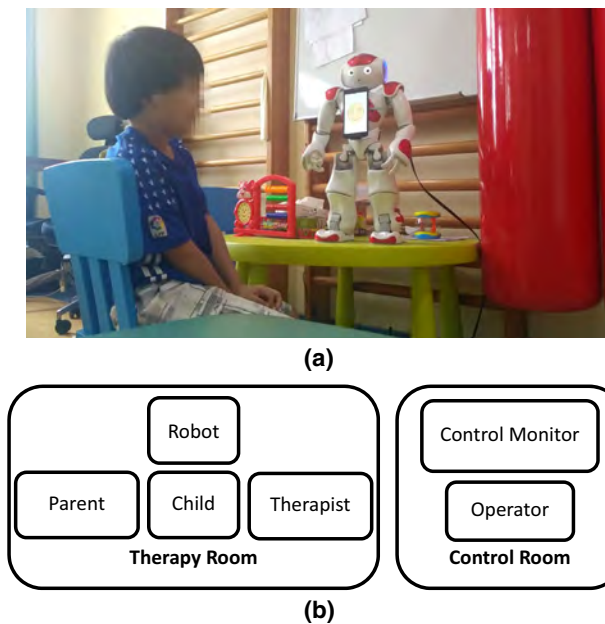


Fig. 6 **a** Therapy and interaction room setup. **b** Schematic diagram of experimental setup

feedback form. The therapist assessment sheet was an attention scale from 0 to 100. The therapist rated the attention of the child every 30 s and marked the value on the assessment sheet; subsequently, the values on this sheet were compared to the assessment system scores. The parent feedback form had two scale questions: “How do you rate your child interaction with the robot today?” and “How do you rate your child interaction at home?,” as well as a third Yes or No question: “Do you think robot therapy is helpful?” and a blank space with the heading “Any Comments.” The parent feedback was important to rate the interactions of the children with the robot compared to their interactions with their families, and to detect whether interacting with the robot influenced their interactions at home. The robot interacted with the child for 5 min (specified time limit), after which the session ended. At

the end, the robot thanked the child, encouraged him or her to come again for subsequent sessions, and said “bye bye.”

3.2 Real-Time Score Visualization

The accumulated attention, joint attention, and sound response scores were presented in real time in a separate single compact plot, as shown in Fig. 7, to convey the level of interaction between the patient and the robot to the operator. The plot shows the sound response request instants and joint attention requests instants as well. Sound response request is generated when scenario phrase has sound response option. Joint attention request is generated when scenario phrase has joint attention option. The algorithm saved a copy of the detailed results at the end of each session in a pickle file for processing. The scoring system GUI allowed the operator to plot the detailed results immediately after each interaction session.

4 Results

4.1 First Impression

Patients were included in a “one trial” experiment to assess their first impressions of the robot. Since adaptation depends on previous interaction history, all patients had no history at the first session, which means that the system had no adaptation capabilities yet. Sample score results generated after the session of one patient are shown in Figs. 8, 9 and 10.

The normalized attention scores in Fig. 8 show that the attention level increases over time, with minor drops representing gaze distractions. In addition, the joint attention level increases two times, representing two successful joint attention events, where the robot talks about and points towards an object to either the right or left and the patient is expected to follow. Moreover, the emotion color map shows several detected happy facial emotions, mainly at the beginning of

Fig. 7 Sample real-time plot showing the attention score (red line), joint attention score (blue line), sound response requests (yellow bars), i.e., time instants at which a sound response was expected, and joint attention requests (green bars), i.e., time instants at which a joint attention cue was expected. The white numbers above the red line represent accepted sound responses in units of seconds, which are less than the response-time-out parameter values. (Color figure online)

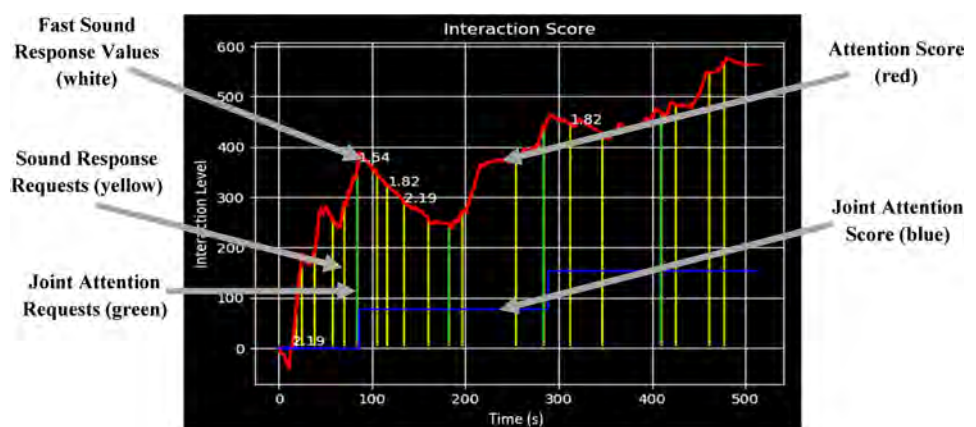


Fig. 8 Normalized attention (red) and joint-attention (blue) scores with emotion color bands. Scores were normalized by scaling data to have the values between 0 and 100, by dividing values by the maximum value and multiplying by 100. The attention scores and predicted emotions could be mapped for each time instant. An emotion result of “None” represents an instant at which emotion prediction is not possible.
a Highlights a case of attention score drop with sad emotions, then joint attention score increase with happy emotions.
b Highlights a case of attention score increases with surprise emotions. (Color figure online)

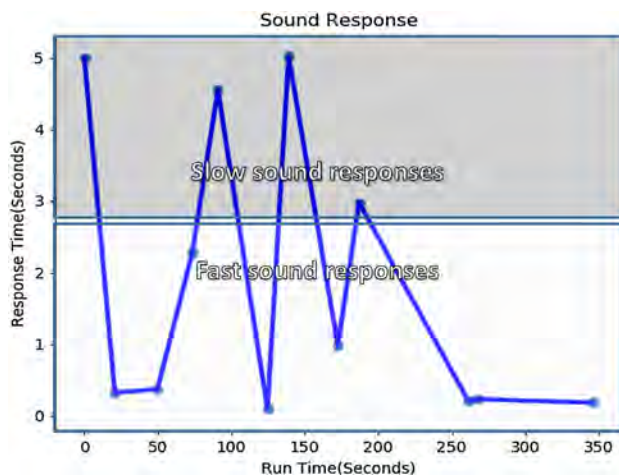
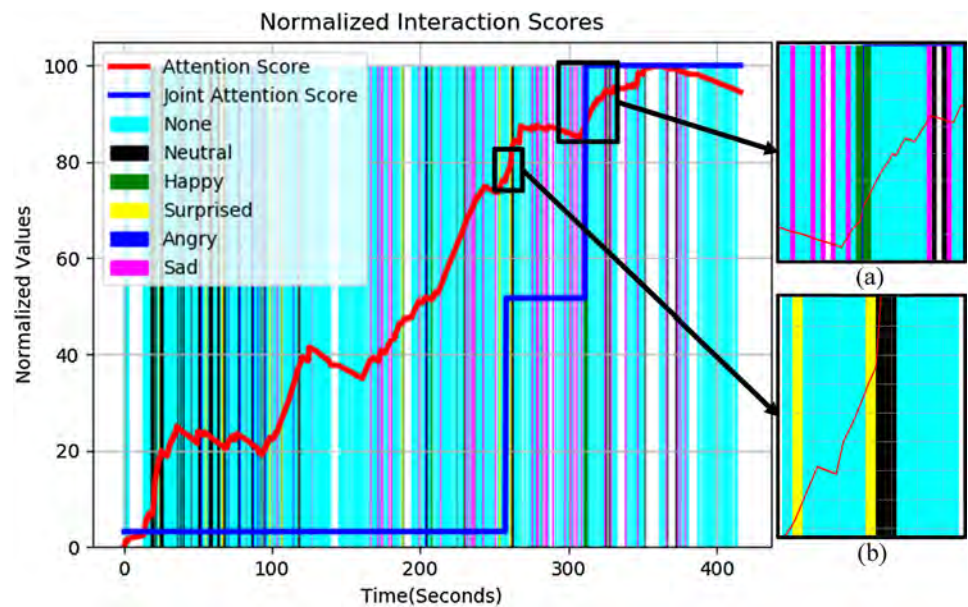


Fig. 9 All recorded sound response values. Only fast responses (equal to or less than the Response-time-out value) are plotted in real time. Response-time-out and several other parameters explained in Sect. 2 can be adjusted to suit the needs of each patient for better cue detection

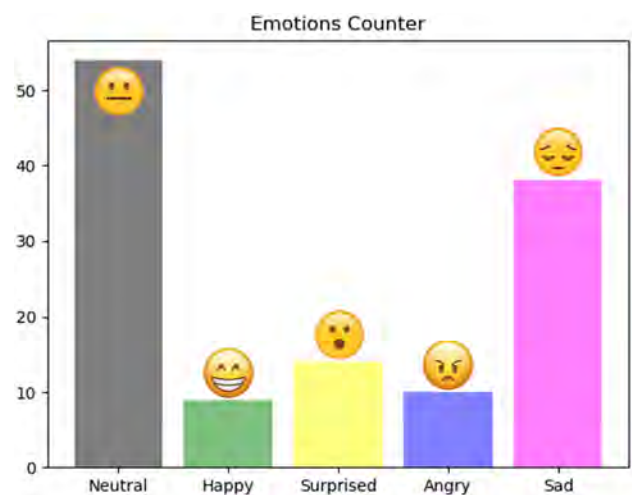


Fig. 10 Emotion counts, showing how many times each emotion was detected

interaction and at the instants of joint attention, as shown in Fig. 8a. Some sad and angry emotions were detected, where the patient could have been experiencing some discomfort. Thus, these emotions mainly occurred after drops as shown in Fig. 8a, where the patient was not looking at the robot and hence no emotion data were obtained, or during a disturbed gaze. Some surprise was detected during steadily increasing attention phases as shown in Fig. 8b, when the robot regained the attention of the patient using an attractive and surprising action.

The sound response values in Fig. 9 are divided into slow and fast sound responses based on the Response-time-out value. This patient had more fast sound responses than slow

sound responses (only four slow sound responses), which means that the patient was very responsive.

The emotion counter in Fig. 10 shows that the neutral emotion is the most dominant and that sadness in discomfort instants lasted for some time. It also shows that happiness was detected 10 times and surprise 15 times, corresponding to excitement instants.

The average attention score of all patient trials was found and is plotted in Fig. 11. The data show that most of the patients had positive first impressions and that their attention levels increased during the interaction. However, some patients did not follow this pattern, which explains the large standard deviations.

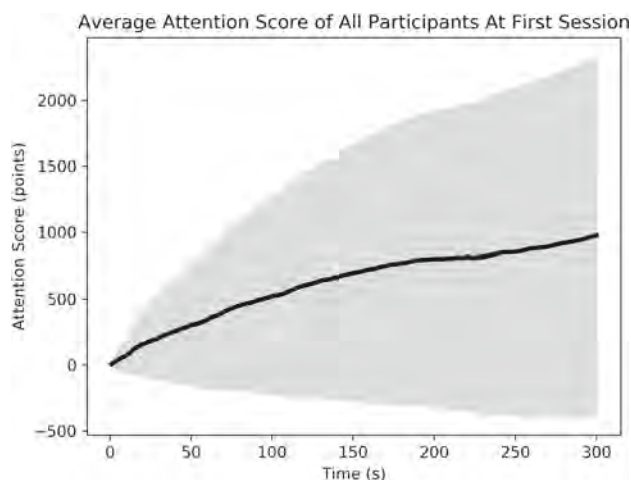


Fig. 11 Average attention scores of all patients in the first session

4.2 Interaction Progress

Six out of the 11 patients were able to continue in a long-term progress study. Further experiments involving these six patients were performed during the following 7 weeks to examine their interaction progress and the variations in the obtained weights of the scenario sections.

The session attendance of the patients over the 7-week period is illustrated in Fig. 12. Some patients could not come to all the sessions. Although some patients had more sessions than others, each patient had at least four sessions.

The weight changes of the scenario sections over the first 4 weeks are shown in Fig. 13. The scenario was the same for all patients in the first session, so they have the same section weights. In the second week, the weights took two different sets of values. It took up to the fourth week for six different

Fig. 12 Session attendance by week

Patient No. 1	X	X	X	X	X		
Patient No. 2	X	X	X	X	X	X	
Patient No. 3	X	X	X	X	X	X	
Patient No. 4	X	X	X	X	X	X	X
Patient No. 5	X	X	X	X	X	X	
Patient No. 6	X	X	X	X			
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7

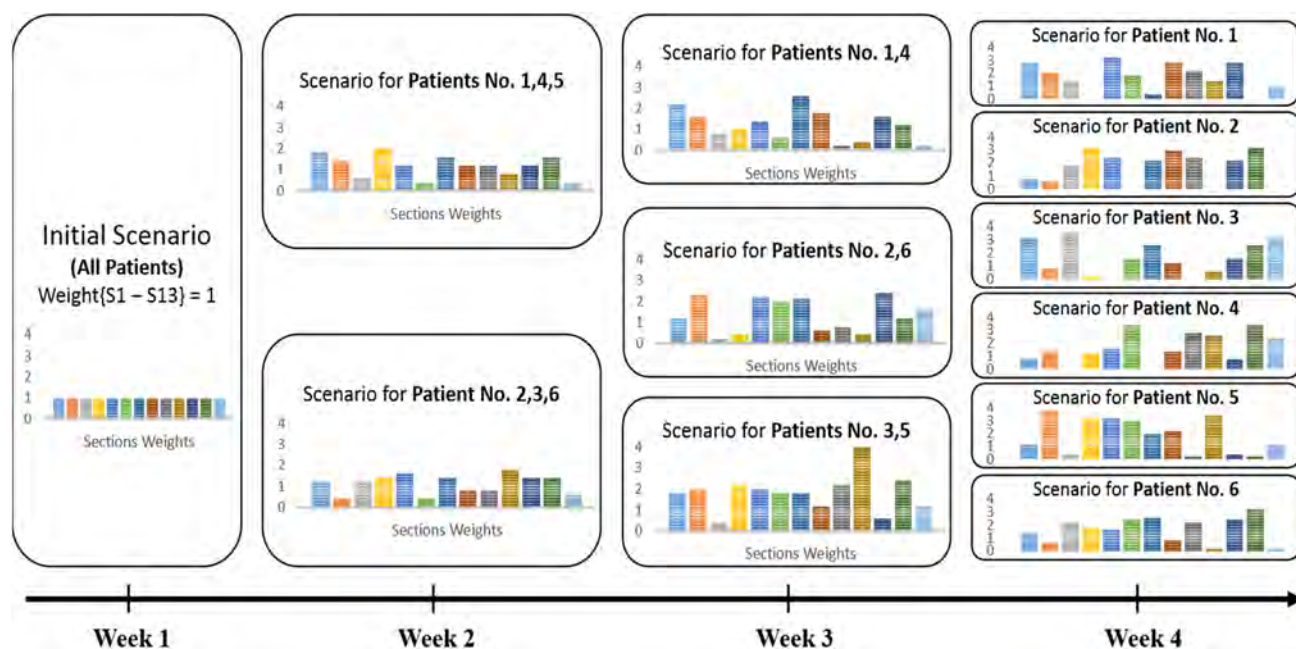


Fig. 13 Scenario flow and sections weight diagram. The weights change in each session based on the attention scores of the patients. Any “zero weight” section ceases to appear. The weight could not be zero from the first two sessions. S1–S13: section 1 to section 13

sets of weights to emerge; thus, each patient had different personal weight values at this point.

The results of the empirical sessions for all patients are shown in Fig. 14, which compares the progress in terms of attention scores, therapist assessment, and parent feedback. The results show that all six participants portrayed a trend of increasing attention scores. However, the therapist and system assessment trends are similar for most of the patients.

5 Discussion

The average attention score in the first impression trials generally increases for 9 out of 11 patients from the beginning of interaction until the end of the session. The other two patients show decrease at some points. This finding demonstrates that, on average, patients had positive first impressions and could complete the 5 min of interaction. Since not all scenario sections were familiar to all patients and scenario personalization requires at least one previous interaction history to be active, attention discontinuities occurred at the times of unfamiliar scenario sections. Consequently, attention results varied among patients, and the average attention score had a large standard deviation.

The later sessions had increasing (patients 1, 2, and 3) or fluctuating (patients 4, 5, and 6) attention levels with some drops. These drops have several possible causes. The primary possible cause is hyperactivity, which the therapist explained as being due to the patient not taking energy-draining sessions for a long time or not taking the prescribed medicine. Another possible cause is the patient's interest in small scenario sections (i.e., those with few defined statements), which decreases gazing time. In this case, the system prompts the operator to resolve the issue and define more statements for the subsequent sessions in the sections of patient interest.

The long-term interaction progress results show that robot intervention in autism therapy is highly beneficial for autistic patients. They enjoy the treatment sessions and give enough attention to learn or enhance a skill in every session. Noticeable changes in patient behaviors and skills took a few weeks, until the patients became familiar with the robot voice and moves. Breaking the ice took one or two sessions on average, until the patients interacted with the robot freely and excitement reached a steady level.

The autistic patients built strong bonds with the robot as a friend who encouraged them to withstand the therapy time. One of the patients once brought some friends to meet the robot as a new friend. We observed that the therapists at the hospital went further by encouraging their patients to do energy-draining physical therapy exercises by offering a session with the robot as a reward. The therapists reported that the patients asked about the robot every day and looked forward to the day of the robot session.

The parent of patient 3 reported that the interaction of the patient at home increased noticeably after taking sessions with the robot. The parent commented:

He has improved verbally since he has been interacting with the robot... There has been an impressive progress with his communication and interaction at home. He has a long way to go. But each step has been a joyful success for us. Thank you.

Moreover, most parents reported that their children mimicked and repeated many of the robot responses at home, and some of the children asked their parents at home the same questions that the robot had asked them during the session. Five out of six parents believed that their children performed as well as at home or even better in some sessions (patients 2 and 5).

The therapist assessment shows an increasing trend, with all patients demonstrating an increased attention level over time. Four out of six drops in session attention progress levels detected by the therapist were detected by the assessment system as well. Overall, six mismatches, highlighted with red circles in Fig. 14, in the attention score were found between the system and therapist assessments out of 34 sessions (82.4% accuracy).

Emotion predictions are of great assistance in understanding the emotional states of children in different scenario sections. Mapping the predicted emotions using the attention score and associated scenario section enables understanding of the facial responses of patients to specific topics and could facilitate the detection of difficulties in emotional responsiveness and other similar autism characteristics.

The objective of this study was to develop a simple robot-mediated assessment and interaction technique to prepare such systems for long-term presence in autism rehabilitation centers. This system can be used by therapists and parents after short training. The proposed system is easily replicable due to its simplicity and ease of use, and can cope with a large number of patients simultaneously. In addition, it may play a complementary and coadjutant role in the therapy of the patients who cannot continue traditional treatment sessions or need sessions with a frequency higher than rehabilitation centers can offer due to a lack of therapists.

Because of the heterogeneity of ASD disorders, one predefined intervention scenario cannot possibly address the needs of all patients. The proposed adaptive dynamic scenario and weighting technique allows customized interactions for each child. Such adaptive techniques have proven effective in sustaining social engagement during long-term children–robot interactions [28]. The employed techniques maximize engagement, which is one of the strongest predictors of successful learning [29, 30], using a ludic mobile robot to stimulate social skills in ASD children. Moreover,

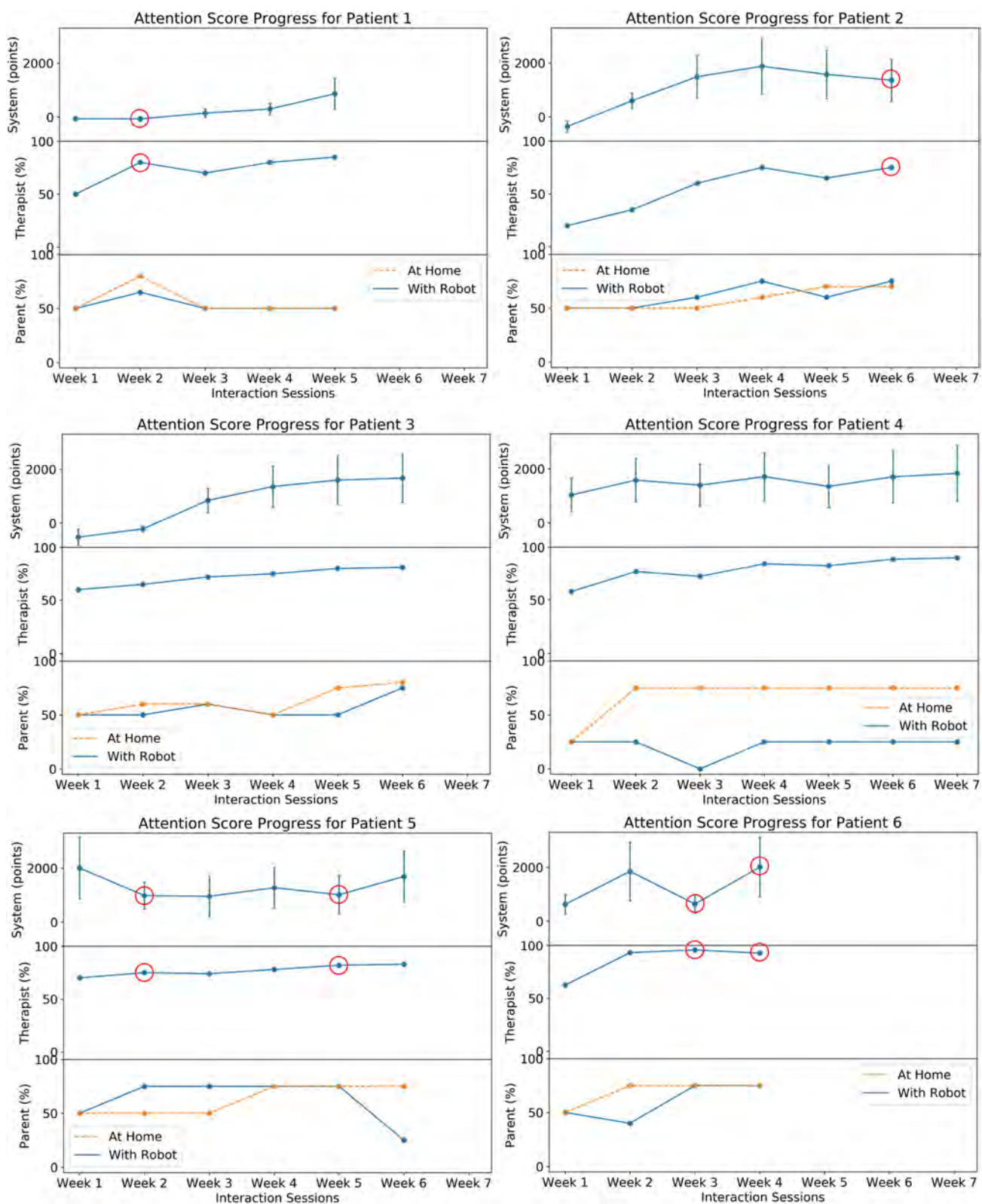


Fig. 14 Session results over several sessions for all patients, showing the average system attention scores (upper plots), the average therapist attention assessments (middle plots), and parent feedback when asked

about the interaction level after each session with the robot and at home (lower plots). Red circles show the mismatches between therapist and system results. (Color figure online)

the proposed system reduces the therapist's subjectivity in assessment and allows for early intervention.

The current study has several limitations that need to be highlighted with a view to address them through future developments and integrations. First, only one-on-one interaction is possible, to allow the child to focus on the robot only and to allow the system to capture the child's attention cues. Moreover, the study included a relatively small number of patients, and more results may be revealed when it is applied to more patients. Also, the proposed interaction system is only applicable to patients with moderate severity and who have at least minimal verbal response capabilities. Some children may become distracted by the mobile phone display on the robot, since they are used to playing games with such devices, which may lead to drops in attention score. This issue occurred only at the beginning of the early sessions where the children were exploring the robot features and it has been partly addressed by fixating the display in a way that it cannot be moved by the children.

6 Conclusions

An adaptive robot intervention system for ASD assessment and therapy was designed for clinical use and tested empirically on six ASD patients in an autism rehabilitation center. The results demonstrate that the proposed assessment system can accurately represent the attention levels of patients with mild to moderate ASD and simple verbal communication skills, matching over 80% of therapist assessments. The proposed adaptive, dynamic interaction system yielded remarkable improvements in the attention levels of most of the patients in long-term therapy.

Based on these outcomes, our hypothesis is that a properly designed robot intervention system can increase the attention levels of ASD children insofar as it enhances their engagement and, in so doing, helps them improve their communicative and social skills. Moreover, the same system can facilitate the assessments of autism symptoms providing therapists with a useful set of reliable and objective quantitative methods. The proposed system is so flexible, robust, user-friendly, and easily customizable that we infer it could be utilized without effort by parents in domestic environments. Not only does not the system require any previous technical experience, but—thanks to the scalability of robotic intervention—it enables the efficient treatment of a large number of patients, increasing the frequency of the sessions that can be administered to the children while implementing exactly replicable protocols.

Some authors maintain that exposure to digital technology can aggravate the social symptoms of autism in children, worsening their deficits and possibly increasing the chances of developing obsessive compulsive behaviors [27]. The out-

comes of our research mitigate these worries. Our research provides robust anecdotal evidence that proper design and supervised application of robots in autism therapy has the potential to make ASD subjects feel spontaneously engaged in basic forms of social interaction and, at least apparently, significantly less anxious than during the typical interactions with other humans. We speculate that one of the contributing factors in achieving this positive result is the fact that robots allow effective forms of pseudo-social interactive engagement while decreasing the complexity of the social context and hence reducing the excessive emotional and cognitive burden that autistic children typically have to process. Our hypothesis is that this technology offers ASD children an opportunity to familiarize with social interaction in a context that they find conducive and reassuring with a pace that they can comfortably control by means of tasks that they can repeat at will without hurry. That is why we believe our robot-based interventions have the potential to support the development of social skills in autistic subjects and can teach them new ones, consolidating their ability to interact with other humans.

Future developments of the proposed system would have to aim to increase the assessment accuracy and further enhance the patient's engagement with the robots broadening the set of available types of interaction and increasing the degrees of freedom that define such interactions. Multiple cameras could be employed where a fixed observation setup is possible (considering the specificities of the clinical setting) to preserve the robot's mobility while broadening their interactive capabilities. Furthermore, a large set of patients would be desirable to test more extensively all the functionalities of the system and tune them for performance improvement. Finally, it would be useful to test whether a virtual avatar displayed on a tablet could reduce the costs involved in the use of embodied mobile robots, simplifying the use of the interactive system in the home setting: however, we anticipate that the quality of the interaction with a virtual avatar might be inferior in terms of both quality and length to the interaction established by the children with a physically embodied robot.

Funding This study was funded by a 31R188-Research AUA- ZCHS -1–2018, Zayed Health Center.

Compliance with Ethical Standards

Conflict of interest The second and fourth authors are coeditors of the special issue to which this paper is submitted, however do not anticipate to be involved in the review process for this paper. The other authors declare that they have no conflict of interest.

References

1. Rotheram-Fuller E, Kasari C, Chamberlain B, Locke J (2010) Social involvement of children with autism spectrum disorders in elementary school classrooms. *J Child Psychol Psychiatry* 51:1227–1234
2. Moriuchi JM, Klin A, Jones W (2016) Mechanisms of diminished attention to eyes in autism. *Am J Psychiatry* 174:26–35
3. Freeth M, Foulsham T, Kingstone A (2019) What affects social attention? Social presence, eye contact and autistic traits. *PLoS One* 8(1):e53296. <https://doi.org/10.1371/journal.pone.0053286>
4. Roddy A, O'Neill C (2019) The economic costs and its predictors for childhood autism spectrum disorders in Ireland: how is the burden distributed? *Autism* 23(5):1106–1118
5. Huerta M, Bishop SL, Duncan A, Hus V, Lord C (2012) Application of DSM-5 criteria for autism spectrum disorder to three samples of children with DSM-IV diagnoses of pervasive developmental disorders. *Am J Psychiatry* 169:1056–1064
6. Volkmar F, Siegel M, Woodbury-Smith M, King B, McCracken J, State M (2014) Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry* 53:237–257. <https://doi.org/10.1016/j.jaac.2013.10.013>
7. Llana DC, DeLuke SV, Batista M, Crawley JN, Christodulu KV, Frye CA (2010) Communication, interventions, and scientific advances in autism: a commentary. *Physiol Behav* 100:268–276
8. Paul R (2008) Interventions to improve communication in autism. *Child Adolesc Psychiatr Clin N Am* 17:835–856
9. Costa S (2014) Robots as tools to help children with ASD to identify emotions. *Autism* 4(1):1–2
10. De Bel-Air F (2015) Demography, migration, and the labour market in the UAE. *Gulf Labour Mark Migr* 7:3–22
11. Aresti-Bartolome N, Garcia-Zapirain B (2014) Technologies as support tools for persons with autistic spectrum disorder: a systematic review. *Int J Environ Res Public Health* 11:7767–7802
12. Pennisi P, Tonacci A, Tartarisco G, Billeci L, Ruta L, Gangemi S, Pioggia G (2016) Autism and social robotics: a systematic review. *Autism Res J* 9:165–183
13. Alahbabi M, Almazroei F, Almarzogi M, Almeheri A, Alkabi M, Al Nuaimi A, Cappuccio ML, Alnajjar F (2017) Avatar based interaction therapy: a potential therapeutic approach for children with Autism. In: IEEE international conference on mechatronics and automation, pp 480–484
14. Zheng Z, Zhang L, Bekele E, Swanson A, Crittendon JA, Warren Z, Sarkar N (2013) Impact of robot-mediated interaction system on joint attention skills for children with autism. In: IEEE international conference on rehabilitation robotics, pp 1–8
15. Robins B, Dautenhahn K, Dubowski J (2006) Does appearance matter in the interaction of children with autism with a humanoid robot? *Interact Stud* 7:479–512
16. Zhang K, Liu X, Chen J, Liu L, Xu R, Li D (2017) Assessment of children with autism based on computer games compared with PEP scale. In: 2017 international conference of educational innovation through technology (EITT), pp 106–110
17. Rudovic O, Lee J, Dai M, Schuller B, Picard RW (2018) Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci Robot* 3:eaa06760
18. Ferrer EC, Rudovic O, Hardjono T, Pentland A (2018) Robochain: a secure data-sharing framework for human–robot interaction. *arXiv preprint arXiv:180204480*
19. Anzalone SM, Tilmont E, Boucenna S, Xavier J (2014) How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot. *Res Autism Spectr Disord* 8:814–826
20. Kim ES, Paul R, Shic F, Scassellati B (2012) Bridging the research gap: making HRI useful to individuals with autism. *J Hum Robot Interact* 1:26–54
21. Bensalem S, Gallien M, Ingrand F, Kahloul I, Thanh-Hung N (2009) Designing autonomous robots. *IEEE Robot Autom Mag* 16:67–77
22. Billard A, Robins B, Nadel J, Dautenhahn K (2006) Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assist Technol* 19:37–49
23. Kozima H, Nakagawa C, Yasuda Y (2005) Interactive robots for communication-care: a case-study in autism therapy. In: IEEE international workshop on robot and human interactive communication, pp 341–346
24. Schopler E, Van Bourgondien M, Wellman J, Love S (2010) Childhood autism rating scale—second edition (CARS2): manual. Western Psychological Services, Los Angeles
25. Alnajjar FS, Renawi AM, Cappuccio M, Mubain O (2019) A low-cost autonomous attention assessment system for robot intervention with autistic children. In: 2019 IEEE global engineering education conference (EDUCON)
26. Pokress SC, Veiga JJD (2013) MIT app inventor: enabling personal mobile computing. *arXiv preprint arXiv:13102830*
27. Wilson PI, Fernandez J (2006) Facial feature detection using Haar classifiers *J Comput Sci Coll* 21:127–133
28. Ahmad MI, Mubin O, Orlando J (2017) Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *Int J Hum Comput Interact* 33:943–962
29. Marcos-Pablos S, González-Pablos E, Martín-Lorenzo C, Flores LA, Gómez-García-Bermejo J, Zalama E (2016) Virtual avatar for emotion recognition in patients with schizophrenia: a pilot study. *Front Hum Neurosci* 10:421
30. Powell S (1996) The use of computers in teaching people with autism. In: Autism on the agenda: papers from a National Autistic Society conference, London, pp 128–132

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Fady Alnajjar is Assistant Professor of Computer Science and Software Engineering at the College of IT of UAE University. Dr. Alnajjar is also the director of AI and Robotics Lab at the same university. He is also a visiting research scientist at the Intelligent Behavior Control Unit, BSI, RIKEN. He received a M.S. degree from the Department of Human and Artificial Intelligence System at the University Of Fukui, Japan (2007), and a Dr. Eng. degree in System Design Engineering from the same university in 2010. Since then he had interest on brain modeling in aim to understand higher-order cognitive mechanisms. He also started to study motor learning and adaptation from the sensory and muscle synergies perspective in order to suggest practical robotics applications for neural-rehabilitations.

Dr. Massimiliano Cappuccio is Deputy Director of the Values in Defence and Security Technology group at University of New South Wales Canberra. The group investigate the normative frameworks that regulate human-robot teams and autonomous weapon systems in the military and the law enforcement contexts. Max also holds a secondary affiliation with UAE University, the national university of the United Arab Emirates, where he collaborates with the AI & Robotics Lab, developing a line of interdisciplinary research in neuro-cognitive robotics, robotics in education, and child-robot interaction. In this context, Max has used humanoid robots to conduct empirical studies and

test neuro-cognitive models of choking effect and social interactions of autistic children. Max's research has three main thematic foci: human-robot interaction and social cognition (human-robot teams, normative frameworks for social coordination); technology ethics (including the ethics of AI and virtue ethics applied to human-agent interaction); human performance and embodied cognition (skill acquisition, performance under pressure). This research is interdisciplinary in nature as it integrates phenomenological analyses, empirical experimentation, and synthetic modelling. Max's early work, conducted in Italy, France, Netherlands, and the US, has focused on mirror neurons theory, joint attention and deictic pointing, the frame problem of artificial intelligence, the theoretical foundations of computationalism. Max is the organiser of international academic events like the annual Joint UAE Symposium on Social Robotics in Abu Dhabi and the TEMPER workshop on human performance in Canberra.

Eng. Abdulrahman Renawi has M.S. in mechatronics from the American university of sharjah. Currently working as a research assistant at the AI and Robotics lab in UAE University. His main research interest is to utilize AI techniques to design automatic assessment system for attention record level of autistic children in classroom.

Dr. Omar Mubin is a Senior Lecturer in Human Computer Interaction at the School of Computer, Data and Mathematical Sciences at Western Sydney University, Australia. Dr Mubin's primary research interests are human robot interaction and human-agent interaction. Specifically, he studies social robotics and their applications and consequently interaction with humans in educational, public spaces and information dissemination scenarios. Dr Mubin is also the general chair for the Human Agent Interaction (HAI2020) conference. His current h-index according to Google Scholar is 16 with a more than 1400 citations. Dr Mubin is involved in teaching and supervising (undergrad and post-grad) students in the broader area of Human Computer Interaction, Mobile Computing and Health Informatics.

Professor. Dr. Chu Kiong Loo completed his Bachelor of Mechanical Engineering degree with honours at the University of Malaya in 1996. He then received his PhD degree at the University of Science Malaysia in 2004, specialising in neurorobotics and its applications in Connected Healthcare. Currently he is a full Professor at the Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, University of Malaya. His main research interests are neuroscience inspired machine intelligence. He has led numerous research projects as Principal Investigators of several international funding projects including Joint Research Program between UAE University and Institutions of Asian Universities Alliance (AUA). From these research activities, he has published more than 200 peer-reviewed journal and conference papers on neurorobotics and machine intelligence. Prof. Loo is also the recipient of JSPS fellowship award from Japan and Georg Forster Fellowship award from Humboldt Foundation, Germany. He has been visiting professor of University of Perfectual Osaka University, Japan and King Mongkut's Institute of Technology Ladkrabang, Thailand.

Chapter 3, Part 0: Preamble

3.0.1 Themes and problems

Chapter 3 investigates (i) how human-robot (pseudo-)relationships are shaped by the attitudes and beliefs of the human interactants; (ii) how these attitudes have a moral significance, as they can express (or promote) virtuous or rather vicious behaviours in humans; (iii) how the embodiment and the behaviour of robots contribute to shape such attitudes and belief, thus indirectly impacting the moral character and the ethical behaviour of the human users. The user's attitudes I will consider are four, in accord with what I call the "ATTA-bot" model of human-robot relationships: Acceptance, Trust, Tolerance, and, importantly, Anthropomorphism.

The notion of Anthropomorphism captures the human tendency (often implicit) to attribute intentions and experiential states to inanimate robots. Our spontaneous tendency to anthropomorphise robots, forming (pseudo-)empathic relationships with them, explains why we are destined to perceive humanoid robots as moral patients, capable to suffer pain and injustices. Because of this inevitable bond, relationships between humans and robots call for a particular kind of ethical consideration for the latter, regardless that they are inanimate beings without rights or entitlements: first, our behaviours and thoughts over robots reveal the kind of moral agents we are, and whether our moral character is good or bad; secondly, the quality of our recurrent interactions with robots and the way we habitually treat them, might in the long run affect our moral character, promoting virtuous or rather vicious inclinations in the immediate users and in the society at large.

My point is that anthropomorphism, empathy, and moral consideration for robots are rooted in the embodied and dispositional features of the human user. The kind of aesthetic and moral attitudes that determine our predilection for or refusal of robots are embodied in the specific sense that, like sensorimotor skills and social skills, they are underpinned by pre-reflective dispositions: while they are recurrent forms of thought and judgment, attitudes and beliefs also derive from habitual patterns of perception, emotion, and action, thus involving the embodied dimension of HRI. This dimension shapes the normative aspects of human-robot relationship, orienting our perception of what is right or wrong in them. For example, a robot attending a class of students may solicit in them self-attention and self-control, which are key to promote moral awareness and disciplined behaviours.

3.0.2 Research outcomes covered in this section

Chapter 3 expounds both a theoretical achievement and an applied outcome. The theoretical achievement, illustrated in **CP6** and **CP7**, is a comprehensive model of HRI that maps the key attitudes of human interactants toward robots. This model reveals the reciprocal connection of three important attitudes (Trust, Acceptance, and Tolerance) and their communal co-dependency from anthropomorphism, thus recognizing the specific causal contributions of embodied and non-embodied factors in determining adoption propensity. This model is, for convenience, referred to as “ATTA-bot”, where the acronym ATTA stands for Acceptance, Trust, Tolerance, and Anthropomorphism. Building on it, my colleagues and I have developed a virtue-ethical approach to robot ethics, that is an ethical framework that interprets our relationship with robot in the lights of the traditional philosophical notions of “virtue”, “moral character”, and “good life”. This ethical framework, illustrated in **CP8**, **CP9**, and **CP10**, is particularly important to understand the moral function that robots play during the developmental age: robots specifically designed to have a positive impact on the education of children can solicit moral attitudes that eventually may contribute to form their character and cultivate its positive qualities (Cappuccio et al 2021b [**CP8**]).

The second research outcome achieved by this chapter, documented by **CP11**, puts to the test the intuitions of virtue ethics in the applied domain: the outcome is a prototype of socially interactive agent (a humanoid robot) designed to monitor students undergoing an examination in school environment. This robot “invigilator” is expected to monitor and report suspicious behaviours, thus promoting discipline, honesty, self-restraint with its judgmental presence. Based on Schellen, Bossi & Wykowska (2021) and several other studies, we assumed that the monitoring functioning of a robot gaze (a particular application of ‘social attention’) may solicit certain virtuous behaviours (being focussed and dedicated to one’s task) while deterring vicious behaviours (copying and having an undisciplined conduct during the exam). The data collected during our experiment only partially met our expectations: the presence of the robot made the students more excited and inclined to misbehave during the exam. We impute this partial disappointment to the defective embodiment of Nao, which is too friendly and not sufficiently imposing to command discipline. This was useful to understand the implications of robot embodiment in the educational context.

3.0.3 Author's publications used in this chapter

In Chapter 3, I will review and discuss the contents of 12 of my publications (7 “Core Publications” and 5 “Support Publications.”)

Core Publications

- [CP5] Cappuccio M.L., Galliot J.C., Lanteri A., Eyssel F. (2024)
- [CP6] Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021)
- [CP7] Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2020)
- [CP8] Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2021)
- [CP9] Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (eds.) (2021)
- [CP10] Alnajjar F., Bartneck C., Baxter P., Belpaeme T., Cappuccio M., Di Dio C., Eyssel F., Handke J., Mubin O., Obaid M. (2021)
- [CP11] Mubin O., Cappuccio M.L., Alnajjar F.S., Ahmad M., and Shahid S. (2020)

Support publications

- [SP15] Alotaibi M., Alnajjar F., Cappuccio M.L., S. Khalid (2023).
- [SP16] Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021).
- [SP17] Cappuccio M.L., Devitt K.S., Cara Wrigley (2020), redacted.
- [SP18] Cappuccio M.L., Peeters A., McDonald W (2019).
- [SP19] Sandoval E.B., Billingham M., Cappuccio M.L. (2024).

Chapter 3, Part 1 - Human Attitudes Towards Robots: Embodied and Disembodied Components

This section is centred on Core Publications #5 and #6 [CP5] [CP6]

3.1.1 Towards an embodied understanding of human-robot relationships

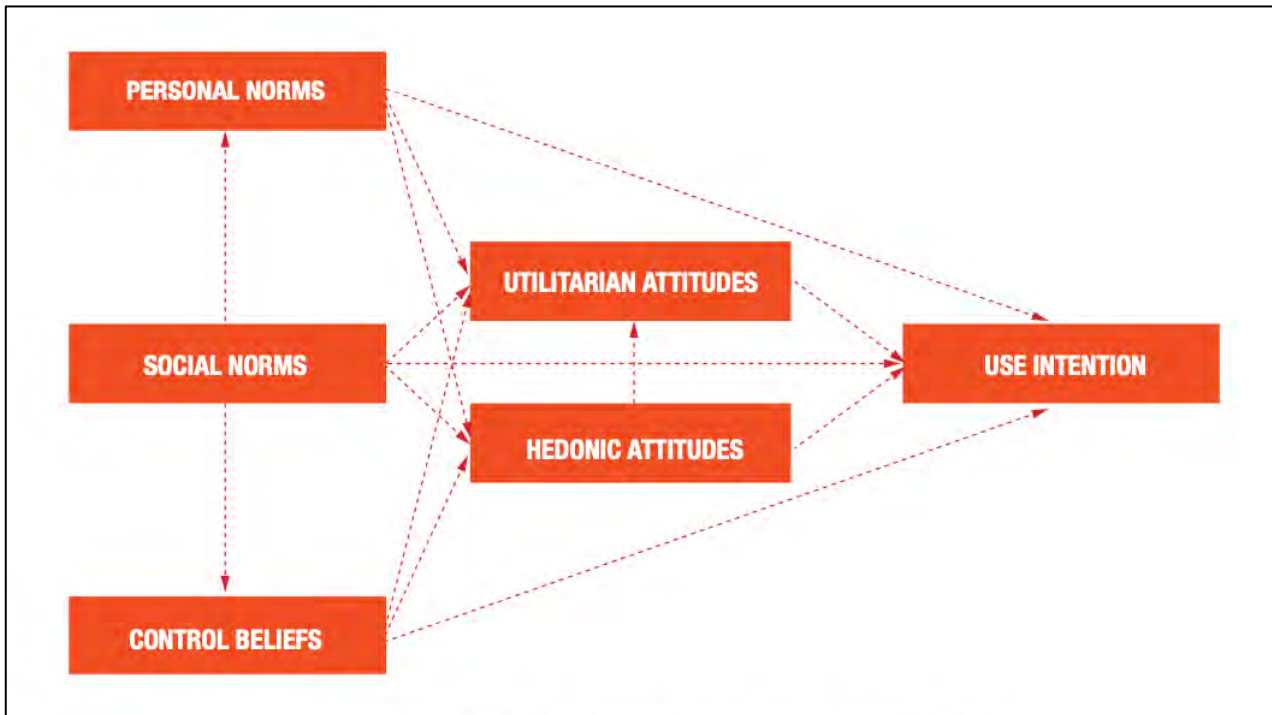
Part 1 of Chapter 3 deals with the human attitudes towards intelligent robots (including, but not limited to, social robots), and Autonomous Artificial Agents (or “Agents”) in general. This topic is investigated by HRI and human factors & ergonomics researchers under the label of “attitudes and beliefs towards autonomy” (Naneva et al 2020). The primary aim is to characterize attitudes based on how they impact positively or negatively HRI, influencing the propensity of final users and managers to adopt Agents and their expectations about establishing satisfactory interactions. More specifically, in accord with the principles of ECT and consistently with the work documented in Chapter 1 and Chapter 2, I intend to highlight the relation between attitudes towards autonomy and embodiment. I will consider two different dimensions of this relationship: embodiment as a precondition and a concrete aspect of attitudes (quality) vs embodiment as the thematic focus of attitudes (scope).

The first dimension captures the crucial distinction between attitudes that are embodied (in that they are rooted into the pre-reflective, habitual, and perceptual dispositions of the human users) and those that are not embodied (as they are based on more intellectual forms of abstract reasoning and reflection of said users). In our 2024 paper on technology resistance [CP5], my colleagues and I have argued that embodied attitudes (like Trust and Acceptance) influence user’s choices and behaviour in ways that are highly characteristic and different from the non-embodied attitudes (like Tolerance). Some empirical findings motivated me to clarify this distinction, acknowledging its (too often neglected) importance, and operationalize it within the practice of HRI research. The second dimension (explored in Part 2 and Part 3 of this Chapter) reflects, conversely, how the embodiment of robots affects the attitudes of the human users and whether some attitudes are more sensitive to the details of the presence and appearance of robots, revealing the importance of the physical implementation and behaviour of robots. In our 2021 paper on human-robot teams [CP6], we argue that some attitudes towards autonomy –in particular anthropomorphism– depend on the embodied qualities of the robot in a way that is profound and pervasive but also subtle, articulated, and often non-obvious.

What exactly are the ‘attitudes towards autonomy’ that concern HRI experts in particular? Clarifying that attitudes govern every variety of human-robot relationship, covering even the most impersonal, is absolutely crucial, rather than merely emphasizing the deeply emotional, personal ties that form between humans and social or humanoid robots: whether a robot is intentionally designed for social engagement with humans, to look human-like, intelligent, pleasant, kind, etc., the stances of its human operators very significantly affect usage patterns, revising the expectations linked to its presence and its adoption (Cappuccio, Galliot & Sandoval 2021a [SP16]; Damiano & Dumouchel 2018). Inherent in the fundamental concept of autonomy is the unwavering general expectation that any autonomous agent, irrespective of its design or outward form, should be regarded as an interactive entity amply furnished with the essential means to effectively traverse its environment, undertake goal-directed planning, devise its own actions, and decide autonomously or with human partners – unlike older factory robots, which exhibit very limited adaptability and were originally engineered as implements performing predetermined tasks within a specifically organized workspace (Duffy 2003; Epley et al 2008; Asada 2020).

Although robot companions are not mere tools, their function remains fundamentally assistive because they are not fully autonomous yet: humans must complement or oversee their activities as direct partners in interaction and collaboration. In this scenario, people neither merely command robots like slaves nor employ them as objects in actual practice; to synchronize effectively with robots, the human operators must establish a collaborative bond with them, keenly discern their intentions, and thoroughly assess their endeavors to fulfill those intentions (Hoffman & Breazeal 2004). Therefore, the effective implementation of Agents hinges on their ability to meet the users’ expectations, which may prove extremely varied (subverting ordinary assumptions), implicit (concealed even from the users), and ambivalent (colliding with alternative expectations). Consequently, people’s dispositions toward Agents align with impressions and emotional responses triggered by the Agents’ outward form, the user’s level of familiarity with Agents (i.e., skill set and prior experience with analogous systems), and assumptions about Agents (i.e., concrete information and varying degrees of validity, encompassing beliefs derived from logical analyses or from illogical or quasi-logical biases). In HRI, studies have repeatedly revealed that these inclinations hinge on multiple psychological, sociocultural, and skill-related elements (Eyssel & Kuchenbrandt 2012; Bartneck et al 2019a). Such determinants are identifiable and quantifiable via self-reports (the explicit approach) or external monitoring (the implicit approach): the former encompasses verbal outputs documented prior to, during, or following engagement (i.e., spontaneous remarks and replies to surveys or interviews); the latter

involves the behavioural and affective manifestations displayed during or after encounters with Agents (e.g., posture, gestures, gaze, facial expressions).



3.1.2 Making space for Tolerance

While the attribution of intelligence, autonomy, and pleasantness significantly influences HRI, Acceptance and Trust are arguably the two most extensively investigated attitudes in this field. To accurately identify these attitudes among human users, HRI researchers implement specialized tools, such as dedicated rating scales. Nevertheless, my research in this domain has underscored the necessity of introducing an additional attitude, which my collaborators and I name “Tolerance,” to supplement Acceptance and Trust and rectify their shortcomings. Tolerance (for Agents) is a theoretical construct crafted to capture the resilience or insusceptibility of human users to alienating beliefs and dystopian narratives about autonomous artificial agents. Its core concept emerged with my co-author Jai Galliot, who framed Tolerance as a measure of immunity to the anxiety and hostility that frequently accompany the introduction of autonomous technologies (Galliot 2018). While I was undertaking a research project on technology resistance, which was funded by the U.S. Department of Defence under the Minerva initiative, I recognized a pressing need to fully elaborate the notion of Tolerance for Agents and to devise a consistent methodology for examining the psycho-social ramifications of exceptionally low or high levels of this factor.

The first time I started realizing this need was when I was asked to conduct field research using qualitative methods as part of a project sponsored by the Royal Australian Air Force (RAAF). This unclassified research concerned the impact of a newly introduced autonomous robot in an Australian military airbase. As reported by RAAF in official press releases, in 2020 a Konica Minolta logistic bot was introduced in Amberley airbase to release the ground crew members of the 36th Squadron from the time-consuming burden of transporting bulky components from one side of a very long hangar to the other. By temporarily joining an ongoing research effort coordinated by the University of Sydney, a team formed by me and two other HRI experts was recruited to study the attitudes of managers and final users in relation to the newly introduced log bot. While the details of our final report are not significant to this discussion, it is worth mentioning that, quite unsurprisingly, the human responses varied significantly among the users. This offered us the opportunity to examine and classify a broad array of different attitudes and perceptions (Cappuccio, Devitt, Wrigley 2020 [SP17]). From my point of view, this experience was important to develop a more mature and nuanced methodological approach in relation to the general issue of human attitudes towards autonomy, forming an original conceptual perspective about the way such attitudes universally impact adoption propensity and utilization across various sectors and fields of application.

The insights gained from this experience helped me elaborate the fictional scenario that, with a different team of co-authors and with an entirely different aim, I have described and discussed in the IJSR paper “Autonomous systems and technology resistance: New tools for monitoring Acceptance, Trust, and Tolerance” (Cappuccio et al 2024 [CP5]). While the scenario described in this paper is largely speculative, it is also very plausible, as it is consistent with several bodies of empirical evidence available in the scientific literature and with personal observations. In this article, we propose a “thought experiment” to envision a hypothetical situation concerning a log bot introduced in an airbase to assist the ground crew. In this hypothetical scenario, over a few weeks of seamless integration of the autonomous system, supervisors report that personnel have ceased using the bot almost entirely, and some have even attempted to sabotage it. We imagine that a series of semi-structured interviews with the crew uncovers a phenomenon that is apparently counterintuitive but that perfectly makes sense if one considers the complexity of human responses: the crew does not want to use the bot despite not disliking or distrusting it. In fact, they admit finding it useful and dependable, with all crew members appreciating its friendly interface and its interactive features. Most of them show sympathy towards the bot, while a smaller segment express indifference. So, what dynamics motivated the

their hesitation to use the bot over time? Our analysis of this hypothetical scenario suggests that the crew, although not genuinely opposed to the bot, may have held concerns over the potential effects on relationships between two teams stationed on opposite ends of the hangar. They feared that a more frequent use of the bot would reduce opportunities for social interaction, thereby straining their professional and personal bonds. This shift in perception altered their view of their roles within the airbase, triggering latent tensions. Moreover, despite no apparent plan to replace human workers, the crew grew apprehensive about the possible redundancy they could face if additional bots were deployed. Finally, some managers noted that assigning logistics to a bot might raise moral concerns, as they believed that such specific tasks should be reserved to humans, as it has always been.



The fictitious airbase scenario illustrates how recalcitrant users may reject a new technological solution despite having had a favourable first impression of its applications upon its introduction. The initial positive impression about the log bot can be perfectly explained using the existing frameworks of Acceptance and Trust, two key metrics often used in social robotics research to gauge user willingness to adopt new technology: the idea is that crew members are pleased with the bot's interaction and assistance in daily tasks.

However, we must also consider that the overall inclination to use the bot was low, overall. Despite their positive impressions, the workers also harboured anxieties, manifested as a

negative judgment that inform their attitudes towards the bot. Crucially, this negative view does not replace their earlier positive thoughts; instead, it added a layer of negativity that overtook the initial positive response. In our thought experiment, the users maintained their previous belief that the bot (1) was pleasant to work with and (2) helped with tasks. However, they quickly came to form a third belief: that, despite their positive interactions, the bot's presence disrupted social roles and threatened job security, leading to anxiety about their positions. Arguably, this underlying negative sentiment did not emerge suddenly but rather constituted a latent belief, previously dormant, ready to surface under the right conditions.

Despite its fictional nature, this scenario highlights the objective need for a new metric in HRI. That is why, in our 2024 IJSR paper [CP5] on technology resistance, I and my co-authors started delving into the conceptual elaboration of notorious anti-tech theorist Ted Kaczynski: a controversial, militant promoter of the anti-tech ideology, Kaczynski had advocated the violent rebellion against intelligent systems to prevent that autonomous technologies could devoid human life of its meaning and purpose. Despite Kaczynski's hostile position, he was himself a user and a developer of technology and was not against the idea of using technological artifacts for instrumental purposes, i.e. to establish the control of human power over nature. He was specifically concerned with the possibility that those artifacts could annihilate human agency by replacing human decision and responsibility (Galliot 2018), a possibility that he deemed even more concrete and worrying when the artifacts in question were deemed worthy of human Trust and Acceptance.

But (as I asked myself) if Trust and Acceptance were unable to reflect the reasons of his anti-tech rejection of autonomous technologies, then how can we capture, with a rigorous and consistent scientific methodology, the fact that people like Kaczynski advocate the total destruction of the technological apparati that govern human societies while being perfectly comfortable with using technology and interacting with advanced technological artifacts? Clearly, the task required us to differentiate between Acceptance and Trust, on the one hand, and a third attitude –that we called “Tolerance”– on the other.

3.1.3 The ATTA-bot model of Attitudes towards robots

Drawing on the outcomes of these two explorations, my work on Tolerance illustrates that resistance to technology does not inevitably emerge from simplistic emotional or behavioural reflexes automatically elicited by certain Agent characteristics: rather, antagonism toward autonomous agents frequently stems from “an assertive posture primarily motivated by articulate normative judgments and rather elaborate, albeit rigidly categorical, theorizations about Agents in general.” In a way, this violates the popular assumptions of HRI researchers, who focus straightway on the moment of the interaction itself. The concept of Intolerance (i.e., a relatively diminished Tolerance value) expresses how the conceptual frameworks of a person may motivate reluctance or even overt opposition to any plan of adopting Agents. Hence, those beliefs strongly affect users’ readiness to collaborate with advanced systems.

Intolerant mindsets and worldviews can indirectly erode an individual’s Adoption Propensity; crucially, such a stance need not involve overtly hostile statements or conspicuously anxious actions, which can remain hidden or unvoiced. Diminished Tolerance (or Intolerance) directly aligns with “Autonomy Estrangement,” a mental condition marked by uncanny sensations and disquieting (potentially obsessive or paranoid) interpretations concerning autonomous technologies: although having a robust Tolerance for Agents lowers one’s likelihood of attaining this state, an Intolerant viewpoint may lead to Autonomy Estrangement, prompting the gravest behaviours linked to it (e.g., avoiding or vandalizing robots, even committing sabotage). Ultimately, underlying scepticism proves decisive in shaping negative outcomes for advanced systems and can undermine technology acceptance.

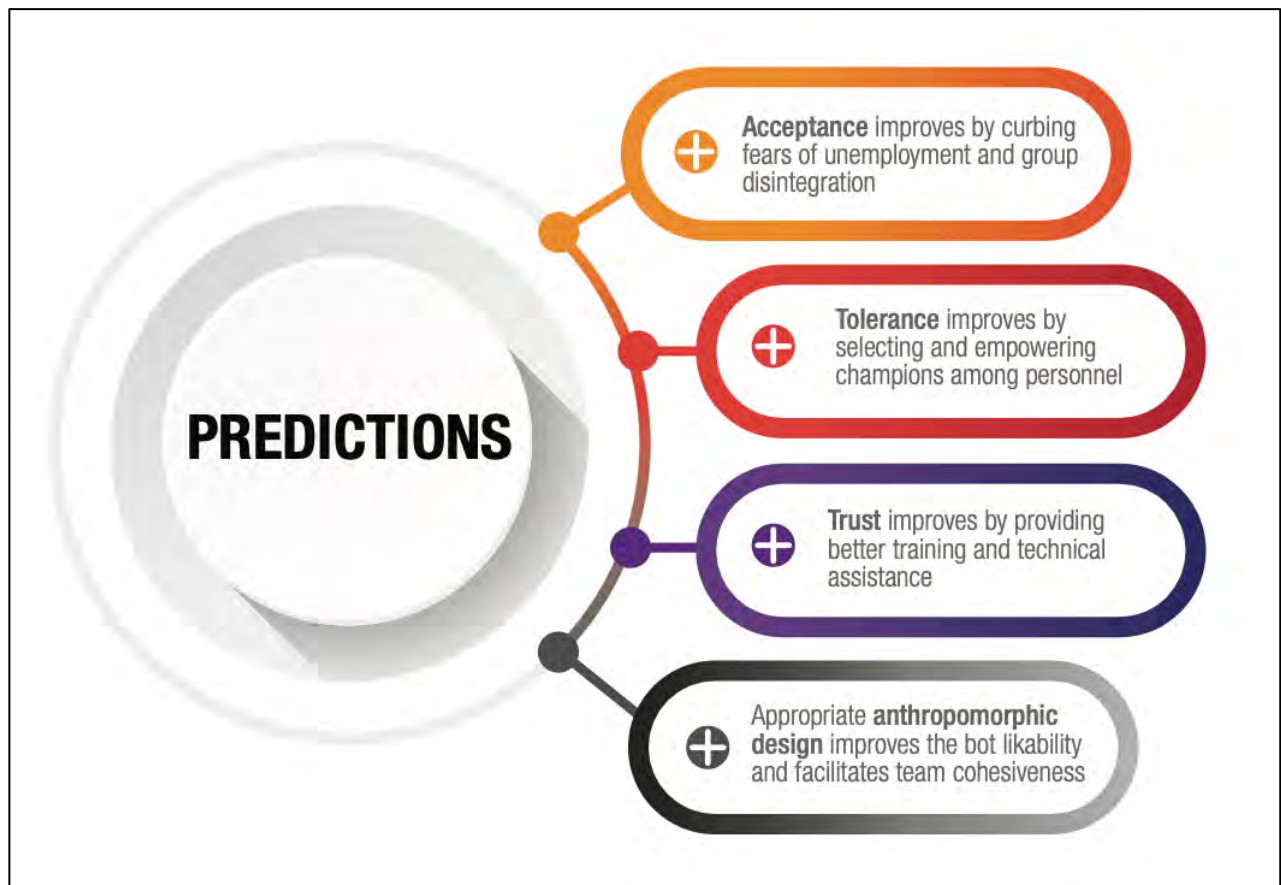
Like Acceptance and Trust, Tolerance (and, in turn, Intolerance) constitutes one of the user dispositions toward Agents that HAI specialists investigate to enhance the assimilation of autonomous technologies in human contexts. Acceptance and Trust gauge whether the user’s expectations, requirements, and priorities are fulfilled during engagements with certain Agents; Tolerance, however, concerns the sweeping philosophical perspectives and frameworks users hold concerning all Agents. These convictions typically manifest through narratives bearing varied degrees of intricacy and plausibility. The narratives that drive Tolerance remain largely detached from the firsthand encounters that foster Acceptance and Trust. Yet our study established that a minimum of three distinct attitudinal components influence Adoption Propensity, underscoring the multifaceted interplay underpinning users’ acceptance or rejection of innovative systems. Generally, HRI research does not thoroughly

separate these three dispositions, as Trust and Acceptance are presumed individually adequate to account for Adoption Propensity.

This distinction eventually brought me to advance what I call the “ATTA-bot” model of attitudes towards autonomy, where the acronym A-T-T-A stands for the four key attitudes that reinforce or diminish our propensity to use a certain technology: “Acceptance”, “Trust”, “Tolerance”, and “Anthropomorphism” (Cappuccio, Devitt, Wrigley 2020 [**SP17**]). Later, in our 2024 IJSR paper [**CP5**], we treat these attitudes as distinct constructs and we offer concrete indications to measure them by means of rigorous metrics (scales) and data-collection protocols (questionnaires).

Table 6 Synoptic table of the differences between key attitudes toward Agents

	Tolerance	Acceptance	Trust
Key question	Is it fair and right that the Agent occupies roles and takes tasks usually assigned to humans?	Does the Agent belong in the work/family environment and is it perceived as part of the team?	Does the Agent operate in an efficacious, efficient, reliable, secure, and safe manner when it conducts a certain job?
Core idea	The Agent deserves to join our group and/or to work in a certain role.	It is nice (pleasant, comforting, rewarding, effortless, etc.) to work with the Agent.	The Agent works well (efficiently, reliably, according to expectations...)
Scope	Categorical normative judgements and beliefs (possibly including unmotivated prejudices and irrational biases) about A.	The quality of the relationship between the Agent and its users, as it is experienced by the users through direct interaction.	The outcome of the work done by the Agent, and whether or not the Agent performed well, based on the users' expectations.
Values	Authenticity, autonomy, self-reliance, agency, purpose.	Familiarity, harmonisation, engagement, inclusive relationship.	Predictability, explainability, reliability, traceability, safety.
Expectation	A meaningful, fulfilling, and self-determined human existence	An appropriate, pleasant, and stable interaction.	An outcome considered desirable and needed and hoped for.
Indicators	Quality of life outside of HAI: self-reliance, purpose	Quality of HAI: familiar, harmonious, inclusive, sanctioned	Quality of the outcome of HAI: expectations and hopes about performance.
Worries and concerns	Concern that Agents may alter the users' customs and practices (A1).	Working with the Agent is a frustrating and irritating experience (as opposed to rewarding and fulfilling).	The information provided by the Agent is irrelevant and difficult to understand (as opposed to comprehensible and clear).
	Concern that Agents may alter the socio-political relations among users (A2).	Interacting with the Agent feels uncanny and creepy (as opposed to familiar, confident and comfortable)	The Agent responds to orders slowly and imprecisely (as opposed to promptly and accurately).
	Concern that Agents may replace humans in their jobs and key functions (B1).	The Agent is an isolated and solitary entity (as opposed to a well-integrated member of the team).	Interaction with/handling of the Agent is unsafe and not secure (as opposed to safe and secure).
	Concern that Agents may undermine the users' authority and decisional power (B2).	The Agent is disliked by people and attracts negative feelings like antipathy and hostility (as opposed to sympathy and friendship).	Previous interactions with Agents were typically unsuccessful or useless (as opposed to successful and useful).
		The Agent's attitude is unhelpful and unavailable (as opposed to helpful and available).	Opacity 1: The reasons/plans motivating the Agent's behaviour/decisions are difficult to understand (as opposed to easily explainable).
	Concern that Agents may question the human exceptionalism, challenging their unique status/entitlements (B3).	The Agent appears opaque and alien, it is difficult to relate to it (as opposed to: Agent solicits empathy, transfer, projection, i.e. people see themselves reflected in it).	Opacity 2: the intended effects of the Agent's behaviour/decisions are difficult to guess (as opposed to easily predictable).
		The Agent's features and behaviours appear rigid, inflexible, stereotyped (as opposed to: they adjust to the team members and their preferences and habits).	The Agent appears intelligent, perceptive, and knowledgeable about its assigned task (as opposed to insensitive, ignorant of the context).
		The Agent displays no collaborative intent and willingness to help (as opposed to being collaborative and involved in the task)	The Agent displays no collaborative intent and willingness to help (as opposed to being collaborative and involved in the task)
			The Agent requires little or no maintenance, technical expertise.
Areas of overlap	No overlap	The relationship with the Agent is affected positively/negatively by the fact that Agent operates more or less reliably and in a clearly explainable/predictable manner.	
Areas of overlap	The relationship with the Agent is affected positively/negatively by the fact that the user has certain biases and prejudices about their status and role.		No overlap
Areas of overlap	The trust in the Agent is affected positively/negatively by the user's biases and prejudices.	No overlap	The trust in the Agent is affected positively/negatively by the user's biases and prejudices.



According to the ATTA-bot model, Acceptance describes the user’s experience of engagement with an Agent and the willingness to recognize said Agent as a legitimate interaction partner or co-worker (De Graaf & Ben Allouch 2013; Han & Conti 2020; Heerink et al 2009; Lee et al 2018; Moradi & Bayat 2018). Trust, in turn, tracks how the user acknowledges the Agent’s capability to reliably operate toward the achievement of a certain goal and the user’s willingness to assign tasks and delegate responsibilities to the Agent. Our model emphasizes not only the distinctiveness, but also the complementarity of Acceptance and Trust: “while Acceptance serves to mapping Adoption Propensity in relation to *the quality, robustness, and easiness of the interaction/engagement* with the user, Trust serves to mapping Adoption Propensity in relation to *the efficacy, efficiency, reliability, and safety of the Agent* in performing their tasks” [CP5].

They remain inevitably blind to some specific kinds of anxiety and hostility that only a notion like Tolerance can effectively account for. That is why we specifically indicate with the term “Tolerance” (for Agents) the user’s insusceptibility or resilience to Autonomy Estrangement, that is the suffocating sense of isolation, displacement, and frustration that human users may experience when they think (for right or wrong reasons) that their lives are (or risk being) controlled by Agents, i.e. by intelligent systems that, while created by

humans, operate independently of human control. Autonomy Estrangement is best theorized in terms of the (real or just feared) impoverishment of the human life allegedly caused by the dominating presence of automated monitoring and control systems. When suffering from Autonomy Estrangement, technology users experience anxiety/disorientation over disruptive changes (configured as disruption of *modus operandi* and way of living and/or disruption of the network of socio-political practices) and/or hostility/frustration against technological dominance (configured as fear of being made redundant, losing authority, and/or losing human uniqueness).

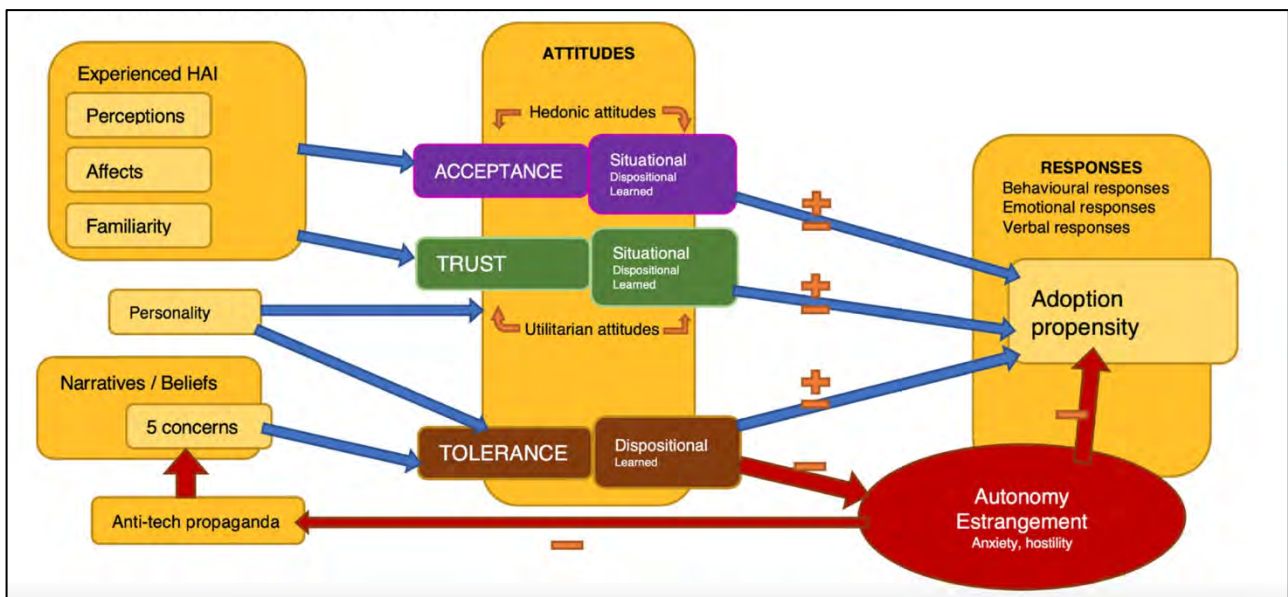
3.1.4 The risks associated with Intolerance

For convenience, we use the term “Intolerance” to indicate the property inverse to Tolerance, that is the user’s susceptibility or proneness to Autonomy Estrangement. Thus, Tolerance and Intolerance are inverse representations of the same phenomenological continuum, with Intolerance increasing when Tolerance decreases and vice versa. Neither Tolerance nor Intolerant tendencies are inherently good or bad: it depends on whether these tendencies are epistemically justified and correctly motivated by a correct evaluation of facts. However, after establishing the in-principle neutrality of the concept of Tolerance, we must also remark that one of the most compelling reasons to introduce Tolerance in HRI research is that, in our world, extremely Tolerant or Intolerant responses *can be symptoms of biased, irrational, and possibly fanatic attitudes*.

Thus, Tolerance is a useful instrument to red flag the kinds of extreme antagonistic attitudes that deserve to be investigated in greater depth: low Tolerance levels increase Autonomy Estrangement, which in turn discourages the adoption of autonomous technologies. Our conjecture is that Autonomy Estrangement negatively correlates with Adoption Propensity and tends to indirectly co-vary with Tolerance for Agents in combination with other variables. Our analysis of Intolerant tendencies towards Agents brought us to formulate six different hypotheses: (i) the most extreme judgments about Agents originate from militant discourses; (ii) Tolerance correlates with some aspects of the individual’s psychological profile, such as personality (personality traits like “openness” and “neuroticism” and personality-related variables like “affinity for technology interaction” or “technology commitment”); (iii) Intolerant attitude tends to produce less obvious and more covert behavioural manifestations, compared to low Acceptance or Trust, as those holding Intolerant beliefs are often wary of publicly manifesting their concerns and anxieties concerning technology; (iv) Intolerance for Agents shares a common root with intolerance

for humans and possibly resonates with it, although these two tendencies have very different ethical implications.

Another reason to introduce a new metric in HAI research is that the distinction between Tolerance, on the one hand, and Acceptance and Trust, on the other, reflect a key distinction between different kinds of stakeholders: the *users*, i.e., those whom will have to work with (and possibly be replaced by) Agents; and the *owners*, i.e., decision-makers like employers or political authorities, who purchase Agents because they find their adoption convenient. Their perspectives do not necessarily align and often tend to be diametrically opposed. Whereas an account that relies only on Acceptance and Trust might conceal this divergence, a Tolerance-based account is likely to show any polarisation between these two groups.



My characterization of Tolerance reflects how this attitude is non-embodied, unlike other attitudes, such as Acceptance and Trust, which are clearly embodied. First, Tolerance reflects the hostile or anxious anti-tech *narratives*, while Acceptance and Trust reflect the *interactive experiences* of humans with robots. Second, Tolerance is concerned with *categorical judgments about Agents*, while Trust and Acceptance concern the *particular instances of interaction with an Agent*. Third, Tolerance concerns a subject's *normative* beliefs, while Acceptance and Trust concerns Agents' appearance and behaviour. Fourth, Tolerance is a *context-independent dispositional attitude* acquired primarily through learning, while Acceptance and Trust are primarily determined by the *situational context* of HRI. Fifth, Tolerance forms in an *offline* (possibly long-term) *dimension*; while Acceptance and Trust form in the *online dimension* (a *short-term time scale*) considering temporally specific events. Sixth, Tolerance expresses a contentful *knowledge-that*, while Acceptance and Trust primarily attest a *direct performative*

familiarity (know-how) developed through practical interaction with robots. In short, Tolerance tracks the subject's normative, categorical, acontextual narratives and beliefs that reflect their factual knowledge of Agents in a detached and atemporal dimension; conversely, Acceptance of Trust track the subject's descriptive accounts of contingent, practical instances that reflect the subject's interactive competence in a situationally engaged and temporally specified experiential dimension.

To sum up, Trust, Acceptance, and Tolerance are key for the methodology of social robotics and HRI research: Trust indicates the expectations related to the robot's intelligence and general ability to competently carry out its tasks; Acceptance measures how likeable and credible a robot can be as a social companion and a co-worker; Tolerance refers to the easiness of transferring to technology the authority and responsibility of decisions traditionally made only by humans. While inherently qualitative in nature, these psychological aspects of HRI must be carefully considered by engineers if they want to design efficacious applications of AI. This is particularly obvious in the domain of robotics in warfare: the way Agents are perceived by human troops makes a difference to what they can do together as part of a hybrid military soldier-robot team, as trust, acceptance, and tolerance contribute to define the function and the scope of their interaction.

The three attitudes examined so far presuppose and depend on a more fundamental phenomenon, called anthropomorphism, which deeply influences them but it also fundamentally distinct from all of them (Bishop et al 2019). Anthropomorphism is the fourth key attitude in my ATTA-bot model of HRI. It is well documented that implicit anthropomorphic tendencies emerge spontaneously and pervasively from HRI (Hegel et al 2008): anthropomorphism is first of all a universal stance, a structural, very basic feature of the human perceptual system (Marchesi & Wykowska, 2024), and only secondarily a judgment deriving from our system of beliefs (Dacey 2017). As such, it is not revisable and human users have little power over their perceptual experience, although they certainly have control over the emotional effects of anthropomorphising a machine (Bruni, Perconti, & Plebe 2018). In this sense it is very different from Tolerance, which largely depends on one's deliberate, conscious adhesion to a set of beliefs and articulated world views.

Table 5 Designing a scale for Autonomy Estrangement—sample questions

	Questionnaire to measure Autonomy Estrangement “How often do you experience the following feelings/wishes?” 5 possible answers (score 0–4): “Never” (0), “Just once or twice” (1), “Occasionally” (2), “Often” (3), “Consistently” (4)
Queries	(I.1) I feel that autonomous technologies make my work repetitive and boring
group I	(I.2) I feel that autonomous technologies make my activities riskless and predictable
Anxiety and disorientation	(I.3) I feel that autonomous technologies make my life empty and meaningless (I.4) I feel that autonomous technologies make my social relationships superficial and fake (I.5) I feel that autonomous technologies make my competency and skills useless and obsolete (I.6) I feel that autonomous technologies make my decisions pointless and insignificant
Queries	(I.7) I feel that autonomous technologies make my actions powerless and ineffective
group II	(I.8) I feel that autonomous technologies make my experiences insipid and flat
Hostility and frustration	(I.9) I feel that autonomous technologies make my role redundant and unneeded (I.10) I feel that autonomous technologies make my judgment disrespected and unauthoritative (I.11) I feel that autonomous technologies make my projects irrelevant and inconsequential (I.12) I feel that autonomous technologies make my dreams and desires futile and worthless (II. 1) I wish engineers and designers stopped developing new Agents (II. 2) I wish manufacturers stopped selling new Agents (II. 3) I wish families stopped using Agents as home assistants (II. 4) I wish managers stopped adopting Agents in their offices (II. 5) I wish businesses stopped relying on Agents (II. 6) I wish a global industrial crisis dramatically reduced the production of Agents (II. 7) I wish the parliament passed stricter laws to limit the use of Agents (II. 8) I wish citizens were more aware of the threat represented by Agents (II. 9) I wish consumers associations took action to limit the diffusion and the usage of Agents (II. 10) I wish a pirate virus disabled many AI-based devices in the world, including Agents (II. 11) I wish all Agents were forbidden and all their manufacturers went out of business (II. 12) I wish movements of activists started to systematically inactivate Agents everywhere (II. 13) I wish engineers and designers were punished for creating Agents
Autonomy Estrangement measured on a 5-degree scale	<i>Measurement method (tentative): sum the scores (from 0 to 4) of each of the 25 answers</i> <i>Calculate the total score (from 00 to 100)</i> 00 to 20 No Autonomy Estrangement 21 to 40 Sporadic or minimum Autonomy Estrangement 41 to 60 Minor Autonomy Estrangement 61 to 80 Major Autonomy Estrangement (may involve mild robo-phobia/techno-phobia) 81 to 100 Extreme Autonomy Estrangement (may involve severe robo-phobia/techno-phobia)

3.1.5 A peculiar attitude: Anthropomorphism

What exactly is Anthropomorphism and what are its cognitive underpinnings? According to the most common definition, anthropomorphism is the tendency to attribute to non-human entities the characteristics of human ones, transferring to the former psychological attitudes and forms of engagement that are typically reserved to the latter (Epley, Waytz & Cacioppo 2007; Damiano & Dumouchel 2018). Among the characteristics that can be projected onto a non-human, possibly inanimate, entity there are both basic features (goal-oriented, adaptive behaviour and the capability to selectively sense and track contextually relevant cues) and more cognitively complex features like knowledge states (beliefs), objectives (desires), and plans to achieve them (intentions), which arguably involve propositional and semantic content. Social robotics research has highlighted that behavioural anthropomorphism correlates to higher levels of likeability and perceived intelligence of robot which, in turn, predict more resilient trust and robust interest (Bartneck et al 2009b). Anthropomorphism is not always intentional and explicit. Rather, it often emerges at a pre-reflective level, embedded in a set of preferences or subconscious behavioural dispositions²⁷.

The ATTA-bot framework recognizes that Anthropomorphism differs markedly from the other three attitudes because, unlike Acceptance, Trust, and Tolerance, it does not exert a straightforward, uniformly positive or negative influence on adoption propensity. Instead, it amounts to the spontaneous tendency to perceive human or animal-like traits in inanimate objects (Złotowski et al 2015). In my studies on this topic, I have consistently noted that such a tendency is constantly at work, shaping a significant portion of our disposition toward robots, especially humanoid ones (Cappuccio, Galliot & Sandoval 2021a [SP16]). Consequently, when anthropomorphism is high—i.e., when we strongly ascribe human-like properties to a machine (intelligence, emotions, desires, etc.)—it can influence how we interpret and judge that machine's appearance or behaviour in a variety of ways, including amplifying our favourable or unfavourable assessments of its performance. Because the

²⁷ Anthropomorphism can be either *implicit* or *explicit*: it is regarded as implicit when a subject pre-reflectively responds to an inanimate entity as though it were animated, thereby acknowledging through their behaviour or conduct the bestowal of animacy or agency (Ziemke 2023); anthropomorphic attributions are deemed explicit if they are expressed through verbal behaviours (Złotowski et al 2017). The former can operate automatically, at the level of rapid, immediate perceptual and habitual responses of which the agent may remain unaware, whereas the latter operates through conscious, reflective regulation, at the level of judgment and rational evaluation. Explicit anthropomorphism may also quite frequently originate from the very intentional pursuit of distinct human-like characteristics—for example, as part of the personification of nature, which crucially underlies figurative or poetic creation.

ways in which anthropomorphism affects Trust and Acceptance can vary widely and remain largely unpredictable, its real impact on adoption propensity requires empirical, case-by-case investigation (Epley, Waytz & Cacioppo 2007; Eyssel & Kuchenbrandt 2012; Waytz, Cacioppo & Epley 2014; Kaplan, Sanders & Hancock 2019). Certain anthropomorphic effects resist reproduction under laboratory conditions, so they must be examined either in real-world settings or by relying on anecdotal evidence.

In the article, “Saving Private Robot: Risks and Advantages of Anthropomorphism in Soldier-Agent Teams” (Cappuccio, Galliot & Sandoval 2021b [CP6]), I explored this phenomenon specifically within the domain of soldier-robot teaming, i.e., military uses of HMI. Produced during my work for the Minerva Initiative and published in the *International Journal of Social Robotics*, the paper reviews anecdotal accounts from soldiers stationed in Afghanistan and Iraq who grew unexpectedly attached to the explosive ordnance disposal robot they employed. In such contexts, a military robot’s pseudo-social dimensions (whether humanoid, zoomorphic, or an unmanned vehicle) can critically shape how a combined human-machine team fulfills shared objectives, contingent on how well they select suitable modes and signals to facilitate communication.

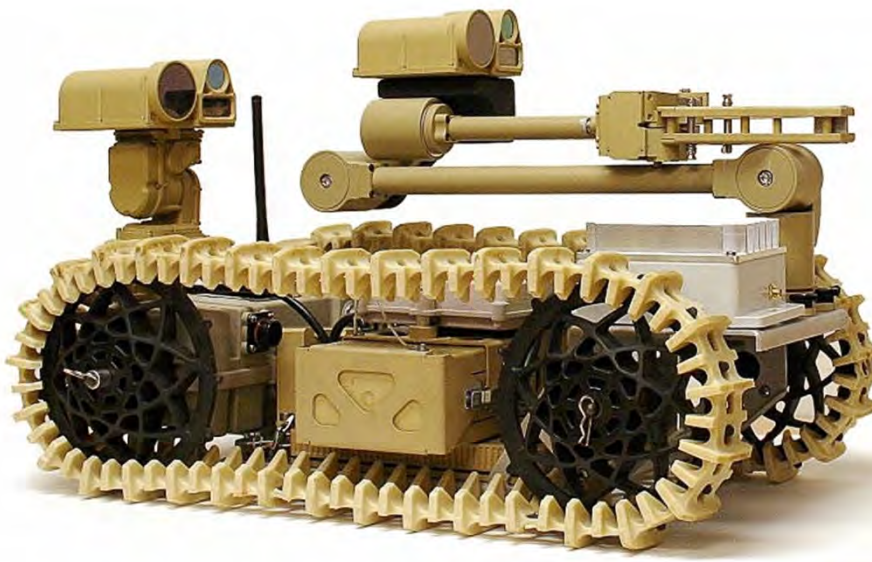
Although it is vital to track anthropomorphic tendencies within a human-robot team to prevent unwarranted attributions of trustworthiness or intelligence, endeavouring to eliminate such tendencies outright is neither feasible nor desirable given their pervasiveness and tenacity. Accordingly, in CP6, I address the potential hazards and gains of anthropomorphism for military teams composed of human soldiers and autonomous artificial agents. In these scenarios, “System 1” processes (fast, automatic, pre-reflective, to use, according to Dual Processing theory) govern natural communication (through body language, signals, etc.), which is crucial for swift, practical, cognitively efficient coordination among the human members and their synthetic teammates.

Embodied perception of anthropomorphic characteristics is more basic and primordial than inference-based or rational deliberation (Marewski, Gaissmaier & Gigerenzer 2010). Thus, it does not compete with higher-order cognition but rather operates in parallel, supporting it. Indeed, embodied and perceptual responses (including anthropomorphic ascriptions) do not necessarily stifle contextual, adaptive intelligence; on the contrary, by enabling empathy as an embodied, situated mode of understanding others, such perceptions enrich social and emotional intelligence. They also furnish the direct involvement that, for instance, allows a soldier to coordinate smoothly with teammates in the field. As an outcome of processes rooted in our embodied perceptual capacities,

anthropomorphic attributions strongly predict social reactions, and social engagement requires at least some rudimentary form of anthropomorphic perception. Identifying an agent as anthropomorphic is a necessary, though not sufficient, condition for genuine social interaction: authentic “social” engagement arises only when it connects entities that are—or appear to be—alive (Coeckelbergh 2011, 2018; Damiano & Dumouchel 2018).

3.1.6 The dark side of anthropomorphism: unsafe relationships

Anthropomorphism is often accused to fuel the tendency to misattribute mental states to robots and overstate their capacities? These tendencies are often indicated as a risk factor, because they could allegedly create inflated expectations in the human users, undermining the effectiveness of HRI and even creating operational dangers for those who work in proximity with robots that are neither intelligent nor autonomous as the human users believe. We have seen that the inflated expectations about robots’ performance may conduce to catastrophic outcomes in delicate tasks that require a perfect real-time coordination of intentions and actions among all the interactants. But in addition to this, there is another risk: that users may establish delusional bonds with robots, projecting inappropriate emotions towards their synthetic companions and teammates. This is often indicated as a detrimental and risky downside of anthropomorphism, to the point of discouraging the use of robots with humanoid features (not only heads and faces, but also bipedal posture) in sensitive team scenarios (Scheutz 2012).



The Increment 1 Platform is an unmanned ground vehicle (UGV) prototype developed by Northrop Grumman Corporation and U.S. Navy as part of the Advanced Explosive Ordnance Disposal Robotic System (AEODRS) program. This vehicle can operate autonomously and is designed to carry out tasks like the disposal of explosive ordnance and improvised explosive devices and the handling of hazardous materials

For example, in the literature concerning military soldier-robot teaming, we can find frequent mentions of cases of service members allegedly developing some kind of attachment for the robots they use (Scharre 2018): this includes soldiers mourning the accidental destruction of explosive-disposal robots during their dangerous operations or even soldiers concerned for the integrity of robots during deployment in battle scenarios, to the point of putting at risk their own safety to prevent that the robots could be damaged in combat. While scarce and prevalently anecdotal, such cases attracted the attention of military decision makers and aroused the imagination of the public, highlighting a very serious concern: that our subconscious anthropomorphistic tendencies could elicit in us emotions and judgments towards robots that machines do not deserve and are not worthy of, hence creating unwanted and problematic consequences. The main worry is that, when unrecognized, this kind of subconscious impulse could make users prone to an unnecessary consideration for the welfare of robots, motivating a misuse of the robot (Dacey 2017). This is the kind of safety concern that we explore in the paper “Saving private robot” (Cappuccio, Galliot & Sandoval 2021b [CP6]), where the bond established by soldiers with robots is not only unmotivated (the robots in question were not built to have a social function or to entertain affective relations) but also physically harmful for the users. However, in this paper we try to highlight that anthropomorphism has also a bright side, in addition to the dark one.

3.1.7 The bright side of anthropomorphism: advantageous applications

Military professionals can and often do subconsciously attribute anthropomorphic features to Agents (with the possible attachment feelings that may or may not derive from this projection), even if they perfectly know, at the conscious level, that robots and AIs are just mechanical devices devoid of emotional states, proper motivations, and purpose. How to make sense of this (apparently contradictory) duality? I have proposed a “pragmatic,” approach to anthropomorphism, that radically oppose two other approaches that are prevalent in the literature (I call them the “lax” and the “anxious” approach, respectively). The Pragmatic Approach allows us to understand the distinction between conscious (explicit) and the subconscious (implicit) levels of anthropomorphism not as a problematic gap between two independent systems, but as two different moments of the same integrated process subserving social cognition in HRI. When we subconsciously attribute anthropomorphic attributes to an Agent, this process is neither intrinsically irrational nor the delusional by-product of hyper-stimulated imagination: like empathy, anthropomorphism is rather an informative modality of aesthetic perception, and as such it can be wisely used by designers (DiSalvo & Gemperle 2003). Through it, intersubjective coordination and embodied communication articulate themselves (Airenti 2018).

Approaches to anthropomorphism/defining features	The 'Lax' approach	The 'Anxious' approach	The 'Pragmatic' approach
Key examples in the literature on military Agents	Barnes and Evans [17]	Carpenter [22]	This paper
Psychological processes underlying anthropomorphic attributions	Conscious level: imaginative processes with semantic (verbal and/or symbolic) content	Subconscious level: emotional (irrational, instinctual, and likely non-verbal) processes	Multi-level: it includes both empathic (perceptual and embodied, analogic) and cultural (symbolic, verbal) processes (e.g., narratives of sacrifice)
Outcome	Metaphoric projections, symbolic imagery	Development of attachment feelings	May vary significantly. Social-interactive infrastructure for communication and coordination
Problem with anthropomorphism	An excess of cognitive activity: overthinking, overimagining	A lack of cognitive activity: absence of rational judgment	A constitutive component of cognitive activity: it enables judgment and recognition
Epistemic value of anthropomorphic attributions:	A-rational (no connection between perception and judgment)	Irrational (open conflict between perception/feeling and judgment)	Pre-rational (perception/feeling can scaffold judgment and shape its content)
Underling cognitive architecture	Cartesian model of cognition: dualism of perception and judgement	Cartesian model of cognition: dualism of perception and judgment	Embodied model of cognition: integration of perception and judgment
Practical value of anthropomorphic attributions	They are neither risky nor beneficial	They expose tactical operations to significant risks	They are indispensable to support communication and coordination in HAI. They have only very modest risks
Recommendation for designers and developers to improve HAI	Indifferent to anthropomorphic design elements	Avoid or minimize anthropomorphic design elements	Empirically investigate and selectively implement the advantageous aspects of anthropomorphic design

Contrary to mainstream safety narratives, designing agents capable of communicating effectively with civilians (who may be unprepared for high-pressure encounters, see Scharre 2018) is both possible and desirable. Therefore, these agents' appearance, behaviour, and communicative elements must be developed with careful attention to a broad range of human factors, such as local customs, socio-cultural norms, and ethical standards, both for the military personnel who deploy them and the civilians they engage. Critically, accounting for anthropomorphic tendencies in the design process can amplify benefits while reducing unwanted consequences. Potential benefits to military users include multiple forms of coordination and synergy: (i) *Coordination Between Same-Unit Members* – Agents can generate and interpret information conveyed through bodily, emotional, purposeful signals, instead of relying solely on explicit commands. Gesture, gaze, mimicry, proxemics, posture, movement style, and vocal tone are some cues agents can use to support cooperative actions and joint tasks. These features help sustain synergy; (ii) *Coordination Between Multiple Allied Units* – In large multi-front operations, Agents can maintain optimal situational awareness by automatically integrating critical information from various battle fronts, and such systems are already under development; (iii) *Coordination Between the Military Unit and Civilians* – Robots deployed in peace-keeping tasks, for example at checkpoints or during inspections, leverage anthropomorphism to convey suitable emotional cues, stabilize interactions with civilians, and ultimately foster trust with them in the process; (iv) *Communication with Adversaries* – Properly designed devices can deter opponents or elicit specific responses for deceptive or diversionary aims, such as strategically distracting and delaying them when required.

Therefore, despite the alleged risks, anthropomorphism remains key in the military applications of human-robot teaming because artificial agents, no less than human troops, must communicate correctly (which means naturally, spontaneously, quickly, efficiently) with subjects that are untrained in emotionally loaded situations that can easily escalate.

Recent applications of social robots as conflict mediators might be useful to achieve this objective. Shape, friendly look, size, behavioural patterns, mimicry, velocity, and mechanical features resembling faces, eyes, or other human or animal body parts can be anthropomorphized at any time and therefore indicate a threatening or friendly presence. In general, the impact of anthropomorphism on ethical behaviour is not to be underestimated, as specific anthropomorphic affordances (i.e., robotic features that tend to trigger anthropomorphic attributions in the users) may produce complex patterns of benefits and risks in the moral domain. The following table schematically describes some of these anthropomorphic affordances²⁸.

Summary table: Anthropomorphic affordances and ethical behaviour

Aim / Application (example use)	Anthropomorphic affordance (what's used)	Anticipated benefits for virtue cultivation	Main ethical risks & concerns	Design notes / mitigations	Representative references
<i>Moral attention & habit scaffolding</i> (daily prompts for prosocial acts)	Friendly face/voice, social presence, name & turn-taking	Increases mindful self-regulation; supports habit formation and reflective practice (VRx framing)	Dependency, motivational crowding-out; manipulative nudging	Disclose intent; opt-in, adjustable prompt frequency; sunset policies	Cappuccio et al., 2021 (VRx); Fogg, 2003; Berdichevsky & Neuenschwander, 1999; Vallor, 2016
<i>Empathy training & perspective-taking</i> (role-play)	Expressive eyes/gaze, contingent affect	Embodied rehearsal of empathy; moral imagination	<i>Deception</i> ("as-if" care), misplaced trust; emotional overidentification	Clear simulated status; reflective debriefs; human-in-the- loop facilitation	Breazeal, 2003; Coeckelbergh, 2020; Vallor, 2016; Reeves & Nass, 1996
<i>Social skills coaching in autism</i> (e.g., joint attention)	Gaze cues, prosodic feedback, touch responses	Structured, low- stakes social practice; clearer cues improve learning	Overgeneralisation across contexts; miscalibrated feedback	Therapist- supervised protocols; clinical validation;	Wainer et al., 2014 (KASPAR); Alnajjar et al., 2021;

²⁸ In addition to the problems mentioned in this table, the adoption of anthropomorphic robots involves other ethical challenges, such as false expectations, biases, over-reliance and moral deskilling. These risks must be mitigated by designers and developers, adopting specific safeguards or interventions.

Aim / Application (example use)	Anthropomorphic affordance (what's used)	Anticipated benefits for virtue cultivation	Main ethical risks & concerns	Design notes / mitigations	Representative references
<i>Character education in schools</i> (courtesy, cooperation)	Classroom persona, peer-like dialogue	Engagement; concrete exemplars of virtues-in-action	Value imposition/cultural bias; moral deskilling of teachers	Participatory co-design; plural values; rotate "virtue- of-the-week"	Vallor, 2011; Sharkey & Sharkey, 2010; Danaher, 2016
<i>Eldercare companionship</i> (loneliness reduction)	Names, backstory, small talk, empathic backchannels	Enhances relatedness, routine adherence (meds, exercise)	Anthropomorphic deception; dignity/consent; privacy	Identity transparency; data minimisation; consent refresh	Sparrow & Sparrow, 2006; Sharkey & Sharkey, 2012; Riek & Howard, 2014
<i>Ethical reflection prompts</i> (workplace/compliance)	Polite reminders, "Why/what if" questions	Gentle norm cueing; reduces akrasia (weakness of will)	Slippery slope to behavioural manipulation; autonomy concerns	Just-in-time prompts with skip; logging for accountability	Thaler & Sunstein, 2008; Hausman & Welch, 2010; Bryson, 2018
<i>Conflict de-escalation & civility</i> (customer service)	Calm prosody, empathic mirroring	Models civility; reduces escalation	Masking power asymmetries; emotion work offloading	Usage policies; escalation to human agents; audit logs	Coeckelbergh, 2020; Darling, 2015; Nyholm, 2020
<i>Health/fitness adherence</i> (self- control)	Coach persona, praise/commitment devices	Builds temperance, perseverance	Shame/pressure; harmful comparisons	Personalised, non- judgmental framing; user- set goals	Fogg, 2003; Consolvo et al., 2009; Vallor, 2016
<i>Sustainability & civic virtue</i> (pro- environmental nudges)	Social proof messages, "we" framing	Encourages stewardship, collective goods	Greenwashing; normative overreach	Evidence- backed claims; opt- out; impact transparency	Sunstein, 2015; Cialdini, 2007; Vallor, 2016
<i>Child safety & online conduct</i> (digital citizenship)	Mentor persona; stories & moral exemplars	Narrative encoding of norms; transfer to behaviour	Moralising tone; reactance; cultural mismatch	Co-created stories; parent/teacher mediation	Narvaez, 2006; Kim & Hannafin, 2011; Cappuccio et al., 2021

3.1.8 Anthropomorphism for robots as empathy for artificial agents

The fact that soldiers anthropomorphise their robotic teammates is not only to be expected but also, in some measure and under specific constraints, potentially beneficial. This conclusion is based on the theories of empathy, which understand anthropomorphism as an embodied perceptual process based on internal simulation and a projective process of bestowal/attribution. Empathy is a key component of social expertise and an enabler of altruistic behaviour, but it is also the ability to perceive the other's body as alive and sensible. We automatically tend to attribute the emotions, feelings, and intentions to other agents, whether they are human or artificial, so long as they display the behaviour that we use to express the same mental states (Suzuki et al 2015). In the same way, empathy allows us to assign intelligible meanings to inanimate entities, personifying them.

Similarly, anthropomorphism allows us to see robots' bodies as bodies that feel and want and expect (Cappuccio 2009 [SP1], Asada 2016). This "mirroring", or analogical understanding, plays a key role in social cognition, as it represents a distinctively cross-modal and pre-conceptual mechanism for understanding the other's goals, emotional state, or thoughts (Calvo-Merino et al 2006; Gallese 2007; Gallese & Sinigaglia 2011, Dumouchel & Damiano 2011, 2015b; Airenti 2015). Mirroring constitutes also the cognitive basis for anthropomorphism, which is a key aspect of embodiment in the field of social robotics and HRI (Riek et al 2009). If empathy is the perceptual, automatic attribution of goals and mental states to the body of a generic other, anthropomorphism is the analogous process by means of which we project goals, intentions, emotions and other mental states to non-human, possibly inanimate entities. It is perfectly normal that anthropomorphic attributions happen when we observe entities that objectively have an humanoid feature but the truth is it is possible to anthropomorphise also perfectly non-humanoid entities. Anthropomorphism is key in HRI although it intervenes only indirectly in the social interaction process, not changing the living quality of objects in our experience but rather modulating the attitudes of human interactants toward the robots they interact with, thus contributing to determine their level of trust, acceptance, and tolerance and, subsequently, their overall disposition to use the robots and establish a social relationship (or pseudo-relationship) with them (Damiano, Dumouchel & Lehmann 2015a, 2015b). Starting with iCub by Metta and Sandini, several concept humanoid robots have been developed to explore artificial empathy, testing the notion that the motoric competency necessary to effectively perform a certain goal-oriented action can be reused by the system, during the observation of human actions, to facilitate the recognition of the actions' goal types (Metta et al 2006; Asada 2014, 2015a, 2015b, 2016; Damiano, Dumouchel, & Lehmann 2015a).

Compared to the cognitivist theories of the mind, the embodied approaches to social cognition, based on interaction and empathy, are better positioned to account for the phenomenon of anthropomorphism that emerge in the military applications of HAI. As the embodied approaches describe automatic low-level responses that are built in one's pre-reflective perceptual and practical dispositions and don't require explicit cognitive effort, they are able to capture the specificity of fast-paced, emotionally charged, and practically characterized situations, such as the tactical operations in which Agents are anthropomorphized by service members.

3.1.9 Outcomes, limitations, and future research

This section has explored human-robot relationships from the point of view of the embodiments of the agents involved, whether they are human or artificial, to highlight the importance of human attitudes in HRI and, specifically, the moral significance of these attitudes. The two studies explored in this section attest that the moral implications of attitudes towards robots cannot be taken for granted, as too often it is done in the HRI literature: trusting robot is not necessarily a good thing and designing robots that are trustworthy is not enough to dispel the concerns and doubts of the human users; also, anthropomorphizing robots is not necessarily a bad thing, as it is very often assumed, but it is a key component of the process of relation-building that is central in HRI.

The design of intelligent interactive robots must consider not only the diversity of attitudes, their variability across different groups, and their reciprocal, implicit interconnectedness, but also the fact that these attitudes are structural components of a team activity. Thus, instead of simply anticipating how attitudes can be increased or suppressed, the design and the conceptualization of new robots must interpret attitudes strategically, asking in the first place what HRI is intended to achieve and whether and to what extent and for what purpose a proper relationship with a robot is desirable. This is the kind of considerations that researchers in the emerging fields of "value-sensitive design" and "meaningful human control" have already started to explore, unearthing the deep links between technology creation, ethical concerns, and socio-cultural expectations.

The main limitation of these studies consists in the fact that our models haven't been extensively tested on human subjects yet. To identify patterns of Trust, Acceptance and Tolerance in different kinds of users and the adopters of autonomous robots, our future studies, to be conducted by means of surveys, focus groups, and semi-structured interviews, will be crucial, as they will allow us to definitively validate the ATTA-bot model of human

attitudes. This model is anyway based on an extensive review of the existing literature, on a careful analysis of the published results, and a preliminary exploration of attitudes and beliefs in real-life scenarios.

Core Publication #5 [CP5]

Cappuccio M.L., Galliot J.C., Lanteri A., Eysel F. (2024),
“Autonomous Systems and Technology Resistance:
New Tools for Monitoring of Acceptance, Trust, and Tolerance”
International Journal of Social Robotics, 16(1), Springer Verlag.

Reproduced with permission from Springer Verlag.
All rights reserved.



Autonomous Systems and Technology Resistance: New Tools for Monitoring Acceptance, Trust, and Tolerance

Massimiliano L. Cappuccio¹  · Jai C. Galliot⁴ · Friederike Eyssel² · Alessandro Lanteri³

Accepted: 15 September 2023 / Published online: 6 November 2023
© The Author(s) 2023

Abstract

We introduce the notion of *Tolerance* for autonomous artificial agents (and its antithetical concept, *Intolerance*), motivating its theoretical adoption in the fields of social robotics and human–agent interaction, where it can effectively complement two contiguous, but essentially distinct, constructs—*Acceptance* and *Trust*—that are broadly used by researchers. We offer a comprehensive conceptual model of Tolerance, construed as a user’s insusceptibility or resilience to *Autonomy Estrangement* (i.e., the uncanny sense of isolation and displacement experienced by the humans who believe, for right or wrong reasons, that robots can subvert and/or control their lives). We also refer to Intolerance to indicate the opposite property, that is the user’s susceptibility or proneness to Autonomy Estrangement. Thus, Tolerance and Intolerance are inverse representations of the same phenomenological continuum, with Intolerance increasing when Tolerance decreases and vice versa. While Acceptance and Trust measure how the user’s interaction with a particular robot is satisfying and efficacious, the dyad Tolerance/Intolerance reflects how the user’s attitude is affected by deeply held normative beliefs about robots in general. So defined, a low Tolerance (that is a high Intolerance) is expected to correlate to antagonistic responses toward the prospect of adoption: specifically, Intolerant attitudes predict the kind of anxious and hostile behaviours toward Agents that originate from the concerns that autonomous systems could deeply disrupt the lives of humans (affecting their work cultures, ways of living, systems of values, etc.) or dominate them (making humans redundant, undermining their authority, threatening their uniqueness, etc.). Thus, Negative beliefs and worldviews about Agents are the cause of the Intolerant attitude toward Agents, which predicts Autonomy Estrangement, which in turn correlates to low Adoption Propensity and avoidance and rejection behaviours.

Keywords Technology acceptance model · Trust · Tolerance · Human–robot interaction · Autonomy estrangement · Alienation

1 Tolerance: a New Tool for Human–Agent Interaction Research

Rightly or wrongly, many people believe that the unrelenting rise of autonomous technologies will transform our lives in disturbing and alienating ways: technological unemployment [1], depersonalization of relationships [2], and threats to human uniqueness [3], are but some of the dystopic scenarios evoked by such narratives [4]. Whether these worries are well founded or not, their adverse impact on innovation is being taken seriously by the experts of human–machine interaction [5]. Nonetheless, a metric to rigorously measure

✉ Massimiliano L. Cappuccio
m.cappuccio@outlook.com

Jai C. Galliot
jaichristiangalliot@gmail.com

Friederike Eyssel
friederike.eyssel@uni-bielefeld.de

Alessandro Lanteri
alelanteri@gmail.com

¹ Trusted Autonomy Group, School of Engineering and Information Technology, University of New South Wales Canberra, Canberra, ACT, Australia

² Center of Excellence for Cognitive Interaction Technology, Universität Bielefeld, Bielefeld, Germany

³ ESCP Business School, Turin, Italy

⁴ Faculty of Business & Law, University of Wollongong, Northfields Ave, Wollongong, NSW, Australia

the negative influence of these worries and concerns on the prospective adopters of autonomous systems is still lacking.

In this paper, we introduce “Tolerance”, a theoretical construct designed to measure the resilience or insusceptibility of human users toward the effects of alienating convictions and dystopian narratives about *autonomous artificial agents* (henceforth referred to as “Agents”¹). Agents are intelligent interactive systems, such as mobile robots, unmanned vehicles, and digital avatars, capable to independently make decisions and interact (physically and/or socially) with complex environments. Many Agents, like co-bots and social robots, are designed to assist humans in their daily activities, communicate, and complete joint tasks with them [6, 7]. Our analyses motivate the adoption of the concept of Tolerance in the context of social robotics and *human–Agent interaction* (HAI)² [8]. In these fields, Tolerance could effectively complement two contiguous, but essentially distinct, notions—*Acceptance* and *Trust*—that are broadly used by researchers [9, 10].

Tolerance for Agents was first conceptualized by Galliot as a measure of the resilience or immunity to the sense of anxiety and hostility that often accompanies the prospect of adopting autonomous technologies [11]. Galliot et al. [12] examined, from a security perspective, the rise of neurotic, possibly paranoid, attitudes towards highly automated and autonomous systems, highlighting the potential risk of non-compliance among the defence personnel who adverse the prospect of operating with or under such systems. This investigation pointed to Theodor Kaczynski’s theoretical elaboration as an epitome of low Tolerance for Agents. Both Kaczynski’s infamous manifesto and his disgraceful actions represent a lasting testament to how estranging narratives and intricate cogitations about automatization can motivate a radical, and in the most extreme cases even violent, rejection of all autonomous technologies [2, 13].³

If there is one thing to be learned from Kaczynski’s notorious case, it is that this kind of radical form of rejection (normally reduced by sociologists and social psychologists

to “technology resistance” [15]) is not a stereotyped emotional and behavioural reaction mechanically triggered by the objective features of this or that particular Agent, but rather an *assertive posture primarily motivated by articulate normative judgments and rather elaborate, albeit rigidly categorical, theorizations about Agents in general*. The notion of Intolerance captures how the content of these theorizations motivates the subject’s unsupportiveness or even active antagonism toward any prospect of adopting Agents. By soliciting or disinhibiting negative responses, the content of Intolerant beliefs and worldviews indirectly contributes to undermine the user’s Adoption Propensity; importantly, such an attitude may, but does not necessarily have to, be accompanied by openly hostile verbalizations and visibly anxious behaviours, which may, and often do, remain covert or unexpressed.

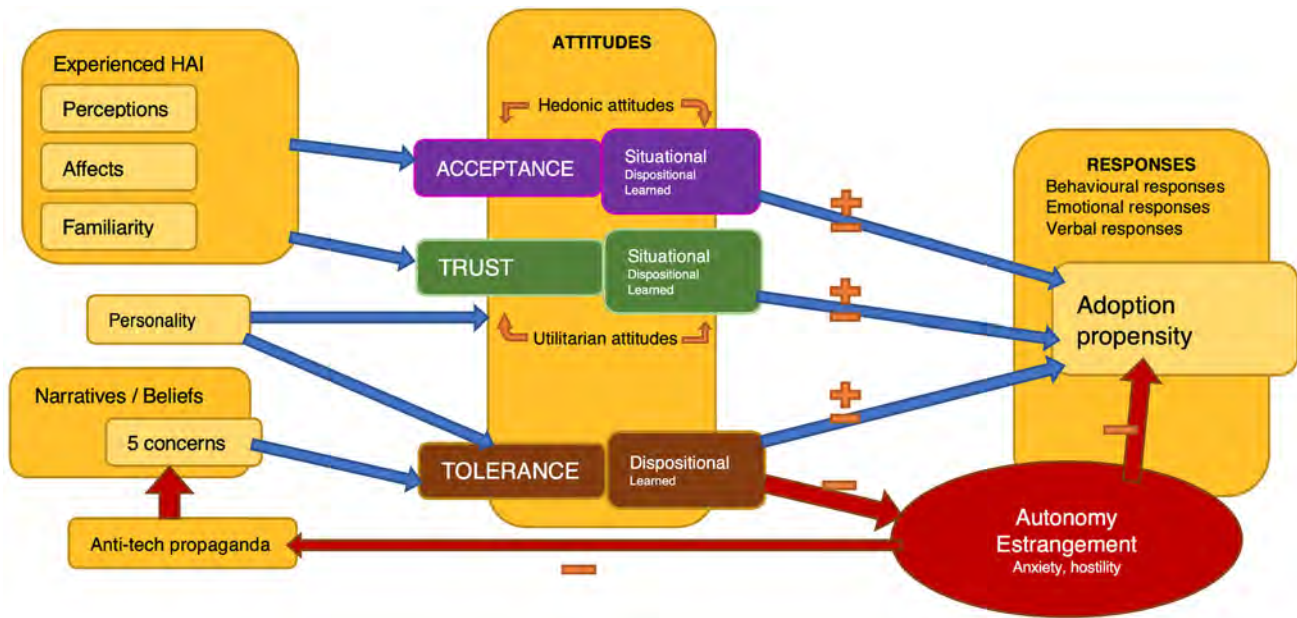
According to its own proponents, the notion of technology resistance requires a systematic revision, as it is currently unable to meaningfully account for the beliefs and worldviews that deeply motivate some people to resist technology adoption [15]. Comparatively, the notion of Tolerance we are proposing is more useful to HAI researchers because it can help them monitor subjects’ proneness to what we call “Autonomy Estrangement”, a state of mind characterized by uncanny feelings and disturbing (possibly obsessive or even paranoid) elaborations about autonomous technologies: While a high Tolerance for Agents reduces one’s chances of developing such a psychological state, an Intolerant attitude may open the door to Autonomy Estrangement inviting the most negative behaviours associated with it (e.g., robot avoidance, vandalism, sabotage). At the same time, we argue that Intolerance should not be confused with other negative attitudes toward Agents (e.g., lack of Trust and lack of Acceptance) that may induce analogous behaviours through different psychological mechanisms [16] (Table 1).

Like Acceptance and Trust, Tolerance (and, subsequently, Intolerance) is one of the attitudes toward Agents that HAI researchers study to facilitate the integration of autonomous technologies in human environments [16–18]. While Acceptance and Trust are well-researched metrics, Tolerance describes an aspect of HAI that has been investigated only scarcely so far and thus requires specific analyses: Acceptance and Trust reflect whether the user’s expectations, needs, and preferences were satisfied by their interaction with some Agents; Tolerance, on the contrary, reflects the categorical beliefs and general worldviews that a user holds about all Agents. These beliefs and worldviews are usually conveyed by narratives characterized by varying levels of complexity and vraisemblance. These narratives, in turn, may, and often do, circulate via popular accounts and influential stories, *independently of real interaction with actual Agents*; these beliefs and worldviews can be rather resilient to revision, especially when they are based on conspiracy narratives,

¹ We capitalize this word only when it indicates *autonomous artificial agents* like mobile robots or digital avatars, to distinguish them from any other kind of agent (e.g., human, animal, natural, chemical, mechanical yet non-autonomous, agents).

² As the definition of Agents used in this paper includes also humanoid and zoomorphic robots, any time we refer to HAI in this context we implicitly refer also to fields like human–robot interaction, human–machine interaction, and social robotics.

³ Kaczynski’s controversial manifesto [2] alleges that the progressive transfer of decisional processes to impersonal, de-humanizing systems and procedures, like those mechanically implemented by autonomous technologies, prevent individuals from realising and pursuing their fundamental needs and desires, including the need to freely exert agency in accord with their nature and to intentionally face the risks associated with their decisions [14].

Table 1 A revised model of Adoption Propensity for Agents

deeply held prejudices, and irrational or pseudo-rational stereotypes. Ultimately, Tolerance reflects the user's categorical beliefs about Agents in general (whether real or just imagined), not the quality of the user's experience of interaction with some particular existing Agents.

When we try to disentangle Tolerance from Acceptance and Trust, the main methodological challenge is that the former has always been embedded within the latter two: while existing models of Acceptance and Trust may at times capture some aspects of Tolerance, they never acknowledge the confines of Tolerance and its specific scope. The overlap between different attitudes is misleading because it conceals Tolerance's distinctive impact on HAI: The narratives that fuel Tolerance are largely independent of the kinds of direct experiences that motivate Acceptance and Trust. Therefore, the fact that Acceptance and Trust map some aspects of the behavioural expression of Intolerance (e.g., proneness to anxiety and hostility) suggests more a methodological confusion among the definitions of the existing constructs than the objective capability of the existing notions to effectively account for the phenomena described by Tolerance. As each of these three constructs offers a unique contribution to categorise the complex attitudes of humans toward Agents, our goal is to bring to light their specificity and complementarity.

As a first step toward this goal, Sect. 2 problematises the way Acceptance and Trust are prevalently used, arguing that these two notions are not properly distinguished in the literature and, moreover, they do not exhaustively map the attitudes that determine Adoption Propensity. In Sect. 3 we illustrate our concerns through a hypothetical scenario requiring that Tolerance is added to the picture to overcome the deficiencies of Acceptance and Trust. In Sect. 4, we articulate a

phenomenology of Tolerance as a form of anxious/hostile attitude towards Agents in general, motivated by narratives that have a normative valence, not experiential engagement.

Section 3 illustrates the deficiencies of Acceptance and Trust through a hypothetical scenario that can be made sense of only by adding Tolerance to the picture. In Sect. 4, we articulate a phenomenology of Tolerance as a form of anxious/hostile attitude towards Agents motivated by narratives that have a normative valence, not experiential engagement, and we fully analyse how such an attitude differs from Acceptance and Trust. In Sect. 5, we outline the principles according to which Tolerance can be recognized and measured, examining five defining concerns/worries to which Intolerant people are particularly prone. This section also introduces two seed questionnaires for measuring different, but closely related, metrics that are relevant to Adoption Propensity: the first questionnaire is specifically designed to assess the level of Tolerance by tracking the presence of said concerns/worries as part of the subject's beliefs about Agents; the second questionnaire, in turn, is designed to assess the level of Autonomy Estrangement, by tracking the subject's feelings and hope in relation to Agents.

In Sect. 6 we discuss the general socio-cultural significance of Tolerance in relation to Adoption Propensity. We explain why such a construct is neither intrinsically positive or negative but has nonetheless a contextual normative valence insofar as it helps us red-flag biased, irrational, and possibly fanatic attitudes. We discuss six hypotheses to be tested empirically, indicating possible trajectories for future research aiming at expanding our seminal investigation: the first hypotheses is the inverse correlation between high Tolerance and Autonomy Estrangement; the second,

that low Tolerance is exacerbated by the militant propaganda of anti-tech ideologies; the third, that Tolerance correlates to personality traits like neuroticism and closedness; the fourth, that Intolerance tends to be more covert and ambivalent than low Acceptance and Trust; the fifth, that Intolerance for Agents shares a common root with intolerance for humans and tends to resonate with it; the sixth that the best mitigation strategies involve early preventive interventions through education and information. Finally, in the conclusive section, we summarise the key points of our proposal, involving a model of Adoption Propensity informed by Tolerance and Autonomy Estrangement, and we highlight the benefits offered by such a model to the field of HAI.

To offer a conceptual model capable to combine Tolerance with the revised definitions of Acceptance and Trust, we will investigate the content of the beliefs and worldviews typically endorsed by the subjects more prone to Autonomy Estrangement. The seminal model presented in this paper is deliberately open-ended as its primary goal is to raise awareness about the important role of Tolerance in HAI, unearthing its psychological, cultural, and theoretical foundations: We aim to provide some general methodological guidelines for its assessment, not a fully-fledged operationalization of the construct. Due to the novelty of the proposed construct and the absence of relevant experimental data on Tolerance for Agents, our theoretical claims primarily build on phenomenological and genealogical analyses corroborated, when possible, by current experimental results in social robotics and social psychology (especially tolerance for humans).

2 Problematising the Attitudes Towards Agents

As the function of today's Agents is still primarily assistive, humans are required to complement or supervise their operations [19]. The successful deployment of Agents depends on their capability to satisfy the *users' expectations*, which can be surprisingly *heterogeneous* (defying common intuitions), *implicit* (unknown to the users themselves), and even *ambivalent* (conflicting with other expectations) [20]. Human attitudes toward Agents correlate with *perceptions* and *affects* evoked by the Agents' appearance [21], the user's *familiarity* with Agents (i.e., aptitude and previous exposure to similar technologies [22]) and *beliefs* about Agents (i.e., factual knowledge and more or less accurate judgments, including both judgments based on rational inferences and judgments based on irrational or pseudo-rational preconceptions [23]). These attitudes depend on various psychological, socio-cultural, and aptitudinal factors [24, 25] and are recognizable and measurable through self-reports (explicit method) or external observations (implicit method):

the former include the *verbal responses* generated before, during, or after interaction (i.e., casual utterances and explicit answers to questionnaires or semi-structured interviews); the latter include the *behavioural and emotional responses* produced during or after interaction with Agents (e.g., posture, gesture, gaze, facial expressions) [23].

Acceptance and *Trust* arguably are the two most studied attitudes in this field. HAI research scrutinizes both attitudes in the users, designing ad hoc assessment tools (like rating scales) to appropriately measure them [10]. Combining Acceptance and Trust systematically, though, proves difficult, as these two constructs were conceived independently in different contexts for different types of application.

2.1 Acceptance

The notion of Acceptance for Agents is prefigured by the older Technology Diffusion Theory [26], which explained the adoption of innovative technologies with a five-stage process that begins with the public "awareness" of technology's impact and terminates with the institutional "confirmation" of technology's adoption. In 1989, Davis introduced his Technology Acceptance Model (TAM) to explain how the *propensity to adopt* a certain technology (the "behavioural intention" that predicts the subsequent "use" of said technology) is determined by two factors: how that technology seems to improve task performance ("perceived usefulness") and how effortless its intervention appears ("perceived ease of use") [27]. TAM's second iteration enriched the model introducing additional "job-related variables" (job relevance, output quality, result demonstrability, user's experience) and "personal implications" (social norm, voluntariness of usage, image) [28]. Its following led to a "Unified Theory of Acceptance and Use of Technology" (UTAUT), which again distinguishes two families of factors ("performance expectancy" and "effort expectancy", corresponding to the previous "perceived usefulness" and "perceived ease of use", respectively) and introduces additional intervening factors like "social influence" and other "facilitating conditions" [29].

The earliest Acceptance models focussed prevalently on the societal/public dimension of technology adoption. In turn, its more recent iterations explore the personal and group level to enable a more detailed understanding of the user's experience. Also, recent TAMs tend to be tailored on digital and computer technologies, as they describe "the anchoring and adjustment of human decision-making processes" [30] by means of variables such as: "computer self-efficacy", "perception of external control", "computer anxiety", "computer playfulness", "perceived enjoyment", and "objective usability". Several variants of UTAUT have specifically explored the attitudes of users working within narrow technological domains, introducing more and more specific factors, like

those concerning personal interactions with social robots or industrial robots [31] (e.g., investigating Acceptance for a pet-like robot cat [29]).

UTAUT has further evolved under a constant pressure for diversification, branching into multiple narrowly specialized TAMs [32]. This trajectory solicits two remarks: the first is that the need to produce context-specific TAMs is in tension with the cross-disciplinary generalizability of said models, due to the inevitable trade-off between *distinctness* and *transferability* of the evaluations across different technological domains (i.e., an evaluative template tailored on a narrow domain becomes more granular when transferring to other domains); the second is that, despite the increasing diversity of TAMs, the polarization between *utilitarian factors* (objective benefits, usefulness, outcome) and *hedonic factors* (subjective experience, aesthetics, effort) remains constant throughout consecutive iterations.

2.2 Trust

As a HAI construct, Trust for Agents serves a purpose analogous to Acceptance's: identifying causal factors that motivate Adoption Propensity as a positive, welcoming attitude. Trust, though, had a different historical genesis: It was first introduced in social psychology to study the emergence of alliances and mutual interpersonal dependencies in human–human (not human–Agent) relationships [33, 34]. The game-theoretical concept of Trust applies to any cooperative scenario in which uncertain, high-stake decisions require confident predictions of the co-operators' true intentions and capabilities.

This applies to both human and robotic co-operators [35], which is why Trust could soon be incorporated into HAI research: Biros and colleagues define Trust as the attitude of “having confidence in the autonomous system to do the appropriate action” [36]; Lee and See as the expectation that an Agent will “help the user achieve their individual's goals in a situation characterized by uncertainty and vulnerability” [37]; Wagner as the “belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk” [38]; Oleson and colleagues as the user's reliance that the Agent will not undertake “actions prejudicial to the well-being” of the user [39]; finally, Hald and colleagues define Trust as the “the combination of feeling physically safe around [the robot] and being able to predict the robot's action in the context of a shared task” [40].

The concept of Trust for Agents remains, however, elusive due to (i) the inherent vagueness of analogies with human–human trust (i.e., Do machines have aims? Do they understand whom their actions will benefit?) and (ii) its dual nature (objective and subjective), which ambiguously refers

to an inclination of the trustor and a quality of the trustee at once.

Most definitions highlight the utilitarian component of Trust as they link this attitude with the belief that Agents will function properly, benefit the users, protect their interests, and fulfil their expectations (whether “robot's actions and behaviours correspond to human's interest” or not) while preventing major disadvantages (“avoiding that human has something to lose” in relying on the robot) [41]. However, a merely utilitarian notion of Trust is hardly useful if it reduces HAI to an instrumental relationship. The notion of Trust was not invoked when automated technologies were still conceived instrumentally, as passive tools confined within factory premises [42]. Trust became integral to the emerging paradigm of HAI only since the early 2000s, when service Agents introduced in the home and the office environment to assist people finally started to be seen as assistants and co-operators capable to act and make decisions independently [43].

Thus, while the Adoption Propensity of earlier *automated technological systems* depended exclusively on their more or less successful functioning (tracked by utilitarian variables like “reliability” or “efficacy”), to predict whether *autonomous technological systems* (Agents) will be adopted, we additionally need to know how Trustworthy they are: that is, we must assess whether the Agent both *is capable to help* the users and *authentically aims* to do it, as these two necessary conditions (respectively called “competency” and “integrity” [44]) are key to predict Trust among humans.

Even introducing this clarification, Trustworthiness remains an ambiguous concept. It tracks, at once: (I) a quality of the human–Agent relationship (which is why it presupposes some reciprocity) [45]; (II) an objective quality of the Agent; and (III) a subjective judgment by the user, a judgment that reflects their own *psychological profile* (i.e., personality traits, individual preferences, anthropomorphic tendencies), *background* (e.g., professional experience, familiarity with the technology), and *beliefs* (i.e., what the user thinks of and how they feel about the Agent). That is why Hancock and colleagues [46] distinguished between a “dispositional”, a “situational”, and a “learned” component of Trust, corresponding to whether the trusting attitude was determined by long-term (possibly permanent) inclinations and propensities, transient aspects of an occurring HAI experience, and knowledge about the Agent and their performance. In principle, a similar tripartition can be applied to Acceptance too.

Some authors believe that Trust, like Acceptance, depends on both the *utilitarian benefits* produced by Agents and the *more or less satisfactory experience* associated with using them (e.g., “[Trust] predicts not only the quality of HAI but also how willing people are to use social robots for certain tasks” [47]). Two remarks are necessary also in relation to Trust: first, the notion of Trust evolved to incorporate

an enriched understanding of the relational dimension of HAI; second, Trust presents an internal polarization between objective and subjective factors, which resembles Acceptance's internal dualism between utilitarian and hedonic factors.

2.3 Combining Acceptance and Trust

While the polarization embedded in both Acceptance and in Trust is not wrong in itself, we must note that too often it has fuelled a problematic confusion between these two constructs. For example, Gaudiello and colleagues assume that Trust strongly contributes to determine Acceptance when they claim that the users' "Trust in the robot [can be] considered as a main indicator of acceptance" [48]. Similarly, it is often assumed that Trust can be used to determine the overall Acceptance of a system, in a way that tends to implicitly conflate the two concepts [23, 49]. If Trust is often considered a component or constituent of Acceptance, Acceptance is also frequently considered just another component or constituent of Trust: far from being accidental, this confusion arises because originally Trust and Acceptance were each deemed sufficient to individually explain Adoption Propensity [50].

Unfortunately, this overlap precludes a clear appreciation of the proper scope of these two constructs and confounds their specific contribution to Adoption Propensity. The tendency to conflate them derives, in our opinion, from the fact that the typical model of Acceptance involves an internal dualism between utilitarian and hedonic factors, a dualism that is replicated in the typical model of Trust. Therefore, each attitude comprises the same constitutive components of the other. But, so long as each construct aspires to offer a complete synthesis of both constituents, it will be difficult to differentiate among them. To make Acceptance and Trust complementary, we must reconceptualize either as inherently incomplete and limited in scope. Several multi-dimensional models of Adoption Propensity have been proposed by the literature, introducing an assortment of assessment methods to measure both Acceptance and Trust as distinct but complementary attitudes [23, 51, 52]. The ambiguity (across models) deriving from multiple concurrent definitions of each construct is aggravated by the unintended conceptual overlap (within models) between the two constructs and the underspecified nature of their scope and function.

Shortly, the taxonomical problems are essentially three. First, in most joint accounts, both Trust and Acceptance share a significant number of communal dimensions/indicators [53]. Second, one of the two attitudes is often taken as a dimension or aspect of the other [9, 48, 54]. And third, to add further confusion, *both* Acceptance *and* Trust are typically defined by technology adoption models in terms of the user's "intention", "willingness", or "propensity" to use the

Agents or interact with [23], which in turn reflects the "actual use" of Agents or at least their "perceived usefulness" [27, 55–57]. These ambiguities motivate at least two concerns regarding construct validity: first, how to measure the two constructs with rigorous and specific methods, given that their constitutive components are so similar? Secondly, how to distinguish in which semantic context Acceptance is more relevant than Trust?

To define their distinctive assumptions and methodological ramifications, construct validity of both Trust and Acceptance must be established conceptually, before being tested empirically. First, we must distinguish their respective scopes: Both Acceptance and Trust reflect certain expectations about an Agent, but our previous analyses indicate that Acceptance prevalently focuses on expectations concerning *the interaction* with the user (i.e., intended as a *mode* of using the Agent) while Trust primarily focuses on expectations concerning *the achievement* of a practical goal (the *end* for which the Agent is used). Thus, we suggest that Acceptance is essentially meant to express value judgments about the Agent's relationship with the users (i.e., the *role* played by the Agent as a co-worker, team member, or interaction partner), while Trust is essentially meant to express judgments about the Agent's performance (i.e., the *output* of their autonomous work). It is possible, of course, that both judgments are similarly influenced by identical or very similar sets of empirical circumstances, but there is no a priori reason to assume that these variables systematically co-vary.

We propose a refined model of Acceptance and Trust that values their conceptual independency by minimizing the number of traits and communal dimensions that they share. In our model, Acceptance describes the user's experience of engagement with an Agent and the willingness to recognize said Agent as a legitimate interaction partner or co-worker [27, 58–60]. Trust, in turn, tracks how the user acknowledges the Agent's capability to reliably operate toward the achievement of a certain goal and the user's willingness to assign tasks and delegate responsibilities to the Agent [42, 43, 61–64]. By highlighting their specific relevance to HAI more than their similarities and overlaps, our model emphasizes not only the distinctiveness, but also the complementarity of Acceptance and Trust: while Acceptance serves to mapping Adoption Propensity in relation to *the quality, robustness, and easiness of the interaction/engagement* with the user, Trust serves to mapping Adoption Propensity in relation to *the efficacy, efficiency, reliability, and safety of the Agent* in performing their tasks.

2.4 Tolerance

Our effort to operationalise Trust and Acceptance via a partial reconfiguration of their existing definitions ultimately

aims to make space for Tolerance as a third crucial attitude toward Agents. In fact, the goal of this paper is not to provide a revised account of Acceptance and Trust so much as introducing Tolerance as their indispensable complement. Acceptance and Trust are unable alone to exhaust the complexity of the attitudes toward Agents: They remain inevitably blind to some specific kinds of anxiety and hostility that only the notion of Tolerance can effectively account for. As argued in this paper, Tolerance for Agents is essentially independent of Trust and Acceptance, although the expressions of these attitudes may partly and occasionally align: an increment or decrement in the Trust and/or Acceptance score can situationally, not universally, correspond to increments or decrements in the Tolerance score, and vice versa. Due to its irreducibility to other constructs, we argue that Tolerance is no less fundamental than Trust and Acceptance in any systematic assessment of the users' expectations toward autonomous technologies.

3 Well-Accepted and Highly Trusted Agents may be Poorly Tolerated

To illustrate the tripartition described in Sect. 2, consider the group dynamics unfolding in the following hypothetical, yet plausible, scenario [65].⁴ Let us imagine that a logistic bot is introduced in an airbase to release the ground crew members from the time-consuming burden of transporting bulky components from one side of a very long hangar to the other. After a few weeks of apparently flawless implementation of the new system, the crew supervisors report that the hangar's personnel have almost entirely stopped using the bot and, on one occasion, some of them have tried to sabotage it.

Some HRI experts are recruited to run semi-structured interviews with the crew members to investigate their negative attitude toward the new technology. Their investigation reveals that, surprisingly, the crew members found the bot quite useful and reasonably reliable: all of them admit liking its friendly interface and its interactive functions; the majority of them expresses sympathy for the bot, while a smaller group is rather indifferent to it. So, why are they ultimately reluctant to use it? Further investigations indicate that the crew members, *while not disliking or distrusting the bot per se*, believe that the permanent adoption of bots may deteriorate the professional and personal relationships between two different teams of crewmembers deployed at the opposite sides of the hangar, reducing their opportunities to interact during work hours. This worry alters the perception of their role within the airbase, rousing latent tensions between the two groups.

Moreover, although there was no plan to replace the workers, the crew members progressively grew concerned about the possibility of becoming redundant if additional bots were deployed in the hangar. Last but not least, some of the managers declare that assigning logistic tasks to a bot is morally wrong and thus should be avoided because that responsibility should be given only to humans.

The airbase scenario reveals something significant: The hangar crew had a favourable impression about the bot when it was first introduced. This initial response is perfectly accounted for in terms of Acceptance and Trust. What the crew judged positively was their interaction with the bot and the bot's ability to assist them in their daily duties. However, it soon became evident that their overall propensity to use the bot was not very high: In addition to their positive impressions, the staff members were simultaneously holding a very negative judgment about the bot, which eventually revealed their deeply anxious and hostile attitude toward the machine. Importantly, this second response overridden the first, not by *replacing* the earlier positive opinions about the bot's role and performance, but by *adding* to them a distinct negative judgment: the users never changed their beliefs that the bot was (1) nice to interact with and (2) useful to their work; rather, they expressed a third belief that, despite the positivity of the other two, ended up deteriorating their overall attitude. The positive opinions were *surpassed*, not *suppressed*, by the negative one. The negative opinion emerged from the realisation that, although the logbot never directly affected the social relations among the workers, such relations can be interrupted or displaced by bots, alienating the crew members from their traditional roles and making them anxious about the stability of their jobs. It seems unlikely that this negative attitude formed suddenly: it is plausible that it was already prefigured by some latent beliefs, inactive in the back of the users' mind, so to speak, silently awaiting to be triggered by the relevant circumstances.

Now, *perceived social displacement* and *unemployment anxiety* are phenomena broadly linked to negative attitudes toward robots [66, 67]. After all, one might be inclined to believe that these phenomena are best tracked in terms of Acceptance or Trust. If this hypothesis is correct, then the negative attitudes of the users can still be interpreted as long-term transformations occurring within the space of Acceptance and Trust, without involving a third kind of attitude. At a first glance, this explanation seems parsimonious because it prevents a seemingly unjustified proliferation of constructs. However, such an explanation is problematic because the general attitude toward the bot *radically changed while leaving completely unmodified the attitudes that had already been assessed*: Acceptance and Trust did not change at all even after the general attitude had switched from positive to negative: In other words, the users never felt any urge

⁴ This hypothetical scenario is presented as a way of a thought experiment to illustrate the logical ramifications of a conceptual distinction.

to revise their judgments concerning the quality of the interaction with the bot or the usefulness of its work.

The judgments concerning the interaction with the bot and its performance had already formed, they were consistently positive, and had never been challenged by the negative beliefs expressed later by the users: The crew members frankly admitted that the bot had never stopped to efficiently, efficaciously, and reliably execute the tasks assigned to it, even when unsupervised; the bot promptly served their goals and anticipated their needs, providing helpful information and making appropriate decisions; the bot still looked friendly and behaved in a pleasant, agreeable manner; it quickly responded to human commands no less than before; interaction with it has never ceased to be smooth and enjoyable; moreover, after two months of use, most of the crew members seemed still moderately inclined, like during the first week, to treat the bot as a teammate or companion (although—and this might be crucial—they did not necessarily treat it as the kind of teammate/companion they prefer to work or spend time with).

That is why the concurrent negative judgments about the bot can hardly be explained simply relying on the classic TAM (which tracks responses concerning the *perceived usefulness* and *ease of use* of a specific Agent [27, 54, 58]). A correlation with the personality traits “openness” and “neuroticism” seems plausible and should be investigated [68], but the categorical convictions examined in this context seem only loosely linked to personality-related variables like “affinity for technology interaction” or “technology commitment” (which again can only track inclinations formed through experiences of personal interaction [69]).

We must conclude that the new negative attitude falls outside the scope of widely used constructs like Acceptance and Trust. The bot’s final rejection was not caused by the factors that typically determine a low level of Trust, as *the users rejected the Agent even though (and even more so because) it operated effectively and reliably*: paradoxically, it was exactly when the Agent operated competently and made appropriate decisions like a human that certain users felt overwhelmed or even aggressed. We infer that the users started to ruminate the idea that only humans deserve to make decisions and operate in such a role. Rejection was not caused by a low level of Acceptance, as the quality of the interaction with the bot was never indicated as the source of the problem: thus, we infer that, although the Agent was perfectly integrated within the existing networks of human practices and did not directly interfere with the personal, professional, or social organisation of the users, *the agent’s technological, artificial nature was still identified by the crew as a potential threat or obstacle*.

Our interpretation is that the crewmembers’ refused to use the bot when their pre-existing, deeply held negative beliefs finally manifested themselves, making the crewmembers’

low Tolerance discernible. This dynamic, however, did not correspond to a concurrent decline of Acceptance and Trust: The scenario is easily explained assuming that the users’ low Tolerance became explicit—without affecting their Acceptance and Trust—*when the situation activated their deeply held negative beliefs about the bot*, triggering their latent fears concerning the drawbacks of allowing the bot in the warehouse. A similar interpretation can be extended to any work or home environment in which assistive robots (e.g., artificial butlers or nannies) are introduced. In such scenarios, humans may ultimately refuse to use Agents, although no particular Agent ever did anything wrong or unpleasant.

We need to refine the received theoretical framework if we want to explain the origin and the content of these negative judgments: thus, we introduce the dyad “Tolerance/Intolerance” to unearth the narratives capable to increase or undermine Adoption Propensity by conveying positive and negative beliefs about Agents of a categorical, normative nature. Tolerance is closely linked with the concept of *Autonomy Estrangement*, that we introduce here to better characterise the alienating condition characterized by pervasive and oppressive feelings of hostility and anxiety associated with Agents in general. We need the concept of Tolerance to capture the idea that preventing Autonomy Estrangement and building a resilience to it requires limiting the impact of negative narratives about Agents.

4 Ambivalent Attitudes Motivated by Conflicting Experiences and Narratives

To establish construct validity, the notions mentioned so far need being clearly distinguished. Tolerance is conceptually independent of both Trust and Acceptance in the sense that it does not systematically co-vary with them: A perfectly trusted and accepted entity is not necessarily well-tolerated, and, in turn, a poorly trusted and accepted entity can be perfectly tolerated. Analogies with human–human interactions in the workplace are both methodologically justified and useful, in this context, due to the deep similarities between human–human and human–Agent attitudes [70]. Think, for example, of when poorly accepted and untrustworthy humans happen to be well tolerated: even an employee disliked by his/her colleagues due to his/her obnoxious personality and questionable work ethics can have a brilliant career if the management is not keen on hiring more qualified personnel, unwilling to question the dysfunctional and self-exculpatory work culture, or attracted by the ‘Brilliant Jerk’ type [71].

A similar constellation can arise when humans deal with non-autonomous technologies. Think of a cohesive group of users that, instead of adopting state-of-the-art technologies, intentionally keep relying on outdated, if not unreliable, systems. In this *high Tolerance scenario*, the fact that the

obsolete systems are untrustworthy and unable to assure smooth and enjoyable interactions with the users does not necessarily prompt the hostility of the users, who may be inclined to keep their low productivity and poorly innovative profile. The motivations behind this *innovation resistance* can be extremely complex [72, 73]: often they have to do with a culture of leniency or simple reluctance to leave the familiar comfort-zone. We mention their impact on groupwork to exemplify how a highly tolerant attitude may be prompted by motivational trajectories that are entirely orthogonal to the Agent's inherent trustworthiness and acceptability.

Having dissociated Acceptance, Trust and Tolerance puts us in a better position to account for the user's mixed responses about the Agent, the causes of which might be unobvious to the HAI researchers and opaque to the users' themselves. Attitudes towards robots are often ambivalent: i.e., the same user can have multiple conflicted attitudes toward the same robots [74, 75]. This is not surprising, considering the stratifications of implicit beliefs and dispositions and the multifaceted nature of social dynamics [76]: First, not all users are perfectly rational decision-makers (they may be biased against or in favour of Agents); Second, their professional code of behaviour may conflict with their personal values (the user can share their employer's technology-friendly views, while thinking deep-down that humans should always be prioritized); Third, users may prioritize their personal success over their team's; Fourth, their first-hand experience may contradict their deeply held beliefs (including, but not limited to, prejudices and stereotypes), which they might be unwilling to revise despite contrary evidence [74].

The fourth point needs clarification. Various empirical studies confirm that, as efficaciously portrayed by writer Robert M. Pirsig in his popularly acclaimed novel [77], some individuals experience a mix of discomfort and revulsion when they have to interact with mechanical devices, a phenomenon that sociologists interpret as *technology resistance* [15]. It is well-known that users and bystanders often experience a sense of eeriness and displacement when directly dealing with Agents [78, 79]. Studies on the *uncanny valley* effect indicate that tendencies to repulsion and resistance are even more acute in the case of hyper-realistic autonomous devices with android or zoomorphic features [80]. Crucially, while this kind of negative response can have a dramatic impact on HAI, *it is not an indicator of scarce Tolerance for Agents*. The logistic bot scenario illustrates a case of low Tolerance (that is: "Intolerance") because it involves some covert, unobvious anxiety and hostility. Such a negative attitude does not arise from the unsuccessful performance of a certain Agent in a specific task or from their uncanny appearance; rather, it arises from deeply held general worries concerning how, in the long run, Agents may collectively transform the human civilisation.

The logistic bot scenario supports some inferences about Intolerance. First, Intolerance is an attitude typically associated with *offline* (detached and mediated) rational or pseudo-rational evaluations and normative judgments about all Agents, not from the *online* (directly engaged) experiential encounter with a certain Agent. Secondly, Tolerance can be, but—importantly—*does not need to be* affected by actual interaction experience, while Trust and Acceptance are always, necessarily, primarily motivated by perceptual and interactive experiences. The situational component is crucial and essential for Trust and Acceptance, but only secondary or irrelevant to Tolerance. Thirdly, the distinction between Tolerance, on the one hand, and Acceptance/Trust, on the other, reflects a dissociation between the *experiential level* (the direct perception of and immediate interaction with a particular Agent) and the *narrative level* (the general beliefs about the Agents' long-term impact on the work environment and its broader implications).

Negative beliefs and deeply entrenched convictions about Agents, possibly accompanied by negative feelings within the anxiety/hostility spectrum (e.g., frustration, envy, repulsion, etc.) are both a symptom and a cause of the fast erosion of Tolerance. These beliefs usually spring from and consolidate because of compelling narratives (popular stories, influential reports, and spreading opinions such as "robots are coming for our jobs" and "AI will make the most important decisions for us"), independently, or even in spite, of one's direct interactions with Agents. That is why Tolerant or Intolerant attitudes are more unlikely to fade away despite what the Agent does, says, or looks like, compared to (un)Accepting and (un)Trusting attitudes, which are more situational and context-dependent. The mismatch between these two distinct, possibly conflicting levels of evaluation (interactive experience vs contentful narrative; context-dependent vs context-independent; dispositional and categorial vs situational and particular, etc.) may conduce to more or less severe ambivalences and even result in cognitive dissonance.

To illuminate the crucial, and dramatically overlooked, role of Tolerance for Agents in HAI, we need to develop scales capable to rigorously capture the specificity of Tolerance. Building on our previous inferences, we can identify at least six elements of specificity, to be assessed with appropriate queries.

The first difference is that Tolerant/Intolerant attitudes are assessed evaluating the *content of the beliefs* and worldviews that subjects hold about Agents. That is, Tolerance reflects the compelling force of the *narratives* that convey this content, while Acceptance and Trust are assessed evaluating the *interactive experiences* that humans may or may not have had with certain Agents (based on the frequency and quality of direct engagement).

Second, Tolerance is assessed asking subjects how they agree with *categorical judgments about Agents in general*, while Trust and Acceptance are measured asking subjects to evaluate their *particular instances of interaction with this or that Agent*.

Third, Tolerance is assessed investigating subject's *normative* beliefs, concerning—for example—whether Agents should or should not be designed in a certain way and whether they do or do not deserve to operate in certain roles, based on whether this is desirable, fair, and appropriate; conversely, Acceptance and Trust are assessed asking questions of a *descriptive* nature, concerning Agents' appearance and how they factually operate in certain roles.

Fourth, Tolerance is a *dispositional attitude* to be appreciated prevalently in the *context-independent* dimension, often acquired through learning, with little or no situational component; Acceptance and Trust, on the contrary, are primarily determined by the *situational context* in which the interaction with Agents occurs, and only secondarily by the possible dispositional and learned components.

Fifth, Tolerance is an attitude primarily appreciable in the *offline dimension*, because it is primarily and necessarily motivated by beliefs and judgments formed outside of or anyway *independently of HAI*—which is why Tolerance is assessed primarily in a long-term time scale or without precise temporal references; Acceptance and Trust, on the contrary, are primarily appreciable in the *online dimension*—which is why they are assessed in a *short-term time scale*, considering temporally specific events.

Sixth, due to its narrative and thetic nature, Tolerance primarily expresses a kind of contentful *knowledge-that*, concerning positive or negative facts about Agents; conversely, due to their episodic and experiential nature, Acceptance and Trust, primarily attest a *direct performative familiarity* developed through satisfactory or unsatisfactory instances of practical interaction with Agents.

In short, Tolerance is prototypically assessed by tracking the subject's normative, categorical, acontextual narratives and beliefs that reflect their factual knowledge of Agents in a detached and atemporal dimension; conversely, Acceptance of Trust are prototypically assessed tracking the subject's descriptive accounts of contingent, practical instances that reflect the subject's interactive competence in a situationally engaged and temporally specified experiential dimension. All these prototypical differences are reflected by survey questions to be asked for assessing the user's attitudes toward Agents. Table 2 shows that the questions used to assess Tolerance are substantially different from the questions used to assess Acceptance and Trust. These observations will be useful to draft a rudimentary protocol for the initial assessment of Tolerance.

5 The Assessment of Tolerance: Negative Narratives, Beliefs, and Concerns

While Acceptance and Trust are experientially justified, Tolerance is primarily motivated by the conviction that the nature of autonomous technologies is not only innovative and disruptive but also inherently estranging and dangerous. But what are exactly the negative beliefs about autonomous technologies that reveal a high or low Tolerance for Agents? To answer this question, we must investigate the narratives that motivate Tolerance or Intolerance.

These narratives are schematically mapped by existing technology resistance models [15], which are helpful at least to distinguish two macro-groups of concerns associated with autonomy: (A) *concerns over disruptive societal changes* and (B) *concerns over Agents' dominance*. However, this level of granularity is not sufficient. These general concerns correspond to two narrative frameworks under which we have grouped five specific worries about autonomous technologies, each of them testified by at least five negative beliefs about Agents. The list presented in Table 3 may not (and it is not meant to) be exhaustive, but it comprises a sufficient number of concerns to assess the subject's degree of Tolerance for Agents.

The five abovementioned concerns can be assessed by survey questions explicitly designed to measure whether and how intensely the subject holds the corresponding beliefs (Table 4). The final evaluation is based on a total score calculated by adding up the values of all the individual answers, which range from 0 to 4 to reflect agreement level).

As per our previous analyses, the five concerns examined here do not arise from *direct interaction* (i.e., a situated form of online engagement) with Agents but from a *detached judgment* (including beliefs, narratives, or theories) that the user formed during an undefined period of time and that does not necessarily coincide with the time of real interaction.

All these concerns essentially reflect the same fundamental worry that technological systems could, sooner or later, *govern humanity*, estranging human lives from their natural purposes, responsibilities, challenges, and risks to the point of making human existence meaningless and unrecognizable. This kind of apocalyptic worry, epitomized by Kaczynski's manifesto [2], informs in various manners the relationship with all technologies, but it is particularly forceful in the field of HAI, where it specifically characterizes Intolerance for Agents. That is because, much more than any other kind of technological systems, Agents can be designed to act in an advisory, supervisory, or executive role: Unlike other automated or digital technologies, autonomous systems can explicitly or implicitly change the goals and priorities of humans, thus they also have the power to estrange humans from their pre-existing goals and priorities, luring them with

Table 2 Specific scope of Tolerance and its key differences with other HAI attitudes

Characteristic	Tolerance	Acceptance & Trust	Form of the question (Tolerance vs Trust & Acceptance)
<i>Modality</i>	Narrative-based—it is measured evaluating the content of the beliefs and worldviews that subjects have about Agents in general	Experience-based—it is measured asking questions about the quality of the interactive experience that users had with certain Agents	“Do you believe that Agents are X?” “Do you agree with narrative X about Agents?” vs “How X was your experience of interaction with Agents?”
<i>Extent</i>	General—it is about all Agents in general, as a class or category. It is primarily about the essence or nature of Agents, regardless of interaction experience	Exemplar—it is about individual Agents (or a definite group of them). It is about how some Agents operated/functioned during specific episodes of interaction	“Is property X necessarily found in all Agents?” vs “Did this/these agent(s) display property X?”
<i>Function</i>	Normative—Asks whether Agents should or should not have certain characteristics and whether they should or should not operate in certain roles, based on whether it is desirable/good	Descriptive—Asks what Agents appear like and how they operate in certain roles, based on what they actually do/how they behave	“Should Agents be allowed to operate in role X?” vs “Do Agents effectively operate in role X?”
<i>Context</i>	Context independent—it is prevalently dispositional and learned; it has little or no situational component	Context-dependent—it is prevalently situational, but it can have a learned and dispositional component too	“Agents are X: how do you agree?” vs “How X was Agent Y in situation Z?”
<i>Temporal framework</i>	Offline: primarily underpinned by a temporal or long-term temporal processing occurring outside or independently of HAI events	Online: primarily underpinned by cognitive processes confined within the short timeframe in which HAI events occur	“Agents make me always X: how do you agree?” vs “Did Agent Y make you X at time Z?”
<i>Epistemic validity</i>	Knowledge-that—about the truth and validity of the user’s thoughts and statements about Agents	Knowledge-how—about the user’s readiness and capability to successively interact with Agents	“Which facts prove that Agents can do X in a Y way?” vs “Can you successfully do X in a Y way with Agents?”

fictitious and counterfeited imitations of authentic aspirations [13, 81].

Some anti-tech ideologies [2] prophesize a dystopic scenario: Because of autonomous technologies, our ways of living and working risk becoming more and more dependent on impersonal systems; however, such systems can only mimic or simulate, not share or understand, human goals

and aspirations; thus, the tendency to offload personal decisions on Agents progressively deprives humans of the related responsibility, privileges, and dignity [11, 13, 82]. When chronically exacerbated, this narrative motivates the fear that an alienated fate awaits those humans who have accepted to be governed by autonomous technologies: to become empty, purposeless shells who live risk-free but ultimately inauthentic and meaningless lives [81, 83].

Table 3 Concerns about autonomous technologies evidencing Intolerance for Agents

(A) <i>Concerns over disruptive societal changes</i> Even if Agents don't directly interfere with human decisions, their sheer existence can indirectly subvert the social reality of humans, their political practices (the way humans exert/recognize power, and make decisions collectively), their relationship with the environment (designed for interaction with Agent or only to host humans), and possibly their ethos and values (their priorities and their perception of what is good, right, and fair)	(A1) <i>Concern that Agents may alter the users' practices and habits</i> To maintain and operate the Agent, the human users are forced to modify their lifestyle, modus operandi, customs, traditions to the point that they feel they are serving the machine, rather than reverse. For example, public, private, and intimate spaces might be redefined by the presence of Agents [97]; also, human capacity for love, friendship, and altruism could be affected by the systematic reliance on assistive technologies [98, 102]. Note that this may happen even if the Agent is perfectly recognised as a legitimate and reliable group member, as such recognition does not prevent the Agent from disrupting the patterns of everyday activities (e.g., introducing taxing logistical and training requirements)
(B) <i>Concerns over Agents' dominance</i> Agents may threaten the primacy and the uniqueness of the humans, taking on expert positions, decisional roles, and metaphysical privileges that only humans were typically entitled to have	(A2) <i>Concern that Agents may alter the socio-political relations among users</i> As a variant of AI, the introduction of Agents may affect the existing socio-political equilibriums, forcing the users to question the existing hierarchies and institutions and challenge the norms and values that regulate labour and the productive systems [3, 94] (B1) <i>Concern that Agents may replace humans in their jobs and key functions</i> Agents might make humans redundant or obsolete, as the former effectively take on specific roles, tasks, or function that in the past only the latter could have. This role can be professional (an occupation in the workplace), social/affective (a status of confidant, friend, or caregiver in a private environment), political (a formal or informal leadership position within the group) or purely symbolic (the team's mascot). The fact that the Agent operates in that role/position flawlessly, diligently and potentially even more effectively than its human colleagues does not reduce but increases human hostility and envy, boosting their sense of perceived professional displacement, especially in contexts in which the risk of technological unemployment and the competition with AIs are strongly perceived [59] (B2) <i>Concern that Agents may undermine the users' authority and decisional power</i> Empirical data suggest that humans don't always cope well with robotic teammates in authority roles [100] and often experience unease when required to obey their direct orders [101]. Even if Agents don't take away their jobs, humans may still feel that their authoritative status is undermined, their competences depreciated, their socio-political standing diminished when Agents become capable to make decisions and dispense timely advice as effectively, or even better, than them. The fact that the Agent's knowledge seems comprehensive and accurate beyond human abilities may increase the sense of subjection and inadequacy experienced by the users (B3) <i>Concern that Agents may question human exceptionalism, challenging their unique status/entitlements</i> Even when Agents don't take on the jobs of humans or diminish their authority, humans may still feel that their unique status is being threatened by the very existence of Agents, who seem having at least some of the special characteristics that only humans used to (think to) have, including, importantly, the status of moral agents and moral patients. This collides with the standard anthropocentric assumption that humans deserve special consideration based on their metaphysical status and their moral exceptionalism [103]

Table 4 Designing a scale for Tolerance—sample questionnaire

Beliefs groups corresponding to Concerns that predict low Tolerance	
Group A1 Concern that Agents may alter human customs and practices	Questionnaire to measure Tolerance to Agents—Beliefs underpinning the corresponding Fear/Concern groups 5 possible answers (score 0–4): “strongly disagree” (0), “somewhat disagree” (1), “neither disagree nor agree” (2), “somewhat agree” (3), and “strongly agree” (score 4) (A1.1) If they were massively adopted in our homes, Agents would severely disrupt the human’s customary ways of living and staying together (A1.2) If they were massively adopted in our workplaces, Agents would irreversibly transform the human’s established ways of operating in the professional environment (A1.3) If they were massively adopted in our communities, Agents would profoundly modify the human’s traditional cultures, the meaning of their legacy, the significance of their heritage (A1.4) If automotive vehicles were entirely replaced by autonomous cars, some people may forget how to drive and how beautiful driving used to be (A1.5) If creative media were entirely replicated by generative AIs, people may lose their artistic skills and their capability to appreciate real art (A2.1) If they were allowed to work, Agents would delegitimize the norms and conventions that regulate the relationships among socio-economical groups/classes (A2.2) If they were allowed to make important decisions, Agents would destabilize the existing political order and the hierarchies that govern the human societies and institutions (A2.3) If they were allowed to live with us, Agents would erode the personal bonds among friends or relatives (A2.4) If we keep developing robot companions and care-givers, some users may start forgetting the value of human relationships (e.g. friendship and family) (A2.5) If robot managers were used in the higher administration, many of us may end up taking orders from them
	(B1.1) Agents should not be allowed to take on certain jobs and professional roles that traditionally belong to humans (B1.2) Agents should not be allowed to have certain managerial responsibilities and key leadership functions that customarily only humans can have (B1.3) Agents should not be allowed to take the kind of high-stakes, high-risk, like-or-death decisions that normally only humans can take (B1.4) If robots were allowed to make life-or-death decisions as surgeons or soldiers, many people may forget the meaning and the dignity of the human life (B1.5) If robots were allowed to send people to jail as judges or lawenforcers, many communities may forget the meaning of right
Group A2 Concern that Agents may alter the socio-political relations among humans	
Group B1 Concern that Agents may replace humans in their jobs and key functions	

Table 4 (continued)**Group B2**

Concern that Agents may undermine the authority and decisional power of humans

- (B2.1) In any decisional process involving both humans and Agents, the judgment and preferences of humans deserves a greater weight
- (B2.2) In any decisional process involving both humans and Agents, the real executive power (as opposed to merely consultative power) should stay with humans
- (B2.3) In any decisional process involving both humans and Agents, the authority attributed to human knowledge and expertise should be always greater than the authority attributed to Agents'
- (B2.4) Our professional credentials might lose value if robots started doing our jobs
- (B2.5) If a robot started to make decisions on our behalf, many people may stop respecting our institutions
- (B3.1) Regardless of how sophisticated they are, Agents cannot be given the civil value, the rights, and the entitlements normally attributed to human citizens
- (B3.2) Regardless of how intelligent they are, Agents could never acquire the personal status, the moral consideration, and the special dignity of human persons
- (B3.3) Regardless of how much Agents evolve, they could never acquire the spiritual qualities, the intellectual faculties, and the conscious self-awareness that define humans
- (B3.4) A robot will never be able to truly understand what is important and valuable in life
- (B3.5) A robot should always be someone's property and should never be allowed to own properties

Group B3

Concern that Agents may question the exceptionalism of humans, challenging their unique status/entitlements

Tolerance is measured on a 5-degree scale

- (i) Very Tolerant
- (ii) Somewhat Tolerant
- (iii) Neither Tolerant nor Intolerant
- (iv) Somewhat Intolerant
- (v) Very Intolerant

We propose two alternative methods to measure Tolerance on this scale

*Method I: sum the scores (from 0 to 4) of each of the 25 answers**Calculate the total score (from 00 to 100)*

00 to 20 Very tolerant

21 to 40 Somewhat tolerant

41 to 60 Neither tolerant nor intolerant

61 to 80 Somewhat intolerant

81 to 100 Very intolerant

Method II: sum the scores (from 0 to 4) of the 5 answers comprised in each of the 5 groups. Calculate how many groups score 14 or more (out of 20)

0 groups: Very Tolerant

1 group: Somewhat Tolerant

2–3 groups: Neither Tolerant nor Intolerant

4 groups: Somewhat Intolerant

5 groups: Very Intolerant

Humans experience an alienating psychological state when they are overwhelmed by the prophetic belief that autonomous technologies are making or could make human life meaningless and empty [13, 84]. We call this state “Autonomy Estrangement”, which is characterised by anxious (disoriented, distressed) and hostile (frustrated, aggressive) feelings and wishes in relation to Agents and the dystopic narratives that involve Agents (listed in Table 5). Autonomy Estrangement involves various manifestations of *robot-phobia* [85] and *techno-phobia* [86], with varying degrees of severity ranging from mild scepticism toward Agents to a major paranoid obsession. Whether they are justified or not, Autonomy Estrangement correlate to a radical rejection of all Agents, which may or may not be expressed through distressed emotive responses (anxiety, disorientation, hostility, and frustration). Autonomy Estrangement strongly undermines Adoption Propensity because it fuels hostile, rejective attitude toward Agents motivated by negative judgments and narratives. Discussing whether the content of the underlying beliefs and worries (which are mapped by the Tolerance scale) is reasonably justified or not is beyond the scope of this paper, but it is obvious that the condition of distress and aggressiveness caused by Autonomy Estrangement is inherently undesirable and draws a problematic trajectory for HAI.

While closely related, Intolerance for Agents and Autonomy Estrangement are not conceptually the same: the former is a general attitude conditioned by beliefs, the latter a psychological state expressing an emotional and wishful response to those beliefs. Intolerant beliefs may contribute to cause estranged emotional states; in turn, estranged emotional states per se do not generate new narratives. They can, however, contribute to exacerbate beliefs conveyed by existing narratives. The relationship between Intolerance and Autonomy Estrangement is somewhat analogous to the relationship between a weak immune system and a viral infection. It is possible to inductively infer how the former operates based on what the latter is doing, but not the reverse. A strong immune system (high Tolerance) helps prevent or mitigate the infection, but it never provides a 100% immunity. Also, we must be aware that other conditions (in our metaphor: low Trust and low Acceptance) may at times produce symptoms like those associated with the viral infection, which is why it is important to distinguish between different aetiologies. As we will see in the next section, the symptoms of anxiety and hostility are well evident in the radical ideologies that invoke rebellion against technology, including those endorsed by anti-tech and neo-luddite movements [2, 87].

6 Scope for Future Work: the Irrational Roots of Intolerance

Tolerance for Agents is a value-neutral construct, thus even very low or very high Tolerance values should not be attributed an intrinsically positive or negative valence: they are not meant to attract contempt or approval or be considered inherently deviant or pathological. Apart from patently aberrant cases, HAI researchers must assume methodological neutrality toward the beliefs and narratives that fuel Tolerant or Intolerant attitudes toward Agents, such as “Agents will make our world a happier place” or “AI is likely to kill all of us”. The Tolerance score provides only an indication of whether, and how robustly, such beliefs and narratives were endorsed by the subject. Tolerance is inherently limited in scope because it cannot tell us whether the beliefs endorsed by the subject have rational, justified bases (empirical evidence or valid inferences) or rather irrational, faulty ones (biases, invalid reasoning). If we lived in a Terminator-like apocalyptic scenario (machine Armageddon and killer-robots taking over human civilization), the surviving humans would probably have a very low Tolerance for Agents, and that would have to be considered perfectly justified.

However, we do not live in the Terminator’s world, which is why only a small minority of individuals fear machines as if they were going to take over at any moment. That is why, after establishing the in-principle neutrality of the concept of Tolerance, we must also remark that one of the most compelling reasons to introduce Tolerance in HAI research is that, in our world, extremely Tolerant or Intolerant responses *can be symptoms of biased, irrational, and possibly fanatic attitudes*. This suspicion is supported by independently collected data suggesting that some widespread categorical beliefs and normative judgments concerning Agents are based on preconceptions deriving from popular narratives (e.g., sci-fi representations) [88] that are supported neither by scientific knowledge nor by practical familiarity with Agents [39, 56]. It has been pointed out that the very negative beliefs often happen to be misinformed or even semi-delusional: some catastrophic predictions about autonomous technologies are certainly based on reasonable concerns, motivated by the unprecedented, unpredictable effects of scale production and the rapidly evolving nature of Agents, while other worries seem prevalently based on techno-apocalyptic clichés, which often seem deliberately fuelled by conspiracyist propaganda. Distinguishing between the two is not always easy.

Thus, despite its inherent neutrality, Tolerance is a useful instrument to red-flag the kinds of extreme antagonistic attitudes that deserve to be investigated in greater depth. Our discussion motivates some hypotheses concerning these attitudes. These hypotheses can and should be empirically tested in future studies using the concept of Intolerance.

Table 5 Designing a scale for Autonomy Estrangement—sample questions

	Questionnaire to measure Autonomy Estrangement “How often do you experience the following feelings/wishes?” 5 possible answers (score 0–4): “Never” (0), “Just once or twice” (1), “Occasionally” (2), “Often” (3), “Consistently” (4)
Queries	(I.1) I feel that autonomous technologies make my work repetitive and boring
group I	(I.2) I feel that autonomous technologies make my activities riskless and predictable
Anxiety and disorientation	(I.3) I feel that autonomous technologies make my life empty and meaningless (I.4) I feel that autonomous technologies make my social relationships superficial and fake (I.5) I feel that autonomous technologies make my competency and skills useless and obsolete (I.6) I feel that autonomous technologies make my decisions pointless and insignificant
Queries	(I.7) I feel that autonomous technologies make my actions powerless and ineffective
group II	(I.8) I feel that autonomous technologies make my experiences insipid and flat
Hostility and frustration	(I.9) I feel that autonomous technologies make my role redundant and unneeded (I.10) I feel that autonomous technologies make my judgment disrespected and unauthoritative (I.11) I feel that autonomous technologies make my projects irrelevant and inconsequential (I.12) I feel that autonomous technologies make my dreams and desires futile and worthless (II. 1) I wish engineers and designers stopped developing new Agents (II. 2) I wish manufacturers stopped selling new Agents (II. 3) I wish families stopped using Agents as home assistants (II. 4) I wish managers stopped adopting Agents in their offices (II. 5) I wish businesses stopped relying on Agents (II. 6) I wish a global industrial crisis dramatically reduced the production of Agents (II. 7) I wish the parliament passed stricter laws to limit the use of Agents (II. 8) I wish citizens were more aware of the threat represented by Agents (II. 9) I wish consumers associations took action to limit the diffusion and the usage of Agents (II. 10) I wish a pirate virus disabled many AI-based devices in the world, including Agents (II. 11) I wish all Agents were forbidden and all their manufacturers went out of business (II. 12) I wish movements of activists started to systematically inactivate Agents everywhere (II. 13) I wish engineers and designers were punished for creating Agents
Autonomy Estrangement measured on a 5-degree scale	<i>Measurement method (tentative): sum the scores (from 0 to 4) of each of the 25 answers</i> <i>Calculate the total score (from 00 to 100)</i> 00 to 20 No Autonomy Estrangement 21 to 40 Sporadic or minimum Autonomy Estrangement 41 to 60 Minor Autonomy Estrangement 61 to 80 Major Autonomy Estrangement (may involve mild robo-phobia/techno-phobia) 81 to 100 Extreme Autonomy Estrangement (may involve severe robo-phobia/techno-phobia)

The first hypothesis that deserves to be tested empirically is that low Tolerance levels increase Autonomy Estrangement, which in turn discourages the adoption of autonomous technologies. Our conjecture is that Autonomy Estrangement negatively correlates with Adoption Propensity and tends to indirectly co-vary with Tolerance for Agents in combination with other variables, thus a high score of Autonomy Estrangement predicts a low score of Tolerance. The anxiety and hostility produced as specific expressions of Autonomy Estrangement reinforce negative narratives in opinion groups and ideological movements characterized by low Tolerance, further exacerbating their collective negative attitudes toward Agents.

The second hypothesis is that the most extreme judgments about Agents originate from militant discourses [81, 89] and thus, once examined carefully, might reveal the distinctive mark of an organized effort (motivated by ideological or political agendas), to discredit or undermine autonomous technologies, provoking a public reaction against their adoption. Proponents and sympathizers of the contemporary “anti-tech” discourse, like those who inspired Kaczynski, programmatically spread negative narratives toward autonomous technologies, portraying Agents as capable of and aiming to take over and subvert the social world [2, 84]. Also, the anarcho-primitivist and the neo-luddite movements openly advocate antagonistic and possibly violent reactions against Agents based on anti-modern and naturistic principles that only partly align with Kaczynski’s doctrine [89]. Some social psychologists have interpreted these reactions as symptoms of a claustrophobic anxiety deriving from the primordial perception of technology as a dehumanizing force [90, 91]. The notion of Intolerance is useful because it can help researchers verify whether the strongest negative responses against Agents correlate systematically with conspirative, possibly semi-delusional, discourses about autonomous systems, shedding light on the political and ideological source of Autonomy Estrangement [92]. While strongly corroborated by well-known facts, for now this correlation is just a hypothesis, to be validated empirically.

The third hypothesis is that Tolerance correlates with some of the variables that define an individual’s psychological profile, such as personality. Personality traits like “openness” and “neuroticism” [68] and personality-related variables like “affinity for technology interaction” or “technology commitment” [69] might exacerbate Tolerant or Intolerant attitudes. Alternatively, it is possible that they aggravate only the effects of Autonomy Estrangements. Other psychological and cultural variables (e.g., individual and collective preferences, anthropomorphic tendencies, innovativeness) might also help us illuminate the mechanisms underlying Tolerance.

The fourth hypothesis is that an Intolerant attitude tends to produce less obvious and more covert behavioural manifestations, compared to low Acceptance or Trust. The reasons of this asymmetry deserve to be investigated. We suspect that, in many cases, a person with low Tolerance toward Agents can unconsciously inhibit, deliberately hide, or even explicitly deny the hostile and anxious impulses associated with Intolerance, in a way that resembles how racist people hide their intolerant beliefs over specific outgroups: not only can such beliefs be biased and irrational, but the racist person may also be unaware of holding them, or inclined to deny to hold them because ashamed of their own racism, or simply unwilling to publicly admit to hold such views.⁵ It could also be because Intolerance, unlike Trust and Acceptance, is often conceptualized as a structural weakness, specifically an undesirable susceptibility, an uncontrollable rigidity, or a lack of resilience. Probably, most people are unwilling to admit they have a weakness, rigidity, or susceptibility of this kind.

The fifth hypothesis is that Intolerance for Agents shares a common root with intolerance for humans and possibly resonates with it. This is the kind of intolerance that, unfortunately, has historically tended to target certain ethnic groups and minorities.⁶ Both intolerance for humans and Intolerance for Agents tend to be grounded in dogmatic beliefs, not direct experience: like racist and sexist intolerance is not justified by the characteristic qualities and abilities of the target groups, so Intolerance for Agents is not necessarily motivated by an objective assessment of their functionalities and technical specifications; like racist intolerance is ultimately grounded into the unfortunate persuasion that one race is superior to the other, also technological Intolerance primarily appeals to the dogma of human exceptionalism, even when Agents are more efficient and efficacious than humans in target roles/activities.⁷ Thus, interrogating the irrational source

⁵ Another aspect to be considered is that concern and hope are not entirely symmetrical psychological states: concerns are more likely to reduce hope than hopes are likely to mitigate concerns because, while pessimistic beliefs identify immediate threats that cannot be overlooked in a logic of self-preservation, optimistic beliefs identify opportunities that, even if strategically important, are not necessarily indispensable to survival.

⁶ The word “tolerance”, like its synonym “toleration”, are prominent in the intellectual tradition of philosophical liberalism, where they indicate the ability or willingness to understand and dialogue with others, recognizing and respecting their “Right to difference” [93]. The corresponding vice, “intolerance”, indicates the lack or fragility of these qualities. Historically, intolerant attitudes (towards certain peoples, ideas, or ways of living) have too often manifested themselves in the form of bigotry and hate fuelled by irrational prejudices.

⁷ Immigrants and refugees have often tended to become the targets of the intolerance of autochthonous communities when they are believed to compete with local workers or threaten the group’s traditional identity. Also, immigrants are discriminated as they are believed to not deserve the autochthones’ jobs [94], regardless of how capable they are to do

of intolerance for humans can be useful to identify the dogmatic roots on which at least some forms of Intolerance for Agents are based.

We suspect that research on Intolerance for Agents can learn a lot from existing studies about intolerance for humans. An intolerant attitude is more than just personal antipathy, irritation, or subjective aversion motivated by conflictual circumstances or personal incompatibility: even when intolerance manifests itself through idiosyncratic behaviours, stimulated by extemporary intuitions and superficial feelings, such behaviours, intuitions, and feelings are just the *effects* of a pre-existing discriminatory attitude, not its motivating *cause*. Thus, the truest source of intolerance is to be found in a systematic, immovable rejection that, *even before being validated by the direct encounter with its live object, is already fuelled by the conviction that such an object is essentially inappropriate, deviant, or wrong*. According to Verkuyten, Adelman and Yogeeswaran [79], this conviction may originate either from *prejudice* (i. e., a negative evaluation of an individual or group, based on group membership) or from deliberation (i.e., a misguided inference): either way, intolerance tends to be impermeable to evidence because it presupposes a *normative* (albeit poor) *judgment* that was formed and dogmatically endorsed before direct experience, or even in spite of it. Most often, intolerance is an effect of a negative stereotype about a whole category or a group, i.e. a

Footnote 7 continued

well those jobs. In fact, this negative attitude is exacerbated when said immigrants are highly qualified, hardworking, and respectful of local customs and traditions. Similarly, a new employee or student may attract the intolerance of their colleagues if (s)he tries to modernise the work culture. It is not that they necessarily ignore the value of innovation: whether the colleagues like or not the novelty brought by the newcomer, what triggers intolerance is the belief that it is not up to the newcomer to make such changes. Similar forms of intolerance may occur in the romantic sphere when one of the partners, due to work-related reasons, has to spend time alone with a younger, pleasant, competent, and morally irreprehensible assistant, provoking their spouse's jealousy: again, the problem is not who the assistant is or the kind of assistance that they bring, but the belief that no assistant is entitled to dispense those attentions to the spouse. Thus, this kind of intolerance does not arise from a personal or cultural incompatibility but from a prejudice about the persons and the groups involved and, more specifically, about their status and role, possibly exacerbated by political discourses that explicitly defends the prejudice by leveraging narratives of competition and rivalry [78].

These kinds of biased attitudes correlate with dogmatic narratives fuelling anxious and hostile feelings toward the other, the different, the unexpected. They are more obvious in homogeneous, traditionalist environments that are unreceptive to innovation and adverse to change [79]. Arguably, they can be found more often in highly hierarchical or sectarian environments where recognition, authority, and decisional power are primarily justified by features (such as seniority, wealth, gender, ethnicity, etc.) that, although irrelevant to leadership, conventionally define one's standing in the group more than the objective qualities and capabilities of the individuals. Newcomers, strangers, and competitors may attract even more intolerant reactions if they are perceived as overqualified, overzealous, and too close to the upper management: they can become the target of intolerance even if nobody has a personal issue

poorly motivated or lazy judgment that the intolerant person is eager to confirm without testing or revising [79].

The classical concept of toleration/tolerance developed in the context of human relationships, but it is in many ways capable to illuminate the notion of Tolerance for Agents. It seems very likely that tolerance for humans and Tolerance for Agents share a communal psychological and cultural root and it is even possible that they reinforce one another. This remains to be investigated empirically, using specifically designed protocols. Nevertheless, some fundamental normative differences between Tolerance for Agents and tolerance for humans are probably beyond empirical investigation, being essentially metaphysical in nature or deeply ingrained in our value system: discriminating humans based on their ethnicity or gender is always morally unacceptable, but discriminating between humans and Agents may be an appropriate choice justified by scientific, design, and moral counts.

The sixth hypothesis to be investigated empirically concerns the prevention of irrationally biased Intolerant attitudes and the mitigation of extreme forms of Autonomy Estrangement. The history of racism and discrimination suggests that the intolerant's mind is impermeable to critical revision and self-diagnosis, as intolerance for human neighbours, immigrants or colleagues is unlikely to disappear even when these individuals are perfectly agreeable people and behave impeccably. The source of the intolerant negative biases toward specific groups and races remains beyond utilitarian or hedonistic considerations [54]. Similarly, Intolerance for Agents is neither caused by the Agent's demeanour and appearance nor by an objective incompatibility with the user: its primary cause remains the Intolerant user's narrow perspective and rigid stance. What are the most effective remedies and mitigation strategies for a problem that is more in the mind of the judger than in the objective features (behaviour, character, competences, etc.) of the judged? The first approach, to be tested empirically, consists in countering deeply held negative beliefs through supervised positive experiences with Agents, not only aiming at familiarizing with autonomous technologies but also at dissolving categorical prejudices, deeply ingrained biases, and hostile narratives. Another approach to be tested empirically is the one postulating that the formation and petrification of discriminatory prejudices can be prevented by early educational interventions informed by values like diversity, equity, and inclusion [96]. This assumption would have to be tested against the competing assumption that Intolerance is primarily fuelled by personality traits, which are only partly shaped by education (especially when they

Footnote 7 continued

with them or dislikes what they do. Conversely, disapproval and disrespect do not necessarily accompany or cause intolerance, which may in fact arise also in presence of explicit approval and respect [95].

Table 6 Synoptic table of the differences between key attitudes toward Agents

	Tolerance	Acceptance	Trust
Key question	Is it fair and right that the Agent occupies roles and takes tasks usually assigned to humans?	Does the Agent belong in the work/family environment and is it perceived as part of the team?	Does the Agent operate in an efficacious, efficient, reliable, secure, and safe manner when it conducts a certain job?
Core idea	The Agent deserves to join our group and/or to work in a certain role.	It is nice (pleasant, comforting, rewarding, effortless, etc.) to work with the Agent.	The Agent works well (efficiently, reliably, according to expectations...)
Scope	Categorical normative judgements and beliefs (possibly including unmotivated prejudices and irrational biases) about A.	The quality of the relationship between the Agent and its users, as it is experienced by the users through direct interaction.	The outcome of the work done by the Agent, and whether or not the Agent performed well, based on the users' expectations.
Values	Authenticity, autonomy, self-reliance, agency, purpose.	Familiarity, harmonisation, engagement, inclusive relationship.	Predictability, explainability, reliability, traceability, safety.
Expectation	A meaningful, fulfilling, and self-determined human existence	An appropriate, pleasant, and stable interaction.	An outcome considered desirable and needed and hoped for.
Indicators	Quality of life outside of HAI: self-reliance, purpose	Quality of HAI: familiar, harmonious, inclusive, sanctioned	Quality of the outcome of HAI: expectations and hopes about performance.
Worries and concerns	Concern that Agents may alter the users' customs and practices (A1).	Working with the Agent is a frustrating and irritating experience (as opposed to rewarding and fulfilling).	The information provided by the Agent is irrelevant and difficult to understand (as opposed to comprehensible and clear).
	Concern that Agents may alter the socio-political relations among users (A2).	Interacting with the Agent feels uncanny and creepy (as opposed to familiar, confident and comfortable)	The Agent responds to orders slowly and imprecisely (as opposed to promptly and accurately).
	Concern that Agents may replace humans in their jobs and key functions (B1).	The Agent is an isolated and solitary entity (as opposed to a well-integrated member of the team).	Interaction with/handling of the Agent is unsafe and not secure (as opposed to safe and secure).
	Concern that Agents may undermine the users' authority and decisional power (B2).	The Agent is disliked by people and attracts negative feelings like antipathy and hostility (as opposed to sympathy and friendship).	Previous interactions with Agents were typically unsuccessful or useless (as opposed to successful and useful).
		The Agent's attitude is unhelpful and unavailable (as opposed to helpful and available).	Opacity 1: The reasons/plans motivating the Agent's behaviour/decisions are difficult to understand (as opposed to easily explainable).
	Concern that Agents may question the human exceptionalism, challenging their unique status/entitlements (B3).	The Agent appears opaque and alien, it is difficult to relate to it (as opposed to: Agent solicits empathy, transfer, projection, i.e. people see themselves reflected in it).	Opacity 2: the intended effects of the Agent's behaviour/decisions are difficult to guess (as opposed to easily predictable).
		The Agent's features and behaviours appear rigid, inflexible, stereotyped (as opposed to: they adjust to the team members and their preferences and habits).	The Agent appears intelligent, perceptive, and knowledgeable about its assigned task (as opposed to insensitive, ignorant of the context).
		The Agent displays no collaborative intent and willingness to help (as opposed to being collaborative and involved in the task)	The Agent displays no collaborative intent and willingness to help (as opposed to being collaborative and involved in the task)
			The Agent requires little or no maintenance, technical expertise.
Areas of overlap	No overlap	The relationship with the Agent is affected positively/negatively by the fact that Agent operates more or less reliably and in a clearly explainable/predictable manner.	
Areas of overlap	The relationship with the Agent is affected positively/negatively by the fact that the user has certain biases and prejudices about their status and role.		No overlap
Areas of overlap	The trust in the Agent is affected positively/negatively by the user's biases and prejudices.	No overlap	The trust in the Agent is affected positively/negatively by the user's biases and prejudices.

seem hereditary, consolidated during early development, or continuously reinforced by the family environment).

7 Conclusions

With a view to concretely apply the concepts examined so far, let us summarize the main outcomes of our analyses: we call “Tolerance” (for Agents) the user’s insusceptibility or resilience to Autonomy Estrangement, that is the suffocating sense of isolation, displacement, and frustration that human users may experience when they think (for right or wrong reasons) that their lives are (or risk being) controlled by Agents, i.e. by intelligent systems that, while created by humans, operate independently of human control.

Autonomy Estrangement is best theorized in terms of the (real or just feared) impoverishment of the human life allegedly caused by the dominating presence of automated monitoring and control systems. When suffering from Autonomy Estrangement, technology users experience anxiety/disorientation over disruptive changes (configured as disruption of *modus operandi* and way of living and/or disruption of the network of socio-political practices) and/or hostility/frustration against technological dominance (configured as fear of being made redundant, losing authority, and/or losing human uniqueness).

For convenience, we use the term “Intolerance” to indicate the property inverse to Tolerance, that is the user’s susceptibility or proneness to Autonomy Estrangement. Thus, Tolerance and Intolerance are inverse representations of the same phenomenological continuum, with Intolerance increasing when Tolerance decreases and vice versa. Whether the user has a Tolerant or Intolerant attitude for a certain Agent primarily depends on the beliefs that the user holds about Agents’ capacity to impoverish the human life (i.e., including, but not limited to, negative biased narratives about autonomous technology) and only secondarily on how such beliefs are supported by experiences of direct interaction with Agents (i.e., the positive and negative experiences that respectively justify and undermine the belief that Agent can impoverish the human life). Due to the self-reinforcing nature of discriminatory biases (via selective attention and confirmation), Intolerant beliefs about the Agent are likely to be reinforced by negative interactions without being equally corrected by positive interactions.

The difference between Acceptance, Trust, and Tolerance is crucial. Not all Intolerant users display their negative emotions when dealing with Agents and, conversely, the presence of negative emotions in a user is not a sufficient reason to infer that they poorly tolerate an Agent, as the same emotions could be triggered by other dynamics (e.g., Acceptance and Trust, when the Agent does not operate reliably and is not easy to interact with). Combining different diagnostic heuristics,

we can develop three independent multi-dimensional scales that concretely distinguish between Tolerance, Acceptance, and Trust. Table 6 summarizes their distinctive features, sketching the variables that would need to be considered when developing and validating an appropriate measurement instrument.

An important reason to introduce Tolerance in HAI research is that the distinction between Tolerance, on the one hand, and Acceptance and Trust, on the other, reflect a key distinction between different kinds of stakeholders: The *users*, i.e., those whom will have to work with (and possibly be replaced by) Agents; and the *owners*, i.e., decision-makers like employers or political authorities, who purchase Agents because they find their adoption convenient. Their perspectives do not necessarily align and often tend to be diametrically opposed. Whereas an account that relies only on Acceptance and Trust alone might conceal this divergence, a Tolerance-based account is likely to show any polarisation between these two groups.

HAI Researchers tend to focus on whether Agents function efficiently and reliably in general, as this indicates that they are inherently desirable by different kinds of stakeholders. However, researchers cannot assume that users and owners have the same goals and expectations, as if the satisfaction of users automatically implied the success of the owners and vice versa. This inference is unwarranted because it builds on the tacit assumption that all stakeholders are unbiased, rational agents whose individual interests and goals perfectly align with those of their organisation/group. In many everyday real-life scenarios, individuals may even have personal motivations (not necessarily ethical or rational) to work against their teammates, sabotaging team outcomes. HAI researchers thus should not overlook such individual motivations, because neglecting the influence exerted by envy, fear, egoism, and prejudice, inevitably means becoming unable to account for several socio-psychological dynamics associated with technological unemployment and the conflicts experienced by humans forced to compete with Agents [66, 67].

HAI research needs to introduce the notion of Tolerance precisely because a model relying exclusively on Acceptance and Trust cannot account for the whole spectrum of motivations (including personal preferences, worries, and aspirations) that determine Adoption Propensity, beyond what the users or the employers are willing to explicitly state or confess to themselves. An Agent that typically scores very well in Acceptance and Trust is an agent that an employer and HR manager would find very promising and desirable [25], but it would not necessarily be well-tolerated by other users, precisely because these users know that the goals of the company (profit maximisation, minimization or externalisation of liabilities, reduction of dependency on human labour) do not necessarily coincide with their individual goals. In this

case, the employer has reasons to see the Agent as a competitor or an usurper because the Agent has all the relational qualities and the professional skills necessary to do their job in a way that satisfies the expectations of the employers.

Tolerance is the construct we need to account for these dynamics because it is best equipped to map judgments and beliefs about Agents (including prejudices and stereotypes), independently of how the users express them during their experience of direct interaction with the Agent. An account that tracks Tolerance, in addition to Acceptance and Trust, is preferable because it is in the position to finally represent all parties and interests involved.

Author's Contribution Each named author has substantially contributed to conducting the underlying research and drafting this manuscript and all of the authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This material is based upon work supported by the Air Force Office of Scientific Research and the Minerva Initiative under award number FA9550-18-1-0039. The work has also received the support of the Australian Government.

Data Availability Statement Not applicable, as this is a theoretical position paper, not an empirical study.

Code Availability Not applicable.

Declarations

Conflict of interest The authors have no conflicts of interest.

Research Involving Human Participants and/or Animals Not applicable.

Consent for Participation Not applicable.

Consent for Publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ford M (2015) Rise of the robot: Technology and the threat of a jobless future. Basic Books, New York.
- Kaczynski TJ (1995) The unabomber trial: the manifesto. The Washington Post. Retrieved April 29, 2011.
- Ford M (2021) Rule of the robot: How artificial intelligence will transform everything. Basic Books, New York.
- Cools H, Van Gorp B, Opgenhaffen M (2022) Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers. *Journalism*. <https://doi.org/10.1177/14648849221122647>
- Laumer S, Eckhardt A (2012) Why do people reject technologies: a review of user resistance theories. In: Dwivedi YK, Wade MR, Schneberger SL (eds) *Information systems theory: explaining and predicting our digital society*, 1 (pp. 63–86). Springer, Cham. https://doi.org/10.1007/978-1-4419-6108-2_4
- Beer JM, Fisk AD, Rogers WA (2014) Toward a framework for levels of robot autonomy in human–robot interaction. *J Human-Robot Interact* 3(2):74–99
- Cappuccio ML, Galliot JC, Sandoval EB (2022) Saving private robot: risks and advantages of anthropomorphism in agent-soldier teams. *Int J Soc Robot* [Advance online publication] 14(10):2135–2148. <https://doi.org/10.1007/s12369-021-00755-z>
- Dautenhahn K (2007) Socially intelligent robots: dimensions of human–robot interaction. *Philos Trans Roy Soc Lond Ser B Biol Sci* 362(1480):679–704. <https://doi.org/10.1098/rstb.2006.2004>
- Han J, Conti D (2020) The use of UTAUT and post acceptance models to investigate the attitude towards a telepresence robot in an educational setting. *Robotics* 9(2):34. <https://doi.org/10.3390/robotics9020034>
- Heerink M, Krose B, Evers V, Wielinga B (2009) Measuring acceptance of an assistive social robot: a suggested toolkit. *RO-MAN 2009*:528–533. <https://doi.org/10.1109/ROMAN.2009.5326320>
- Galliot JC (2018) The soldier's tolerance for autonomous systems. *Paladyn J Behav Robot* 9(1):124–136. <https://doi.org/10.1515/pjbr-2018-0008>
- Galliot JC, Baggiarini B, Rupka S (2021) Empirical data on attitudes towards autonomous systems'. In: Galliot J, Ohlin JD, MacIntosh D (eds) *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare*. Oxford University Press, Oxford.
- Kaczynski TJ (2016) *Anti-tech revolution: Why and how*. Fitch and Madison Publishers, Scottsdale.
- Moen OM (2019) The unabomber's ethics. *Bioethics* 33(2):223–229. <https://doi.org/10.1111/bioe.12494>
- Samhan B (2018) Revisiting technology resistance: Current insights and future directions. *Austr J Inform Syst*, 22. <https://doi.org/10.3127/ajis.v22i0.1655>
- Nomura T, Kanda T, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *AI & Soc* 20(2):138–150. <https://doi.org/10.1007/s00146-005-0012-7>
- Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J Soc Robot* 1(1):71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- Breazeal C, Brooks A, Chilongo D, Gray J, Hoffman G, Kidd C, Lee H, Lieberman J, Lockerd A (2004) Working collaboratively with humanoid robots. In: *Proceedings of the IEEE-Ras/RSJ international conference on humanoid robots* (pp. 253–272). Santa Monica, CA.
- Cappuccio ML, Galliot JC, Sandoval EB (2021b) Mapping meaning and purpose in human-robot teams: anthropomorphic agents in military operations. *Philos J Conflict Violence* 5(1):73–94. <https://doi.org/10.22618/TP.PJCV.20215.1.139005>
- Baker AL, Phillips EK, Ullman D, Keebler JR (2018) Toward an understanding of trust repair in human–robot interaction: current research and future directions. *ACM Trans Interact Intell Syst* 8(4):1–30. <https://doi.org/10.1145/3181671>

21. Katz JE, Halpern D (2014) Attitudes towards robots suitability for various jobs as affected robot appearance. *Behav Inform Technol* 33(9):941–953. <https://doi.org/10.1080/0144929X.2013.783115>
22. Bishop L, van Maris A, Dogramadzi S, Zook N (2019) Social robots: the influence of human and robot characteristics on acceptance. *Paladyn J Behav Robot* 10(1):346–358. <https://doi.org/10.1515/pjbr-2019-0028>
23. Nanava S, Sarda Gou M, Webb TL, Prescott TJ (2020) A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *Int J Soc Robot* 12(6):1179–1201. <https://doi.org/10.1007/s12369-020-00659-4>
24. Bartneck C, Nomura T, Kanda T, Suzuki T, Kennsuke K (2005) A cross-cultural study on attitudes towards robots. In: *Proceedings of the HCI international, Las Vegas*.
25. Cohen AJ (2004) What toleration is. *Ethics* 115(1):68–95. <https://doi.org/10.1086/421982>
26. Rogers EM (1962) *Diffusion of innovations*. Free Press of Glencoe, USA.
27. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 13(3):319–340. <https://doi.org/10.2307/249008>
28. Venkatesh V, Morris MG, Davis GB, Davis FD (2003) User acceptance of information technology: toward a unified view. *MIS Q* 27(3):425–478. <https://doi.org/10.2307/30036540>
29. De Ruyter B, Saini P, Markopoulos P, van Breemen A (2005) Assessing the effects of building social intelligence in a robotic interface for the home. *Interact Comput* 17(5):522–541. <https://doi.org/10.1016/j.intcom.2005.03.003>
30. Bröhl C, Nelles J, Brandl C, Mertens A, Nitsch V (2019) Human–robot collaboration acceptance model: development and comparison for Germany, Japan, China and the USA. *Int J Soc Robot* 11(5):709–726
31. Beer JM, Prakash A, Mitzner TL, Rogers WA (2010) Understanding robot acceptance. Technical Report HFA-TR-1103. Georgia Institute of Technology School of Psychology—Human Factors and Aging Laboratory.
32. Venkatesh V, Davis FD (2000) A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manage Sci* 46(2):186–204. <https://doi.org/10.1287/mnsc.46.2.186.11926>
33. Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Acad Manag Rev* 20(3):709–734. <https://doi.org/10.2307/258792>
34. McKnight DH, Cummings LL, Chervany NL (1998) Initial trust formation in new organizational relationships. *Acad Manag Rev* 23(3):473–490. <https://doi.org/10.2307/259290>
35. McKnight DH, Carter M, Clay P (2009) Trust in technology: Development of a set of constructs and measures. *Digit Proceedings*, 10. <http://aisel.aisnet.org/digit2009/10>.
36. Biros DP, Daly M, Gunsch G (2004) The influence of task load and automation trust on deception detection. *Group Decis Negot* 13(2):173–189. <https://doi.org/10.1023/B:GRUP.0000021840.85686.57>
37. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. *Hum Factors* 46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_30392
38. Oleson KE, Billings DR, Kocsis V, Chen JYC, Hancock PA (2011) Antecedents of trust in human–robot collaborations. In: *IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)*, 2011 (pp. 175–178). <https://doi.org/10.1109/COGSIMA.2011.5753439>
39. Hald K, Rehm M, Moeslund TB (2019) Proposing human–robot trust assessment through tracking physical apprehension signals in close-proximity human–robot collaboration. In: *28th IEEE international conference on robot and human interactive communication (RO-MAN)*, 2019 (pp. 1–6). <https://doi.org/10.1109/RO-MAN46459.2019.8956335>
40. Devitt SK (2018) Trustworthiness of autonomous systems. In: Abbass HA, Scholz J, Reid D (eds) *Foundations of trusted autonomous systems*. Springer, Cham, pp 161–184
41. Roff HM, Danks D (2018) ‘Trust but Verify’: the difficulty of trusting autonomous weapons systems. *J Military Ethics* 17(1):2–20. <https://doi.org/10.1080/15027570.2018.1481907>
42. Lewis M, Sycara K, Walker P (2018) The role of trust in human–robot interaction. In: Abbass HA, Scholz J, Reid DJ (eds) *Foundations of trusted autonomy* (pp. 135–159). Springer, Cham. https://doi.org/10.1007/978-3-319-64816-3_8
43. Lyons J, Nam CS (2021) Introduction: the evolution of trust in human–robot interaction. *Trust Hum-Robot Interact*. <https://doi.org/10.1016/b978-0-12-819472-0.09988-3.pp.XXI-XXV>
44. Hancock PA, Billings DR, Schaefer KE, Chen JY, de Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human–robot interaction. *Hum Factors* 53(5):517–527. <https://doi.org/10.1177/0018720811417254>
45. Wagner AR (2009) The role of trust and its relationship in human–robot social interaction, Georgia Tech Theses and Dissertations, 22898 (submitted).
46. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K (2015) Would you trust a (faulty) robot? Effects of error, task type and personality on human–robot cooperation and trust. In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 141–148).
47. Glendinning C (1990) Notes toward a neo-Luddite manifesto. Retrieved February 4, 2022. Anarchist Library.
48. Gaudiello I, Zibetti E, Lefort S, Chetouani M, Ivaldi S (2016) Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Comput Hum Behav* 61:633–655. <https://doi.org/10.1016/j.chb.2016.03.057>
49. Kassim ES, Jailani SFAK, Hairuddin H, Zamzuri NH (2012) Information system acceptance and user satisfaction: the mediating role of trust. *Procedia—Social and Behavioral Sciences*, 57, 412–418. <https://doi.org/10.1016/j.sbspro.2012.09.1205>
50. Bano S, Shah UU, Ali S (2019) Personality and technology: Big five personality traits as descriptors of universal acceptance and usage of technology UTAUT.
51. Malle BF, Ullman D (2021) A multidimensional conception and measure of human–robot trust. *Trust Hum-Robot Interact*, pp 3–25. <https://doi.org/10.1016/b978-0-12-819472-0.00001-0>
52. Ullman D, Malle BF (2018) What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In: *Companion of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 263–264).
53. Mathur MB, Reichling DB (2009) An uncanny game of trust: Social trustworthiness of robots inferred from subtle anthropomorphic facial cues. In: *Proceedings of the 4th ACM/IEEE international conference on human robot interaction* (pp. 313–314), New York. <https://doi.org/10.1145/1514095.1514192>
54. Lee WH, Lin CW, Shih KH (2018) A technology acceptance model for the perception of restaurant service robots for trust, interactivity, and output quality. *Int J Mobile Commun* 16(4):361. <https://doi.org/10.1504/IJMC.2018.092666>
55. de Graaf MMA, Ben Allouch S, van Dijk JAGM (2019) Why would I use this in my home? A model of domestic social robot acceptance. *Hum-Comput Interact* 34(2):115–173. <https://doi.org/10.1080/07370024.2017.1312406>
56. Sandoval EB, Mubin O, Obaid M (2014) Human robot interaction and fiction: a contradiction. In: Beetz M, Johnston B, Williams MA (eds) *Social robotics. Lecture Notes in Computer Science*. ICSR 2014, vol 8755. Springer, Cham. https://doi.org/10.1007/978-3-319-11973-1_6.

57. Moradi M, Moradi M, Bayat F (2018) On robot acceptance and adoption a case study. In: 8th Conference of AI y Robotics and 10th RoboCup Iran open International Symposium (IRANOPEN), 2018. <https://doi.org/10.1109/RIOS.2018.8406626>
58. Brühl C, Nelles J, Brandl C, Mertens A, Schlick CM (2016) TAM reloaded: a technology acceptance model for human–robot cooperation systems. In: HCI international 2016 - Posters' extended abstracts (pp. 97–103). Springer, Cham.
59. de Graaf MMA, Ben Allouch S (2013) Exploring influencing variables for the acceptance of social robots. *Robot Auton Syst* 61(12):1476–1486. <https://doi.org/10.1016/j.robot.2013.07.007>
60. Lotz V, Himmel S, Zieffle M (2019) Acceptance factors for human-robot collaboration in industry. In: Dimitrov D, Hagedorn-Hansen D, von Leipzig K (eds) Proceedings of the International Conference on Competitive Manufacturing Y. M. Mate. (p. COMA2019). Stellenbosch, South Africa, 405–411.
61. Schaefer KE (2016) Measuring trust in human–robot interactions: development of the “Trust Perception Scale-HRI”. *Robust Intelligence and Trust in Autonomous Systems*, pp 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10
62. Schaefer KE, Perelman BS, Gremillion GM, Marathe AR, Metcalfe JS (2021) A road map for developing team trust metrics for human-autonomy teams. *Trust in Human-Robot Interaction*, pp 261–300. <https://doi.org/10.1016/b978-0-12-819472-0.00012-5>
63. Sifakis J (2019) Can we trust autonomous systems? Boundaries and risks. In: Chen YF, Cheng CH, Esparza J (eds) Automated technology for verification and analysis. *Lecture Notes in Computer Science*. ATVA 2019, 11781. Springer, Cham. https://doi.org/10.1007/978-3-030-31784-3_4
64. van Pinxteren MME, Wetzels RWH, Rüger J, Pluymaekers M, Wetzels M (2019) Trust in humanoid robots: implications for services marketing. *J Serv Mark* 33(4):507–518. <https://doi.org/10.1108/JSM-01-2018-0045>
65. Lui J, Narsalay R, Afzal R, Nair Sharma I, Light D (2022) Research: how do warehouse workers feel about automation? *Harvard Business Review*. Retrieved February 11, 2022.
66. Reich N, Eyssel F (2013) Attitudes towards service robots in domestic environments: the role of personality characteristics, individual interests, and demographic variables. *Paladyn, J Behav Robot* 4(2):123–130. <https://doi.org/10.2478/pjbr-2013-0014>
67. Zhang B, Dafoe A (2019) Artificial intelligence: American attitudes and trends. University of Oxford. <https://doi.org/10.2139/ssrn.3312874>
68. Ellul J (1964) The technological society. J. Wilkinson (Trans.) (pp. 133–143). Random House.
69. Bernotat J, Eyssel F (2018) Can't wait to have a robot at home? - Japanese and German users' attitudes toward service robots in smart homes. In: 27th IEEE international symposium on robot and human interactive communication (RO-MAN), 2018 (pp. 15–22).
70. Krämer NC, von der Pütten A, Eimler S (2012) Human-agent and human–robot interaction theory: Similarities to and differences from human–human Interaction. In: *Studies in Computational Intelligence*, pp (215–240). https://doi.org/10.1007/978-3-642-25691-2_9
71. Useem J (2015) Why it pays to be a jerk. New research confirms what they say about nice guys. The Atlantic.com. <https://www.theatlantic.com/magazine/archive/2015/06/why-it-pays-to-be-a-jerk/392066/>, June.
72. Dibrov A (2015) Innovation resistance: The main factors and ways to overcome them. *Proc Soc Behav Sci* 166:92–96. <https://doi.org/10.1016/j.sbspro.2014.12.489>
73. Edison SW, Geissler GL (2003) Measuring attitudes towards general technology: antecedents, hypotheses and scale development. *J Target Meas Anal Mark* 12(2):137–156. <https://doi.org/10.1057/palgrave.jt.5740104>
74. Dang J, Liu L (2021) Robots are friends as well as foes: ambivalent attitudes toward mindful and mindless AI robots in the United States and China. *Comput Hum Behav* 115:106612. <https://doi.org/10.1016/j.chb.2020.106612>
75. Stapels JG, Eyssel F (2022) Robocalypse? Yes, please! the role of robot autonomy in the development of ambivalent attitudes towards robots. *Int J Soc Robot* 14(3):683–697. <https://doi.org/10.1007/s12369-021-00817-2>
76. Ariely D (2011) The upside of irrationality. HarperCollins Publishers, The unexpected benefits of defying logic at work and at home
77. Pirsig RM (1975) Zen and the art of motorcycle maintenance. Bantam Press, American Psychological Association
78. Sibley CG, Duckitt J (2008) Personality and prejudice: a meta-analysis and theoretical review. *Pers Soc Psychol Rev* 12(3):248–279. <https://doi.org/10.1177/1088868308319226>
79. Verkuyten M, Adelman L, Yogeewaran K (2020) The Psychology of intolerance: unpacking diverse understandings of intolerance. *Curr Dir Psychol Sci* 29(5):467–472. <https://doi.org/10.1177/0963721420924763>
80. Mori M, MacDorman K, Kageki N (2012) The uncanny valley [From the Field]. *IEEE Robot Automat Magaz* K.F. MacDorman, N. Kageki (Trans.) 19(2):98–100. <https://doi.org/10.1109/MRA.2012.2192811>
81. Kiberd R (2015) Burn it all down: A guide to neo-luddism. *Gizmodo*, 28 (January), Retrieved February 4, 2022.
82. Fleming S (2022) The Unabomber and the origins of anti-tech radicalism. *J Polit Ideol* 27(2):207–225. <https://doi.org/10.1080/13569317.2021.1921940>
83. Paulsen R (2015) Empty labor. Cambridge University Press, Idleness and workplace resistance
84. Gardenier M (2016) The “anti-tech” movement, between Anarcho-primitivism and the neo-Luddite. *Sociétés d'Afrique and SIDA: Newsletter of the Sociétés d'Afrique and SIDA Network* 131(1):97–106
85. Liang Y, Lee SA (2017) Fear of autonomous robots and artificial intelligence: evidence from national representative data with probability sampling. *Int J Soc Robot* 9(3):379–384. <https://doi.org/10.1007/s12369-017-0401-3>
86. McClure PK (2018) 'you're fired', says the robot. *Soc Sci Comput Rev* 36(2):139–156. <https://doi.org/10.1177/0894439317698637>
87. Roszak T (1994) The cult of information: A neo-Luddite treatise on high-tech, artificial intelligence, and the true art of thinking. University of California Press, California
88. Bartneck C (2004) From Fiction to Science—a cultural reflection on social robots. In: Proceedings of the CHI2004 workshop on shaping human-robot interaction, Vienna.
89. Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. *Hum Factors* 39(2):230–253. <https://doi.org/10.1518/001872097778543886>
90. Nomura T, Kanda T, Suzuki T, Kato K (2008) Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans Robot* 24(2):442–451. <https://doi.org/10.1109/TRO.2007.914004>
91. Nomura T, Suzuki T, Kanda T, Kato K (2006) Measurement of negative attitudes toward robots. *Interaction Studies. Social Behav Commun Biol Artif Syst* 7(3):437–454. <https://doi.org/10.1075/is.7.3.14nom>
92. Cotgrove S (1972) Alienation and automation. *Br J Sociol* 23(4):437–451. <https://doi.org/10.2307/588323>
93. Locke J (1689) A letter concerning toleration. In: Tully (ed). Hackett, 1983.
94. Brader T, Valentino NA, Suhay E (2008) What triggers public opposition to immigration? Anxiety, group cues, and immigration

- threat. *Am J Polit Sci* 52(4):959–978. <https://doi.org/10.1111/j.1540-5907.2008.00353.x>
95. Simon B, Eschert S, Schaefer CD, Reininger KM, Zitzmann S, Smith HJ (2019) Disapproved, but tolerated: the role of respect in outgroup tolerance. *Pers Soc Psychol Bull* 45(3):406–415. <https://doi.org/10.1177/0146167218787810>
 96. Verkuyten M, Yogeewaran K, Adelman L (2020) Toleration and prejudice-reduction: two ways of improving intergroup relations. *Eur J Soc Psychol* 50(2):239–255. <https://doi.org/10.1002/ejsp.2624>
 97. Daza M, Barrios-Aranibar D, Diaz-Amado J, Cardinale Y, Vilasboas J (2021) An approach of social navigation based on proxemics for crowded environments of humans and robots. *Micromachines* 12(2). <https://doi.org/10.3390/mi12020193>
 98. Cappuccio ML, Peeters A, McDonald W (2020) Sympathy for dolores: moral consideration for robots based on virtue and recognition. *Philos Technol* 33(1):9–31. <https://doi.org/10.1007/s13347-019-0341-y>
 99. Avent R (2018) How robots will break politics. Automation is dramatically reshaping the workforce, but we've barely begun to grapple with how it will reshape society. *Politico.com*. <https://www.politico.com/magazine/story/2018/01/05/robots-politics-automation-technology-216220/>, February.
 100. Saunderson SP, Nejat G (2021) Persuasive robots should avoid authority: the effects of formal and real authority on persuasion in human–robot interaction. *Sci Robot* 6(58):eabd5186. <https://doi.org/10.1126/scirobotics.abd5186>
 101. Haring KS, Satterfield KM, Tossell CC, de Visser EJ, Lyons JR, Mancuso VF, Finomore VS, Funke GJ (2021) Robot authority in human–robot teaming: effects of human-likeness and physical embodiment on compliance. *Front Psychol* 12:625713. <https://doi.org/10.3389/fpsyg.2021.625713>
 102. Christakis NA (2019) How AI will rewire us. For better and for worse, robots will alter humans' capacity for altruism, love, and friendship. *The Atlantic.com*. <https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/>, April.
 103. Ferrari F, Paladino MP, Jetten J (2016) Blurring human–machine distinctions: anthropomorphic appearance in social robots as a threat to human distinctiveness. *Int J Soc Robot* 8(2):287–302. <https://doi.org/10.1007/s12369-016-0338-y>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Massimiliano Cappuccio is a Senior Researcher in the School of Engineering & Technology of University of New South Wales Canberra, where he currently is a member of the Trusted Autonomy group and a founding member of the AI Hub, after operating for several years as the Deputy director of the Values in Defence and Security Technology group. As a cognitive philosopher and a technology ethicist, his research on intelligent systems is interdisciplinary and aims to integrate phenomenological analyses, empirical experimentation, synthetic modelling, and strategic policy intervention. His current research focuses on human-machine interaction (social robotics and human-agent teaming), human performance (skill acquisition and disruption), and theory and ethics of artificial intelligence. His earlier work has focused on several aspects of social cognition, embodied cognition, and the theoretical foundations of artificial intelligence. During his career, he has published more than a hundred original scientific contributions in top-tier peer-reviewed journals, collective volumes, and international conference proceedings. He is the principal investigator of the project “Cultural Attitudes in the Age of

Autonomy. Australian Policy and the Ethico-Legal Deployment of Human-Machine Teams in the Indo-Pacific”, sponsored by the Australian Department of Defence. Also, he is the author of the upcoming monograph “The Skilled Mind. Habits, Practices, and Knowledge” for the MIT Press. Also, he conducts an intense activity as an organiser and coordinator of high-impact scientific events, such as: the Pacific Rim Strategic Policy Conference on International Security and the Rise of Autonomous Weapons; the TEMPER workshop on Training, Enhancement, and Military Performance; the ‘Long Road Ahead’ Seminar Series on Autonomous, Connected and Smart Transportation; and the annual Joint UAE Symposium on Social Robotics.

Jai Galliot is Honorary Associate Professor in the Faculty of Business & Law at the University of Wollongong. Prior, he was Director of the Values in Defence & Security Technology Group at the University of New South Wales at Australian Defence Force Academy, where he led a multidisciplinary team of experts on the ethical, legal and social questions associated with military technology, and sat on the Faculty & University boards. He previously held visiting appointments at the Modern War Institute at the West Point in the USA, the Skolkovo Institute for Science & Technology in Russia, and the Centre for Technology & Global Affairs in the Department of Politics and International Relations at the University of Oxford in the UK. He is further a former Royal Australian Navy Officer. Most of his work falls at the intersection of philosophy, technology and security, with a particular focus on AI, autonomous systems, human enhancement and big data. Galliot is currently editing the Oxford Handbook of Remote War (Oxford University Press 2024/5) and has authored/edited the following books: *Military Robots: Mapping the Moral Landscape* (Routledge 2016); *Ethics & the Future of Spying* (Routledge 2015); *Force Short of War in Modern Conflict* (Edinburgh University Press 2019); *Big Data and Democracy* (Edinburgh University Press 2020); & *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare* (Oxford University Press 2021). Galliot has secured over \$10 million as CI/PI in competitive research funding for work on technology ethics, including a Welsh Sêr Cymru II Fellowship for work on the individualisation of war, a US Government Minerva award for research on the soldier's tolerance for autonomous systems, and Australian Government awards for work on the ethics of AI.

Friederike Eyssel is Professor of Psychology and head of the research group “Applied Social Psychology and Gender Research” at the Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Germany. Dr. Eyssel is passionate about basic and applied social psychological research and she is interested in various research topics ranging from social robotics, trust, and acceptance of novel technologies to attitudes and attitude change. She is co-author of various textbooks on social robots, among them “HRI: An introduction (2020, Cambridge University Press), “Robots in Education” (2021, Routledge), and “Theory and practice of sociosensitive and socioactive systems” (2022, Springer).

Alessandro Lanteri is Full Professor of Strategy and Innovation at ESCP Business School, Turin campus. He teaches and directs executive education programs at some of the world's leading business schools. An expert educator, he helps executives and students understand emerging technologies and seize the opportunities of turbulent environments while developing sustainable value propositions. Alessandro works with multinationals, governments, International Organisations, startups and family businesses across five continents. He holds a PhD in Philosophy from Erasmus University Rotterdam, a Master in Economics from Bocconi and also studied at Harvard, MIT, NYU, Oxford, and Stanford. His research regularly appears

in top international publications both academic (Journal of Business Ethics, Philosophical Studies) and professional (Harvard Business Review and MIT Technology Review outlets). His previous book “CLEVER. The Six Strategic Drivers for the Fourth Industrial Revolution” became an international no.1 Amazon bestseller. His latest “Innovating with Impact” is published by The Economist.

Core Publication #6 [CP6]

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021),
“Saving Private Robot: Risks and Advantages of Anthropomorphism
in Soldier-Agent Teams”
International Journal of Social Robotics, 14(1), Springer Verlag.

Reproduced with permission from Springer Verlag.
All rights reserved.



Saving Private Robot: Risks and Advantages of Anthropomorphism in Agent-Soldier Teams

Massimiliano L. Cappuccio^{1,2} · Jai C. Galliot¹ · Eduardo B. Sandoval³

Accepted: 21 January 2021

© The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021

Abstract

Hybrid military teams, formed by human warfighters and autonomous artificial agents, represent the technological future of Defence operations. Both the potential and the inherent limitations of current technology are well-known, but the cognitive-behavioral and motivational aspects of human-robot interaction on the battlefield have yet to be systematically investigated. To lay the theoretical and methodological foundation of this scientific investigation, our position paper critically examines how the military personnel's spontaneous tendency to anthropomorphize artificial autonomous agents can affect operations of hybrid military teams in multiple ways. We will argue that the psychological impact of anthropomorphism on military personnel is neither easily avoidable nor necessarily detrimental. Correctly identifying the multi-level cognitive mechanisms that underpin implicit and explicit forms of anthropomorphism allows us to increase the efficacy of human-agent interaction. We will argue that, within hybrid teams, the capability to communicate with teammates, allies, civilians, and adversaries relies on embodied social cognition processes that are inherently geared toward anthropomorphism and leverage its effects. By updating both the design of autonomous artificial agents and the training of human troops to account for these processes, their reciprocal coordination can be augmented.

Keywords Autonomous artificial agents · Autonomy · Hybrid military teams · Human-agent interaction · Social cognition · Social robotics · Social psychology · Situational awareness · Military drone · Autonomous vehicle · Value-sensitive design

1 Introduction

In future warfare, tactical and operational superiority will importantly depend on the armed forces' capability to effectively deploy hybrid units formed by military personnel and *autonomous artificial agents* (hereafter called simply "Agents") [1–5]. The set of systems indicated as Agents,

also collectively known as "autonomy" [6, 7], is more comprehensive than the domain of mobile, physical, interactive robots (most often android or animaloid in design) typically considered by social robotics [8]. Military Agents comprise a greater variety of synthetic entities controlled by artificial intelligence [9]: unmanned underwater vehicles (UUV); unmanned aerial vehicles (UAV, Fig. 4), including various kinds of drones; unmanned ground vehicles (UGV, Figs. 1, 2), including unmanned quadruped ground vehicles (Q-UGV, Fig. 3); interactive embodied robots; digital avatars.

Military Agents and social robots share some important features as both of them are expected to successfully conduct collaborative activities within teams of humans. However, while engaging with humans is the primary goal of social robots, military Agents are equipped with interactive and communicative features only insofar as they are instrumental to successfully completing teamwork tasks via effective coordination with humans. Most military Agents need only very simple socially interactive features to achieve this goal. That is why they are conceptually

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

Jai C. Galliot
j.galliot@unsw.edu.au

Eduardo B. Sandoval
e.sandoval@unsw.edu.au

¹ Values in Defence and Security Technology Group, School of Engineering and Information Technology, University of New South Wales, Kensington, Australia

² AI & Robotics Lab, United Arab Emirates University, Abu Dhabi, UAE

³ Creative Robotics Lab, University of New South Wales, Kensington, Australia



Fig. 1 The Modular Advanced Armed Robotic System (MAARS®) is an unmanned ground vehicle (UGV) developed by Qinetiq and designed for reconnaissance, surveillance, and target acquisition (RSTA) missions to increase the security of personnel manning forward locations

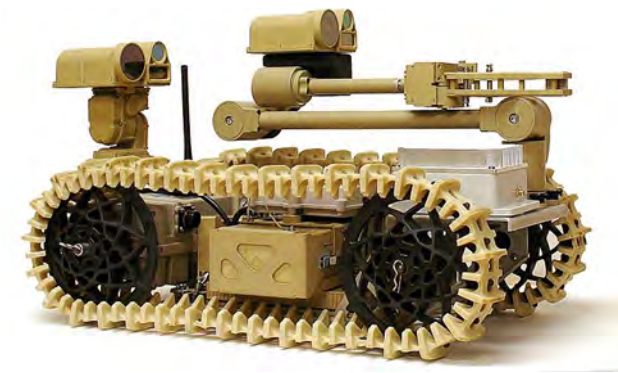


Fig. 2 The Increment 1 Platform is an unmanned ground vehicle (UGV) prototype developed by Northrop Grumman Corporation and U.S. Navy as part of the Advanced Explosive Ordnance Disposal Robotic System (AEODRS) program. This vehicle can operate autonomously and is designed to carry out tasks like the disposal of explosive ordnance and improvised explosive devices and the handling of hazardous materials

similar to “CoBots”, i.e. new-generation industrial robots equipped with basic social features sufficient to adaptively carry out joint tasks (like collaborative montage) in cooperation with human operators [10].

Social robotics research can tell military scientists something important about Agents because, more than any other area of human-agent interaction (HAI), it has documented that the quality and the success of the collaborations between humans and robots importantly depends on how the former perceive the latter and relate to them [11, 12]. Social robotics systematically studies the socio-psychological precondition of successful HAI and, taking advantage of how Agents are generally perceived by



Fig. 3 The Ghost Vision 60 is a quadrupedal unmanned ground vehicle (Q-UGV) developed by Ghost robotics to carry out tasks such as remote inspection, intelligence, surveillance, reconnaissance, mapping, distributed communications, and persistent security



Fig. 4 Boeing Loyal Wingman is a low-cost stealth unmanned aerial vehicle (UAV) developed by Boeing Australia to support manned aircrafts such as the F-35 and perform autonomous missions using artificial intelligence. A squadron of Loyal Wingmen can be controlled by a parent aircraft to accomplish tasks such as scouting or absorbing enemy fire if attacked

humans, aims to advance the civilian applications of HAI through improved design solutions [13].

Both Agents and social robots must produce behaviours that appear purposeful and meaningful, as they are expected to communicate to their human teammates the goals and the reasons of their decisions [14]. Conversely, in order to collaborate effectively with humans, both Agents and social robots must be capable to correctly identify the intentions and the drives that motivate human behaviour [15, 16]. This two-way transparency is essential to establish reliable, agile, and intuitive coordination among military personnel and Agents, which is vital for the satisfactory functioning of hybrid military teams in high-tempo, high-stakes, risky scenarios [1, 17].

Socio-psychological attitudes toward autonomy (e.g., acceptance, tolerance, perceived intelligence, etc.) affect

trust in Agents and impact on the armed forces personnel's ability to establish rich, fluid coordination with them [6, 12, 18]. As confirmed by numerous studies in HAI research, attitudes towards autonomy, including importantly trust, are modulated by a fundamental phenomenon called *anthropomorphism*, i.e. the human tendency to perceive various kinds of non-human agents, including machines, as human-like [19]. It is well documented that anthropomorphic tendencies emerge spontaneously and pervasively in HAI [20].

Thus, anthropomorphism is a communal interest for military and social roboticists and motivates them to learn from each other's work [21]. On the one hand, Defence and security operations conducted by Agent-soldier teams have offered to social roboticists some of the most interesting cases for the study of anthropomorphism [22]. On the other hand, the civilian applications of social robotics offer to military roboticists useful methods and notions to understand how hybrid teams can be perfected. One issue discussed by the literature is whether or not the humans' tendency to anthropomorphize their artificial teammates can damage the work of hybrid teams: in particular, some scholars claim that unexpected anthropomorphic projections could expose military personnel working in hybrid teams to severe, possible lethal, risks [4, 22]; others, underestimating the potential impact of anthropomorphism on the performance of human operators, minimize its functional implications [17].

What is the right approach toward anthropomorphism in the context of military robotics? This question has been repeatedly addressed as part of the ethical dispute concerning whether autonomous weapon systems (AWS) should be banned or rather adopted by the military forces in order to make warfare less risky and inhumane [18, 21, 23–25]. Significantly, both the opponents and the advocates of AWS have blamed sensationalist media for propagating excessively anthropomorphic representations of military Agents, indulgently attributing to AWS human-like capabilities and intentions [21]. Even more interesting, each camp in this debate has criticized the other for having allegedly bought into mythological narratives fueled by some anthropomorphic misrepresentation of Agents: according to the opponents of AWS, the advocates have a misplaced, unmotivated trust in the ability of “killer robots” to apply lethal force proportionately and discriminately because, in subsuming an anthropomorphic representation of the military Agents' intelligence, they have developed exaggerated expectations about the capabilities and trustworthiness of AWS [24]; vice versa, according to the advocates of AWS, the opponents have been deliberately using the fantastic iconography of “killer robots” movie genre to monger unjustified fears over military Agents, representing autonomous technology as a semi-sentient force animated by subversive intentions and an uncontrollable thirst for destruction [23].

In this position paper we will not further discuss the debate on AWS, but we intend to demonstrate that anthropomorphizing Agents does not necessarily mean buying into misleading representations of them, as the two camps seem to take for granted. On the contrary, we will argue that correctly integrating considerations about anthropomorphism and zoomorphism into our scientific methodologies is indispensable to increase our understanding of HAI and improve the design of Agents. We propose first of all to investigate why and how anthropomorphism affects HAI in military teams to identify patterns of opportunities and risks, neither ignoring nor fearing this psychological phenomenon.

In Sect. 1, we review the progress recently made by the Armed Forces in development and testing, recognising the importance of anthropomorphic and zoomorphic design elements in the new Agents. In Sect. 2, we provide a general characterization of anthropomorphism that distinguishes between implicit and explicit attitudes, briefly reviewing a number of relevant empirical studies in social psychology. In Sect. 3, we assess whether anthropomorphism has produced significant adverse effects during the deployment of Agent-soldier teams, suggesting that the risks associated with involuntary anthropomorphism have probably been overemphasized. In particular, distancing ourselves from both the “lax” and the “anxious” approaches to anthropomorphism that we found in the literature, we propose an alternative, pragmatic approach: we argue that designing military robots informed by a correct understanding of anthropomorphism is strategically more convenient than trying to ignore or suppress this phenomenon. In Sect. 4 we will argue that, while monitoring the anthropomorphic tendencies that arise among armed forces personnel is important to prevent unwanted consequences, entirely subduing such tendencies is neither possible, due to their pervasiveness and resilience, nor convenient, due to the inherent advantages that they provide for the effective coordination of soldiers and Agents in tactical scenarios. In Sect. 5, we illustrate these advantages delineating four prototypical tactical scenarios in which anthropomorphism would be critical to establish effective coordination and communication between humans and Agents. The Conclusion summarizes how our pragmatic approach, by promoting a correct scientific attitude toward anthropomorphism, can be beneficial to the design and deployment of future military applications of autonomy.

2 Defence Applications of HAI

The Defence Forces of the United Kingdom and of Australia have recently conducted a large-scale exercise as part of a concerted effort undertaken by the ‘Five Eyes’ (FVEYs) intelligence community. This exercise, called “Autonomous Warrior”, was carried out in Jervis Bay, New South Wales,

in June–July 2018 and in Salisbury Plain, south of England, in January 2019 [26].

Unclassified information about these exercises confirms the crucial importance attributed by military commands to hybrid teams [26]. Apart from combat (Figs. 1, 4), numerous security and Defence-related joint tasks can be assigned to Agents or conducted with their support: reconnaissance, intelligence collection surveillance and perimeter protection (Fig. 3); infiltration and loitering for spying and sabotage; active shielding of military personnel and civilians (Figs. 1, 4); disposal and handling of bombs and chemical, biological and nuclear hazards (Fig. 2); and the provision of medical and psychological assistance to military personnel, including training, team building, counselling, and conflict mediation [2, 22].

A fluid integration of human and artificial workflows is paramount to successfully carry out such tasks [27]. Agents must be equipped with interactive protocols and basic pseudo-social abilities calibrated to the specificities of human teammates, including—inter alia—better methods for identifying humans (e.g., face and voice recognition), interpreting their intentions, and anticipating their actions [2, 9–12, 14–16, 18, 20, 23, 27–44].

In a public interview [26], Jason Scholz, Defence Science & Technology’s Autonomous Warrior 2018 leader and chair of the TTCP (The Technology Co-operation Programme within the FVEYS Community) *Autonomy Strategic Challenge*, reveals that many autonomous systems were evaluated, including “simultaneous uninhabited vehicle assets in the air, land, sea-surface and sea-subsurface domains”. These included, “radar-equipped drones for detecting buried IEDs” (improvised explosive devices), “small two-wheeled robots that [can be] thrown into buildings to search for enemy fighters”, “remote-controlled Warrior infantry fighting vehicle”, and “a lightweight MRZR tactical vehicle”. The tasks performed to assess the interoperability of these devices included “Counter Submarine Incursion, Sea Defence, Counter Piracy, Counter Arms Smuggling, Oil Terminal Blockade by Mining, Oil Terminal Infrastructure Attack, Base Attack, Base Defence, Support to Ground Forces, Containing Civil Unrest, Clandestine Insertion from Sea”. In the exercise subsequently performed by the British Armed forces, more than fifty unmanned systems from industry were deployed to complete similar tasks [34].

Admittedly, although some of the challenges encountered by hybrid teams were technical in nature (e.g., bandwidth management and minimization of reciprocal interferences among Agents), the toughest challenges concerned not the sheer volume and the robustness of data transmission, but the quality of human–machine interfaces. To help their human teammates successfully navigate the so-called “fog of war”, Agents must be able to establish natural, fluent, semantically rich, and effortless communication with them.

That is why the Australian armed forces created and tested a software to advise the units involved in complex operations and hence support “comprehension (assessment of situations) and projection (assessment of threats)” to facilitate reciprocal coordination and strengthen situation awareness. The system was designed to provide “contextual information to aid in user comprehension, and it adapts as the user interacts with the system”, a kind of aid specifically “modelled on a news service paradigm using virtual human characters” [26].

Most interesting for our analyses, the software relies on humanoid digital avatars that speak and act in a simulated environment. With the help of the avatar’s unique expressive features, a real-time speech narrative is generated and integrated with relevant text, graphs, videos, and other contents to help “explain the story” [26]. Remarkably, instead of representing information in an impersonal and neutral format through text and graphs only, the designers have decided to use an anthropomorphic companion. Apparently not worried that it could represent a distraction or cognitive burden, military researchers have deliberately decided to rely on the avatar’s embodied and narrative features, emphasizing its role of teammate and informer [45]. As we will see in the next section, experimental data about anthropomorphism supports this choice.

3 The Cognitive Roots of Anthropomorphism

Anthropomorphism can affect the warfighter’s attentional, behavioural, and emotional responses during hybrid team operations. Anthropomorphism is typically defined as the tendency to attribute to non-human (possibly synthetic and inanimate) entities the characteristics of human ones, transferring to the former psychological attitudes and forms of engagement that are typically or exclusively reserved to the latter [19]. The cognitive mechanics underpinning anthropomorphism are analogous to those of zoomorphism (i.e., the tendency to attribute to inanimate entities the features of animals or animal-like entities), and largely overlap with them [2–5, 11, 12, 14–16, 18, 23–27, 36–60]. Among the human-like characteristics that can be projected onto non-human entities we can distinguish simple cognitive features, that humans arguably share with non-human animals (purposeful, adaptive behaviour and the capability to selectively sense and track contextually relevant cues) and more cognitively complex features involving advanced declarative and linguistic skills, such as contentful knowledge states (beliefs), objectives (desires), and plans to achieve them (intentions), that are traditionally considered exclusively human.

According to some models [34, 35], anthropomorphism depends on three psychological and motivational determinants: (1) *elicited agent knowledge* (i.e., the accessibility and applicability of anthropocentric knowledge); (2) *effectance motivation* (i.e., the motivation to explain and understand the behaviour of other agents); (3) *sociality motivation* (i.e., the desire for social contact and affiliation). Applied to hybrid teams, this theory predicts that military personnel are more likely to anthropomorphize Agents when they are aware of features and functions that justify human–machine analogies, when their survival depends on team cohesion and solidarity among the team members (including possibly Agents), and when they need the presence of a trusted companion to feel less isolated and alone.

Anthropomorphism has both an objective component (*anthropomorphic* design features and structural functions) and a subjective one (*anthropomorphic* attributions via psychological projection). Drawing a clear line among them is not always easy. Social robotics research has highlighted that anthropomorphic attributions can be solicited by certain android design features, but do not strictly covary with them [61]. In turn, anthropomorphic attributions positively correlate to a higher level of perceived intelligence, which, in turn, predicts more robust interest and attention in the users [62]. However, this effect does not consistently correlate to a greater likeability or acceptance: frequently (for example, but not exclusively, when the humanoid appearance is hyper-realistic) the social robots that produce the strongest anthropomorphic attributions also tend to generate the greatest discomfort, mistrust or even repulsion in the human interactants (as documented by the famous ‘uncanny valley’ scenario [63]). Certainly, anthropomorphism modulates the quality and the intensity of HAI, but all of its implications are not obvious: while it is true that some of the most likeable and best accepted robots tend to be strongly anthropomorphized, there is no evidence that a high level of anthropomorphism is per se sufficient to predict better interaction.

One aspect not to be overlooked is that the general attitude toward an anthropomorphic Agent (including the associated responses in terms of likeability and acceptance) importantly depends on the user’s biases toward autonomy and toward technological artefacts, which typically have a cultural origin [31]. For example, it has been remarked that science fiction and popular media often portray Agents with exaggerated anthropomorphic features, which tends to create unrealistic expectations in both users and designers [52], including possibly the military ones. Such expectations are inevitably frustrated when the interaction with Agents proves simple and unsophisticated, damaging the users’ trust in autonomous technologies.

Despite significant individual differences, which may or may not be caused by cultural factors, personal beliefs, and

social representations [64], the general cognitive/behavioural mechanisms underpinning the anthropomorphic experience are seemingly universal, very deep, developmentally early, and most probably innate [20]. Humans show an instinctive tendency to attribute animacy and intentions even to entities that have little or no resemblance at all to animated or living creatures [63]. The fact that anthropomorphism is hardwired into human cognition is compatible with the conjecture that it is strongly tied to fundamental skills in social and cultural cognition, such as empathy, imitation, and perspective-taking [65]. This explains why, for example, various cognitive biases and delusions correlate with anthropomorphic attributions and personifications [29].

Anthropomorphic attribution is not always deliberate and conscious: it is considered either *explicit* when they can be expressed verbally, or *implicit* when anthropomorphic attributions are manifested exclusively by the subject’s behaviour [66]. The former relies on cognitively demanding mechanisms of judgement and deliberation while the latter operates automatically, at the level of fast, embodied responses that the agent might not be mindful of, while the latter operates via conscious control. Explicit anthropomorphism typically arises from a deliberate search for human-like features (e.g., the intentional personification of nature occurring as part of the figurative process of poetic creation). In turn, implicit anthropomorphism is enacted by habitual sensorimotor dispositions and subconscious perceptual preferences. In this respect, anthropomorphism is very similar to empathy, a basic, pre-reflective form of embodied understanding of the others’ intents and emotions. Arguably, both empathy and anthropomorphism derive from the spontaneous tendency to read organised and repetitive patterns of movements in terms of goal-oriented intentional actions [53].

Explicit and implicit attributions are not necessarily consistent to one another: on the one hand, it is possible to anthropomorphize a certain robot at the perceptual level (for example when the robot’s movement conveys animal-like features) while holding the belief that robots are just unintelligent inanimate entities; on the other hand, one can deem a software Agent intelligent and autonomous while feeling completely indifferent to its expressive behaviours.

The implicit/explicit distinction suggests that anthropomorphism is underpinned not by a single mechanism but a multiplicity of cognitive processes that work in parallel and yet are deeply integrate, semi-independent but not encapsulated. More empirical studies are needed to carefully investigate this cognitive architecture if we want to correctly make sense of our complex, and apparently contradictory, relationship with Agents. This investigation begins with the phenomenology of HAI in the context of Agent-soldier teams, which we will review in the following section.

4 Attachment for Robots in Dangerous Missions?

Not only casual users but also highly qualified personnel can develop a psychological bond with robots. This happens more frequently when their work together is long and stressful [50]. In at least a few documented cases, military personnel have developed not only generic anthropomorphic responses but also mild forms of attachment for the robots that assisted them in their missions, even though said robots look barely human-like or animal-like (Fig. 3). Explosive Ordnance Disposal (EOD) operators deployed in Iraq and Afghanistan report at least four remarkable anecdotes about their relationship with the robots that assisted them in carrying out their task in active war scenarios (i.e., UGVs similar to the Increment 1 Platform, Fig. 2), which testify the pervasiveness of anthropomorphism [3, 4, 22, 24, 40]:

- (1) EOD robots were assigned names and gendered identities by the soldiers who worked with them in Iraq and Afghanistan;
- (2) at times, when one of these robots was damaged, its loss was not simply experienced as the destruction of an expensive piece of equipment, but also *grieved* like the death of a teammate, and in some cases it was accompanied by funeral-like rituals;
- (3) when one of these robots was sent to the headquarters for reparation after suffering structural damage, its human mates requested that its mechanical parts were not replaced, but accurately fixed to preserve the robot's individual identity;
- (4) in rare occasions, soldiers have endangered themselves to protect the robot from enemy assaults.

This relationship recalls similar cases of attachment or empathic transfer often encountered in civilian robotics [67]. For example, the members of the Mars Exploration Rover Team report that they unexpectedly started to identify themselves with the robot that they were tele-operating from a distance of millions of miles [68]. Arguably, empathic transfer and anthropomorphic projection tend to be more vigorous and have a deeper moral impact when they arise in the symbolic space of military values, which is distinctively characterized by the notion of individual sacrifice for a collective purpose [69]. Empathic attitudes toward the robot (which may or may not include a sense of attachment for the robot) go hand in hand with narratives picturing the robot's destruction as a sacrifice meant to save human lives. Therefore, anthropomorphic attributions can be fuelled by both empathy and narratives of sacrifice, which reinforce each another's effects: thus, if

robots are portrayed as entities that “sacrifice” themselves, then their destruction feels like a “death”, which in turn reinforces the empathic perception of them as person-like entities.

Since robots are obviously not sentient and cannot reciprocate affection, the attachment developed by military personnel to their robots has been typically deemed useless, unmotivated, and irrational. As such, the anthropomorphising of EODs seems to contradict the notion that military personnel professionally trained for operating Agents are rational agents, perfectly aware of the limitations of their artificial partners, and well-prepared to act responsibly in accord with strict safety protocols.

In the literature on Agent-soldier teams we can recognize two different, and somewhat opposite, attempts to make sense of this apparent contradiction. The first attempt, call it the “Lax approach” to anthropomorphism, is epitomized by Barnes and Evans [17]. This approach sees anthropomorphism as a by-product of social responses mediated by linguistic and symbolic representations. Originating from such subconscious “metaphoric” projections, attachment for robots is explained as the unwanted effect of overthinking, an unfortunate inconvenience produced by intellectual hyper-stimulation. This approach relies on explanations of anthropomorphism as an imaginative process [43, 54, 58, 65]. While the Lax Approach does not explicitly deny that the anxiogenic context that *causes* over-imaginative production can be detrimental to team performance, it fails to acknowledge the significance of the cognitive *effects* of anthropomorphism on social understanding and intersubjective coordination. Hence, it does not offer recommendations usable in the military context because it underestimates both positive and negative implications of anthropomorphism for HAI.

The second approach to anthropomorphism, call it “the Anxious Approach”, is epitomized by Carpenter [22], who has hypothesized that anthropomorphic tendencies predict attachment for Agents, an unwanted psychological effect that inevitably compromises the decisional capabilities of military personnel, conducing them to risky behaviours. For example, when soldiers, feeling subconsciously compelled to protect the Agents, put at risk both their own safety and their mission's outcome. Consequently, Carpenter recommends caution in designing EOD robots (especially now that developers are focusing on biped models) and suggests diminishing the anthropomorphic features that might trigger unwanted attachment in human users.

On the one hand, these two approaches seem antithetic: while the Lax Approach imputes anthropomorphic attributions to an excess of cognitive activity (imagination), the Anxious Approach imputes them to a lack thereof (judgment). On the other hand, both approaches reduce anthropomorphism to same kind of delusional perception that (they

allege) originates from the failure of rational evaluation capabilities and a subsequent misdirection of the affects to an inappropriate object. Both the Lax and the Anxious approaches presuppose a dualism of the human mind: either humans operate rationally and deliberately (and hence efficiently) or emotionally and subconsciously (and hence inefficiently), as if intellect and sensation were entirely opposed and mutually exclusive mechanisms and effective decision making was possible only thanks to the former (Table 1).

We find this dualistic assumption debatable: first, because in many real-life scenarios the intellect/sensation separation is not clear-cut and seems purely arbitrary [70]; second, because such an assumption fails to acknowledge that explicit processes of rational judgement often supervene on implicit dynamics of aesthetic evaluation and expert perceptual discrimination that operate below the threshold of conscious, deliberate cognition [5]. Specifically, anthropomorphism can be construed as a form of expertise that, allowing the perceptual recognition of intentional agents, subserves social cognition and practical rationality in social interaction.

It is because they ignore this complex dependency relationship that, in our opinion, both the Lax and the Anxious approach misjudge the risks deriving from anthropomorphism: such risks are underemphasized by the former approach while being overestimated by the latter. Thus, both approaches ignore that anthropomorphism can positively contribute to the cognitive life of an agent informing their concrete, evaluative skills [71]. Both approaches fail to recognize that anthropomorphism is a constitutive component of HAI, and not a mere accident or by-product, because they both ignore that embodied social cognition abilities—including the ability to empathise and anthropomorphise—constitute a fundamental requirement for collaborative decision-making as they enable mutual adaptation during intersubjective coordination [72]. Anthropomorphism is not just a cognitive bias, but—as we have seen, similarly to empathy—it is also an embodied form of pre-reflective, practical knowledge that enriches our perceptual expertise, emotional engagement, and sensorimotor interaction [38].

We propose a third approach to integrate this idea, which was ignored by the other two. We call it the “Pragmatic Approach”. According to it, anthropomorphizing military robots is neither a failure of the soldiers’ capability of rational evaluation nor simply a subconscious burst of “metaphorical” imagination: the anthropomorphic perception of robots does not obstruct rational judgement because it operates pre-reflectively at a more fundamental, implicit level of cognitive processing, i.e. the level of embodied, embedded, and emotional perception. Embodied perception is more basic and primordial than inference-based judgment and rational deliberation: it does not directly compete with the higher-level forms of cognition but rather operates in

parallel with them supporting them. In fact, embodied and perceptual responses (including anthropomorphic attributions) do not necessarily inhibit situated, adaptive intelligence: on the contrary, by enabling empathy as a form of situated and embodied understanding of the others, embodied perception contributes to articulate our social and emotional intelligence, providing the kind of direct engagement that is needed, for example, by a soldier to coordinate effectively with teammates during field operations [16, 49].

Military professionals can and often do subconsciously attribute anthropomorphic features to Agents (with the possible attachment feelings that may or may not derive from this projection), even if they perfectly know, at the conscious level, that robots and AIs are just mechanical devices devoid of emotional states, proper motivations, and purpose. We propose to understand the conscious (explicit) and the subconscious (implicit) levels of HRI not as two independent systems, but as two moments of the same integrated process subserving social cognition. When we subconsciously attribute anthropomorphic attributes to an Agent, this process is neither intrinsically irrational nor the delusional by-product of hyper-stimulated imagination: like empathy, anthropomorphism is rather the perceptual, experiential medium through which intersubjective coordination and embodied communication articulate themselves [49].

Therefore, while the risks associated with a dysfunctional HAI are not to be underestimated, evidence suggesting that anthropomorphism leads to unsafe behaviours is scarce and merely anecdotal. It certainly does not justify a general ban on anthropomorphic design: first, because, as demonstrated by the uncanny valley example [63], anthropomorphic attributions do not necessarily predict strong attachment to robots (although it is true that attachment to robots implies anthropomorphic attributions); second, while anthropomorphism is a process rooted in perception and embodied interaction with real agents, we don’t have reasons to think that occasional attachment for robots is necessarily robust like the bond between teammates (compassion for robot can be authentic but it could also, in some cases, be of a fictional or confabulatory nature, like “fantasies” or day dreaming); thirdly, even if anthropomorphism was systematically accompanied by robust attachment, its detrimental effects could still be easily countered by appropriately training the EOD operators to monitor and regulate their own emotions. Humans (including soldiers) prepared to recognize their spontaneous anthropomorphic tendencies can easily ignore or control them as the circumstances demand [73].

Our approach provides reasons to believe that the practical advantages offered by a responsible implementation of anthropomorphic features in a military robot’s design largely surpasses the potential risks. Anthropomorphic effects are carefully accounted for and actively exploited by social roboticists to produce positive psychological effects on the

Table 1 Summary of the three approaches toward anthropomorphism

Approaches to anthropomorphism/defining features	The 'Lax' approach	The 'Anxious' approach	The 'Pragmatic' approach
Key examples in the literature on military Agents	Barnes and Evans [17]	Carpenter [22]	This paper
Psychological processes underlying anthropomorphic attributions	Conscious level: imaginative processes with semantic (verbal and/or symbolic) content	Subconscious level: emotional (irrational, instinctual, and likely non-verbal) processes	Multi-level: it includes both empathic (perceptual and embodied, analogic) and cultural (symbolic, verbal) processes (e.g., narratives of sacrifice)
Outcome	Metaphoric projections, symbolic imagery	Development of attachment feelings	May vary significantly. Social-interactive infrastructure for communication and coordination
Problem with anthropomorphism	An excess of cognitive activity: overthinking, overimagining	A lack of cognitive activity: absence of rational judgment	A constitutive component of cognitive activity: it enables judgment and recognition
Epistemic value of anthropomorphic attributions:	A-rational (no connection between perception and judgment)	Irrational (open conflict between perception/feeling and judgment)	Pre-rational (perception/feeling can scaffold judgment and shape its content)
Underlying cognitive architecture	Cartesian model of cognition: dualism of perception and judgement	Cartesian model of cognition: dualism of perception and judgement	Embodied model of cognition: integration of perception and judgement
Practical value of anthropomorphic attributions	They are neither risky nor beneficial	They expose tactical operations to significant risks	They are indispensable to support communication and coordination in HAI. They have only very modest risks
Recommendation for designers and developers to improve HAI	Indifferent to anthropomorphic design elements	Avoid or minimize anthropomorphic design elements	Empirically investigate and selectively implement the advantageous aspects of anthropomorphic design

users (e.g., for domestic, clinical, and educational tasks [30]) and by industrial designers for marketing and branding purposes [32, 74]. Although important differences between civilian and military applications of interactive robots should not be overlooked [12, 15], considering the concrete advantages of anthropomorphism for HAI in military hybrid teams is more promising than looking only at risks whose nature seems largely speculative. In fact, we will concretely illustrate in the next section, anthropomorphic tendencies can be deliberately leveraged to improve human–machine interfaces, creating more reliable and efficient communication inside and outside the hybrid team.

5 Anthropomorphism as a Usability Advantage for Military HAI

HAI research studies intuitive communication methods to control autonomous and semi-autonomous agents, including embodied cues like hand gestures [48], gaze direction (to establish joint attention and facilitate natural action coordination), emotional expressions (through voice or facial expressions), turn-taking and rhythmic synchronization [42].

Such methods offer significant operational advantages in hybrid teams [7]: intuitive control does not require long and demanding training and frees up cognitive and physical resources (working memory, attention and bodily engagement) that the human agent should preferably be free to allocate to parallel tasks [75], for example eye-contact-based communication or bimanual manipulative tasks (like operating a computer).

Cognitive parsimony is crucial in tactical contexts involving an emotionally charged, high-risk, high-tempo component mediated by embodied (perceptual, motoric) engagement [76]. These scenarios prioritize agency over critical analysis and planning, require quick decisions of a practical nature, readiness to generate short-term predictions, and the ability to fluidly transition from one context of action to another.

Anthropomorphic attributions are most likely to impact on the tactical operations that involve value-laden decisions, when the warfighter must rely on reflexive responses developed through training and practice to make fast and effective decisions in uncertain circumstances. Acting and thinking to neutralize possible threats, acquire targets, and overcome obstacles requires quick and trusted coordination between teammates, leveraging on a spontaneous, intuitive two-way signalling activity. Coordination among professional operatives in these circumstances involves reciprocal reading of behavioural and emotional expressions (pain, fear, impatience, intention to maintain one's position or flee, disposition to open fire or wait) that is done automatically as part of a natural set of empathic

abilities that homo sapiens shares with their conspecifics and, to a minor degree, with phylogenetically related species [41].

Soldiers do not need to be taught these informal abilities, although appropriate training can help effectively channel one's embodied know-how to prepare for specific tactical scenarios. When such operations are carried out by hybrid-teams comprised of humans and Agents, anthropomorphism is an intervening factor to be considered carefully because it modulates the empathic communication that is one of the key cognitive components of joint decision-making.

Anthropomorphism theory explains, among other things, how inanimate objects, with their mechanical behaviour, can be part of such natural and fluid coordination, replicating the main features of embodied communication: through automatic anthropomorphic projection mechanisms, the physical and symbolic properties of these objects can evoke the intentional actions and emotional states of living agents [3–5, 14, 16, 24–26, 43, 44, 46–58].

A remarkable case of anthropomorphism is when a particular car driving style conveys information about the driver's intentions, goals, and emotional states [57]. Expert drivers can promptly read another driver's intentions through their car's behaviour, predicting with high accuracy and no apparent effort whether a car is going to *arrogantly* overtake on a highway, or *respectfully* giving us way at a junction, or *patiently* wait that another car will leave a parking to occupy it. When such transparency is established the perceptual features displayed by the car are anthropomorphised.

Notably, knowing that the car does not have intentions of its own does not prevent us from seeing the driver's intention displayed in the very car's movement [24]. As anthropomorphic attribution is automatic and subconscious the observer does not necessarily know how it works, and yet may be able to perfect it and make it more efficacious with appropriate training and expertise development.

Anthropomorphic projection applies to vehicles operated by humans and unmanned vehicle alike. By means of machine learning, the behaviour of UGVs is very likely to capture the fine-grained patterns of human driving styles [47]. This anthropomorphic information (i.e., perceptual information evoking intentions, goals, and emotional states), allows the observer to quickly adjust to the vehicle's behaviour and coordinate with it, while its absence could mislead a human observer habituated to rely on it: confusion could arise if autonomous vehicles were programmed to filter out any emotionally or intentionally-loaded pattern of behaviour. In the military context, this means that the poverty of anthropomorphic information expressed by unmanned vehicles could disrupt the overall performance of the hybrid team [57] because anthropomorphic information does not have only an aesthetic and cosmetic function but plays a functional role in HAI.

Civilian studies on the anthropomorphic features of small social drones have immediate implications to improve the interactions between military personnel and flying drones of varying size and combat role and, indirectly, for other unmanned military vehicles, especially those animaloid in design, such as the Q-UGVs (e.g., the Vision 60 dog-like robot by Ghost Robotics, see Fig. 3). Experimenters have programmed the flying patterns of a small drone to convey information about its “emotional state” and “personality” [77]. The emotions that have been conceived and designed (corresponding to personality stereotypes) include: brave (personality stereotype name: “the Big Boss”), dopey/sleepy (“the goofy comedian”), grumpy (“the detached philosopher”), happy (“the lovable/romantic”), sad/melancholic (“the peaceful artist”), scared (“the sneaky spy”), and shy (“the model student”). The behavioural patterns that express these emotions have been generated calibrating different dimensions that characterize the drone’s movement (slow/fast; stepwise/fluent; instant/delayed; uniform/diverging; constant/inconstant; spatial separation/spatial proximity; approximate/precise; gentle/powerful) and the terms of an “interaction vocabulary” that define the parameters of the drone’s behaviour in relation to the user (the drone’s position and direction compared to the user, speed, rotation angles, altitude, reaction time and compliance to commands).

These emotions map onto five *personality traits* (openness to experience, conscientiousness, agreeableness, extraversion, neuroticism) which were manipulated by the experimenters to define three *personality profiles*, which express complex emotional patterns and correspond to different emotional approaches to the task: the Exhausted Drone (merging dopey, sleepy, and sad), the Anti-Social Drone (merging grumpy and shy), the Adventurer Hero Drone (happy and brave), and the Sneaky Spy Drone (focusing on scared only). The first drone prototype flies with a “drowsy stile” to indicate that the drone’s batteries are nearing depletion; the second flies fast and distant from operators, to indicate unwillingness to engage in interaction with humans; the third flies with high speed, fast reactions, and quick altitude increase to indicate the drone’s readiness to undertake dynamic tasks with determination and resoluteness, communicating a sense of “excitement and bravery” to the observers; the fourth flies slowly and with circumspection (“looks around for danger” and “keeps its distance”).

In the post-experiment survey, participants consistently attributed to drones the emotional and behavioural features that they were meant to evoke, confirming the designers’ expectations. This proves that the device’s anthropomorphic features reliably convey the intended information about its state, function, current task, and mode of operation.

Recognizing and controlling the way military personnel subconsciously project animacy, intentionality, agency, and purposiveness on Agents is not just a stimulus to enforce

discipline, but a tactical prerequisite to make the hybrid team successful as a whole. Our Pragmatic Approach (unlike the other two) acknowledges that anthropomorphic attributions can be a cognitively parsimonious, operationally convenient way to improve coordination and natural communication in hybrid teams. In the following section we illustrate four general types of warfare scenarios in which this idea can be effectively applied.

6 Applications of Anthropomorphism in Defence Operations

Let us generalize these considerations to envision four possible roles anthropomorphism could play in warfare: within the same unit (A), among allied units (B) and during interactions with civilians (C) or hostile forces (D).

6.1 Coordination Between Members of the Same Unit

As special forces operators share the same culture, doctrine, training, and practices, their communication is prevalently implicit because it relies on a broad communal background of shared experiences. The reason these units can achieve superior internal coordination and a robust capability to respond to unanticipated threats is that they rely on natural communication based on gaze, gesture, posture and body language to produce fluid and well-coordinated responses.

If Agents are unable to share the same kind of informal communication, then the hybrid team’s functional and psychological cohesion is undermined. Agents should be designed to produce and respond to information conveyed in a bodily, emotional, and purposeful format, and not exclusively explicit commands [37]. Gesture, gaze, mimicry, proxemics, posture, movement style, and tone of voice are just some of the cues that Agents must learn in order to sustain cooperative behaviours and joint action [51]. This would be hardly achievable without factoring in anthropomorphism, which is prerequisite to natural communication among animated, intentional, purposeful beings.

6.2 Coordination Between Multiple Allied Units

Large multi-front operations demand optimal situational awareness [33] in all the units involved, despite the objective difficulties of collecting, filtering and processing all the necessary information due to distance, security, and other hindering factors. Digital avatars that can automatically integrate the relevant information from multiple battle fronts and units are already under development [9, 26]. Their function is to provide to all units a real-time battle report in the form of a narrative dynamically tailored to each recipient

to increase their coordination and mutual alignment. This narrative is meant to be easily accessible without prior training or background knowledge to quickly apprise decision-makers of contextually relevant information to allow them make informed decisions in real-time.

6.3 Coordination Between the Military Unit and Civilians

In peace-keeping operations, hybrid teams are expected to be used at check points and during inspections of households and factories to reduce the direct risk to military personnel and take advantage of superior scanning abilities of Agents equipped with infrared and other beyond-human sensory capabilities. Anthropomorphism is crucial to convey the kind of emotional information necessary to stabilize the relationship between operatives and civilians, especially in situations where relationships are tenuous and trust has waned.

In this context, avoiding misunderstandings and preventing aggressive responses is critical, and words might not be sufficient. The military ought to be concerned with producing both speech and behaviour that are not simply compatible with a local populace's customs and cultural standards to reassure civilians, but also actively encourages their cooperation, while deterring hostile actions.

Agents, no less than human troops, must be designed to communicate correctly with subjects that are untrained in emotionally loaded situations that can easily escalate. Recent applications of social robots as conflict mediators [25, 55] might be useful to achieve this objective. Shape, (un)friendly look, size, behavioural patterns, mimicry, velocity, and mechanical features resembling faces, eyes, or other human or animal body parts can be easily anthropomorphized [32, 73] and therefore indicate a threatening or friendly presence. These elements, and their moral implications, should be carefully considered to design Agents capable to produce acceptance, trust, and respect in the civilian population [56].

6.4 Communication with Adversaries

The anthropomorphic traits of the Agents also make a difference when enemy armed forces confront one another: if appropriately designed, bulky mechanical devices can impress fear upon one's adversary and swarms of drones flying aggressively towards their target can suggest a resolute intent to attack.

Deterrence is not the only possible application of anthropomorphic Agents, which can be designed to appear aggressive or unaggressive, disposed to advance resolutely or cautiously, inclined to assault, flee, negotiate, explore, or wait and observe. As Agents convey anthropomorphic information about intentions and goals, and this information can be either truthful or false, Agents can be designed to produce

specific responses with a deceptive or diversionary purpose, e.g. to distract the adversary and delay its responsive capability. If appropriately displayed, the "personality traits" of Agents can induce the adversary to misinterpret the size, type, and intent of the enemy forces or to miscalculate the time and place of an incoming attack, incautiously positioning their own forces on the field or lowering their defenses. All strategic options available to humans for deception and diversion can be replicated and amplified by means of Agents, extending and magnifying their offensive and defensive capabilities.

Take aerial combat as an example: with the development of new-generation combat drones equipped with advanced autonomous capabilities, human pilots will soon be confronted by UAVs controlled by expert AIs (e.g., Boeing Loyal Wingman, Fig. 4) [3]. In a series of simulations recently run by DARPA as part of their "AlphaDogfight" project, AI-controlled jet fighters have won several simulated matches against experienced F-16 pilots, consistently outperforming the human adversaries [59]. Arguably, whether they are coordinating with allied combat drones or fighting against the enemy ones, human pilots are advantaged when they are able to understand the AI's goals and anticipate its next moves, and this requires they are able to perceive the pilot's intention by "reading" the vehicle's behaviour in anthropomorphic terms. Conversely, in order to outclass the adversary in dogfight, an AI can be designed, or may learn, to confuse or mislead the human adversary by systematically suppressing a great number of tactically significant anthropomorphic cues (dissimulation) or deliberately generating anthropomorphic cues to convey false intentions and goals (simulation).

In all of these four cases, the ability to produce and decode informal communication is crucial to sustain reliable coordination: in warfare involving Agent-soldier interaction, appropriate action and decision requires effective, parsimonious, and timely coordination, which in turn involves anthropomorphic cues.

7 Conclusion

Despite the growing expectations related to Agents, the cognitive-behavioural and motivational aspects of HAI on the battlefield have yet to be appropriately investigated. Potential or merely imagined risks have, to date, received more attention than the actual opportunities.

But the design of Agents for team embedding must also account for the positive psychological implications of anthropomorphism. Therefore, the phenomenology of HAI must be acknowledged by military researchers if they want to be in a position to realistically judge both the possibilities and risks associated with hybrid teamwork.

Inspired by the research on social robots and CoBots, the pro-active integration of anthropomorphic design motives in military Agents will be a necessary step in the roadmap to the development of efficacious hybrid teams. Military research will have to systematically investigate: the distinctive forms of manifestation of anthropomorphism in military HAI; the risks and advantages associated with anthropomorphism within Agent-soldier teams; the impact of anthropomorphism on the interactions between hybrid military teams and external entities (whether combatant or not); the most effective design solutions to maximize advantages and minimize risks in the future construction of android and animaloid Agents.

In brief, this means that Agents design, troops training, and force structuring should be decided in parallel, aiming towards the development of natural interfaces to optimize human-machine communication, as this could significantly reduce the costs and times necessary to make hybrid military teams operation ready. On the one hand, further empirical studies are needed to validate the (mostly anecdotal) claims regarding risks of anthropomorphism in warfare, prioritizing data collection in active scenarios. On the other hand, to prepare both warfighters and support staff to coordinate with Agents, the training programs for military personnel working in hybrid teams must be informed about and produce awareness of the possible advantage of anthropomorphizing Agents. Such awareness is indispensable to develop an appropriate attitude of trust and acceptance for the new technologies.

Funding This material is based upon work supported by the Air Force Office of Scientific Research and the Minerva Initiative under Award Number FA9550-18-1-0039. The work has also received the support of the Australian Government.

References

- Barnes MJ, Jentsch FJ (eds) (2010) Human-robot interactions in future military operations. Ashgate Publishing, Farnham
- Galliot JC (2013) Unmanned systems and war's end: prospects for lasting peace. *Dyn Int* 8:1–24
- Scharre P (2018) Army of none: autonomous weapons and the future of war. WW Norton & Company, New York
- Singer PW (2009) Wired for war: the robotics revolution and conflict in the 21st century. Penguin, London
- Swiderska A, Küster D (2018) Avatars in pain: visible harm enhances mind perception in humans and robots. *Perception*. <https://doi.org/10.1177/0301006618809919>
- Abbass HA, Scholz J, Reid DJ (2018) Foundations of trusted autonomy: an introduction. In: Abbass HA, Scholz J, Reid DJ (eds) Foundations of trusted autonomy. Springer, Berlin
- Chen JY, Anthony C, Selkowitz R, Stowers K, Lakhmani SG, Barnes MJ (2017) Human-autonomy teaming and agent transparency. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction—HRI '17
- Breazeal C, Dautenhahn K, Kanda T (2016) Social robotics. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer International, Cham, pp 1935–1972
- Cummings ML (2017) Artificial intelligence and the future of warfare. International Security Department and US and the Americas Programme, Chatham House—The Royal Institute of International Affairs, January 2017
- Djuric A, Urbanic R, Rickli J (2016) A Framework for collaborative robot (CoBot) integration in advanced manufacturing systems. *SAE Int J Mater Manuf* 9(2):457–464
- Foster Thompson L, Gillan DJ (2010) Social factors in human-robot interaction. In: Barnes M, Jentsch F (eds), Human-robot interactions in future military operations. Routledge, London, pp 67–81.
- Galliot JC (2016) Defending Australia in the digital age: toward full spectrum defence. *Defence Stud* 16:157–175
- Zlotowski JA, Proudfoot D, Yogeewaran K, Bartneck C (2015) Anthropomorphism: opportunities and challenges in human-robot interaction. *Adv Robot Int J Robot Soc Jpn* 7(3):347–360
- Hoffman G (2012) Embodied cognition for autonomous interactive robots. *Top Cogn Sci* 4(4):759–772
- Galliot JC (2012) Uninhabited systems in the civilian realm: some ethical concerns. *IEEE Technol Soc Mag* 31:13–16
- Hoffman G, Breazeal C (2004) Collaboration in human-robot teams. In: AIAA 1st Intelligent systems technical conference. Infotech@Aerospace Conferences, American Institute of Aeronautics and Astronautics
- Barnes M, Evans AW (2010) Soldier-robot teaming: an overview. In: Barnes M, Jentsch F (eds) Human-robot interactions in future military operations. Ashgate Publishing, Farnham
- Galliot JC (2018) The soldier's tolerance for autonomous systems. *Paladyn J Behav Robot* 9(1):124–136
- Zlotowski JA (2015) Understanding anthropomorphism in the interaction between users and robots. Doctoral thesis, University of Canterbury
- Duffy BR (2003) Anthropomorphism and the social robot. *Robot Auton Syst* 42(3):177–190
- Zawieska K (2015) Do robots equal humans? Anthropomorphic terminology in LAWS. In: 2015 Meeting of experts on lethal autonomous weapons systems, 13–17 Apr 2015
- Carpenter J (2013) Just doesn't look right: exploring the impact of humanoid robot integration into explosive ordnance disposal teams. In: Luppini R (ed) Handbook of research on technoself. Identity in a technological society. IGI Global, Hershey, pp 609–36
- Galliot J, Macintosh D, Ohlin JD (2021) Introduction. In: Galliot J, MacIntosh D, Ohlin JD (eds) Lethal autonomous weapons: re-examining the law and ethics of robotic warfare. Oxford University Press, Oxford, pp 1–15
- Sharkey N (2012) The inevitability of autonomous robot warfare. *Int Rev Red Cross* 94(886):787–799
- Shim J, Arkin R, Pettinatti M (2017) An intervening ethical governor for a robot mediator in patient-caregiver relationship: implementation and evaluation. In: 2017 IEEE International conference on robotics and automation (ICRA), pp 2936–2942
- Slocombe G (2019) Autonomous warrior: major air, land and sea exercise, Asia-Pacific Defence Reporter, September 26. <https://asiapacificdefencereporter.com/autonomous-warrior-2018/>. Accessed 1 May 2019
- Hill SG, Barber D, Evans WA (2015) Achieving the vision of effective soldier-robot teaming. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts—HRI'15 extended abstracts
- Cosenzo KA, Barnes MJ (2010) Human-robot interaction research for current and future military applications: from the laboratory to the field. In: Unmanned systems technology XII

29. Dacey M (2017) Anthropomorphism as cognitive bias. *Philos Sci* 84(5):1152–1164
30. Damiano L, Dumouchel P (2018) Anthropomorphism in human–robot co-evolution. *Front Psychol* 9(March):117
31. Dautenhahn K (2007) Socially intelligent robots: dimensions of human–robot interaction. *Philos Trans R Soc Lond Ser B Biol Sci* 362(1480):679–704
32. DiSalvo C, Gemperle F (2003) From seduction to fulfillment: the use of anthropomorphic form in design. In: *Proceedings of the 2003 international conference on designing pleasurable products and interfaces. DPPI '03*. ACM, New York, pp 67–72
33. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Hum Factors* 37(1):32–64
34. Epley N, Waytz A, Akalis S, Cacioppo JT (2008) When we need a human: motivational determinants of anthropomorphism. *Soc Cogn* 26(2):143–155
35. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. *Psychol Rev* 114(4):864–886
36. Eyssel FA, Pfundmair M (2015) Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: a case study with a zoomorphic robot. In: *2015 24th IEEE International symposium on robot and human interactive communication (RO-MAN)*
37. Fink J (2012) Anthropomorphism and human likeness in the design of robots and human–robot interaction. In: Ge SS, Khatib O, Cabibihan JJ, Simmons R, Williams MA (eds) *Social robotics. ICSR 2012. Lecture notes in computer science*, vol 7621. Springer, Berlin. https://doi.org/10.1007/978-3-642-34103-8_20
38. Gallagher S (2008) 22—Understanding others: embodied social cognition. In: Calvo P, Gomila A (eds) *Handbook of cognitive science*. Elsevier, San Diego, pp 437–452
39. Galliot JC (2015) *Military robots: mapping the moral landscape*. Routledge, New York
40. Garreau J (2007) Bots on the ground in the field of battle (or even above it): robots are a soldier's best friend. *Washington Post*. May 6
41. Harrison MA, Hall AE (2010) Anthropomorphism, empathy, and perceived communicative ability vary with phylogenetic relatedness to humans. *J Soc Evol Cult Psychol JSEC* 4(1):34
42. Heijnen S, de Kleijn R, Hommel B (2018) The impact of human–robot synchronization on anthropomorphization. *Front Psychol* 9:2607
43. Hoffman G (2007) *Ensemble: fluency and embodiment for robots acting with humans*. Ph.D. Media Arts and Sciences, September, MIT
44. Hoffman G, Breazeal C (2007) Effects of anticipatory action on human–robot teamwork efficiency, fluency, and perception of team. In: *Proceedings of the ACM/IEEE international*
45. Turnbull G (2019) Why the British army tested robots in muddy fields, C4ISRnet, January 11, 2019. Accessed 1st May 2019
46. Mirrless T (2016) *Hearts and mines*. UBC Press, The US Empire Cultural Industry, Vancouver
47. Nyholm S (2018) The ethics of crashes with self-driving cars: a roadmap II. *Philos Compass* 13:e12506
48. Obaid M, Kistler F, Kasparavičiūtė G, Yantaç AY, Fjeld M (2016) How would you gesture navigate a drone?: A user-centered approach to control a drone. In: *Proceedings of the 20th international academic mindtrek conference, AcademicMindtrek '16*. ACM, New York, pp 113–121
49. Ono T, Imai M, Ishiguro H (2000) Anthropomorphic communications in the emerging relationship between humans and robots. In: *Proceedings 9th IEEE international workshop on robot and human interactive communication. IEEE RO-MAN 2000 (Cat. No.00TH8499)*
50. Riek LD, Rabinowitch T, Chakrabarti B, Robinson P (2009) How anthropomorphism affects empathy toward robots. In: *2009 4th ACM/IEEE International conference on human–robot interaction (HRI)*, pp 245–246
51. Salem M, Eyssel F, Rohlfing K, Kopp S, Joubin F (2011) Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In: Mutlu B, Bartneck C, Ham J, Evers V, Kanda T (eds) *Social robotics*. Springer, Berlin, pp 31–41
52. Sandoval EB, Mubin O, Obaid M (2014) Human robot interaction and fiction: a contradiction. In: Beetz M, Johnston B, Williams MA (eds) *Social robotics. ICSR 2014. Lecture notes in computer science*, vol 8755. Springer, Cham. https://doi.org/10.1007/978-3-319-11973-1_6
53. Servais V (2018) Anthropomorphism in human–animal interactions: a pragmatist view. *Front Psychol* 9(December):2590
54. Severson RL, Shailee RW (2018) Imagining others' minds: the positive relation between children's role play and anthropomorphism. *Front Psychol* 9(November):2140
55. Shen S, Slovak P, Jung MF (2018) 'Stop. I see a conflict happening' a robot mediator for young children's interpersonal conflict resolution. In: *Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction*, pp 69–77
56. Sparrow R (2016) Robots and respect: assessing the case against autonomous weapon systems. *Ethics Int Affairs* 30(1):93–116
57. Sparrow R, Howard M (2017) When human beings are like drunk robots: driverless vehicles, ethics, and the future of transport. *Transport Res Part C Emerg Technol* 80(July):206–215
58. Tahiroglu D, Taylor M (2018) Anthropomorphism, social understanding, and imaginary companions. *Br J Dev Psychol* 37:284–299
59. Tucker P (2020) An AI just beat a human F-16 Pilot in a dog-fight—again. *Defense One* online. <https://www.defenseone.com/technology/2020/08/ai-just-beat-human-f-16-pilot-dogfight-again/167872/>. Accessed 20 Aug 2020
60. Urquiza-Haas EG, Kotrschal K (2015) The mind behind anthropomorphic thinking: attribution of mental states to other species. *Anim Behav* 109(November):167–176
61. Bartneck C, Kanda T, Mubin O, Al Mahmud A (2009) Does the design of a robot influence its animacy and perceived intelligence? *Int J Soc Robot* 1:195–204
62. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J Soc Robot* 1(1):71–81
63. Carpenter J, Eliot M, Schultheis D (2006) The uncanny valley: making human–nonhuman distinctions. In: *Proceedings of the 5th international conference on cognitive science*, pp 81–82
64. Waytz A, Cacioppo JT, Epley N (2014) Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect Psychol Sci* 5(3):219–232
65. Airenti G (2018) The development of anthropomorphism in interaction: intersubjectivity, imagination, and theory of mind. *Front Psychol* 9(November):2136
66. Złotowski JA, Sumioka H, Eyssel F, Nishio S, Bartneck C, Ishiguro H (2018) Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism. *Int J Soc Robot* 10:701–714
67. Airenti G (2015) The cognitive bases of anthropomorphism: from relatedness to empathy. *Int J Soc Robot* 7(1):117–127
68. Vertesi J (2012) Seeing like a Rover: visualization, embodiment, and interaction on the Mars Exploration Rover Mission. *Soc Stud Sci* 42(3):393–414
69. Baggiani B (2015) Drone warfare and the limits of sacrifice. *J Int Polit Theory* 11(1):128–144
70. Bakardjieva M (2015) Rationalizing sociality: an unfinished script for socialbots. *Inf Soc* 31(3):244–256. <https://doi.org/10.1080/01972243.2015.1020197>

71. Cappuccio ML, Sandoval EB, Mubin O, Obaid M, Velonaki M (2020) Can robots make us better humans? Virtuous robotics and the good life with artificial agents. *Int J Soc Robot* 2020:6
72. Cappuccio ML (2015) Inference or familiarity? *Synth Philos* 58(2):253–272
73. Cappuccio ML, Peeters A, McDonald W (2019) Sympathy for Dolores: moral consideration for robots based on virtue and recognition. *Philos Technol* (February) 33(1):1–23
74. Wodehouse A, Brisco R, Broussard E, Duffy A (2018) Pareidolia: characterizing facial anthropomorphism and its implications for product design. *J Des Res* 16(2):83–98
75. Chen JY, Quinn S, Wright J, Barnes MJ, Barber D, Adams D (2013) Human-agent teaming for robot management in multitasking environments. In: 2013 8th ACM/IEEE International conference on human–robot interaction (HRI)
76. Chérif L, Wood V, Marois A, Labonté K, Vachon F (1918) Multitasking in the military: cognitive consequences and potential solutions. *Appl Cogn Psychol* 32(4):429–439
77. Cauchard JR, Zhai KY, Spadafora M, Landay JA (2016) Emotion encoding in human–drone interaction. In: 2016 11th ACM/IEEE International conference on human–robot interaction (HRI), Christchurch, 2016, pp 263–270

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Massimiliano L. Cappuccio is Senior Lecturer and Deputy Director of the Values in Defense and Security Technology Group at the University of New South Wales. He also holds a secondary affiliation with UAE University, the national university of the United Arab Emirates, where he had been Associate professor of Cognitive science and director of the Cog Sci Lab (August 2011 - December 2018). As a cognitive philosopher and a technology ethicist, his research is interdisciplinary and aims to integrate phenomenological analyses, empirical experimentation, and synthetic modelling. His work focuses on theoretical and applied issues in human performance (skill acquisition and skill disruption), human-robot interaction (social and assistive robotics), philosophy and theory of artificial intelligence (frame problem, computational theory of mind), from a perspective that integrates embodied, enactive and extended approaches to social cognition, virtue ethics,

and complexity theory. He conducts an intense activity as organizer of academic events, including interdisciplinary workshops, research seminar series, and international conferences (like the TEMPER workshop on Training, Enhancement, and Military Performance at ADFA, and the annual Joint UAE Symposium on Social Robotics in Abu Dhabi).

Jai C. Galliot leads the Values in Defence & Security Technology Group within the University of New South Wales at the Australian Defence Force Academy. As a former Royal Australian Navy officer and Australian Army Research Fellow, his recent books include *Military Robots: Mapping the Moral Landscape*; *Ethics and the Future of Spying: Technology, National Security and Intelligence Collection*; *Super Soldiers: The Ethical, Legal and Social Implications*; *Commercial Space Exploration: Ethics Policy and Governance*; and *Force Short of War in Modern Conflict: Jus ad Vim*. He is nonresident fellow at the Modern War Institute at West Point and visiting fellow with the Centre for Technology and Global Affairs at the University of Oxford.

Dr Eduardo B. Sandoval is Scientia Associate Lecturer at the University of New South Wales (UNSW) since 2018. Dr Sandoval has an interest in the design of machines that are capable of meaningful social interactions with humans. Dr Sandoval's work incorporates insights from behavioural economics and social psychology to explore different approaches in social robotics. His pioneering work has evolved around the topics of Reciprocity in Human Robot Interaction (HRI), Robot Design, and Robot Games. Similarly, he has explored the use of robots in education and healthcare and future addiction to robots in the academia and industry. Dr Sandoval has produced thirty-four peer-review publications in relevant venues related to HRI, a robot design register in 2019, and his work in Human-Robot Interaction has circulated in international media. Similarly, he has been appointed as co-chair, guest editor, and reviewer for several conferences and journals related to HRI, and Social Robotics and has collaborated closely with industry in the design and development of robot prototypes. In 2014, The National Autonomous University of Mexico awarded Dr Sandoval with the Medal Alfonso Caso as the best student of the Postgraduate Program of Industrial Design at the National Autonomous University of Mexico. In 2016, Dr Sandoval obtained his PhD in Human Interface Technology at the HITLab NZ, University of Canterbury, New Zealand. His future work involves multidisciplinary research in social robotics as robot games and no-visual natural interfaces aiming future human's well-being.

Chapter 3, Part 2 - Ethical Attitudes towards Robots: Robots that Cultivate Character

This section is centred on Core Publication #7, #8, and #9 [CP7] [CP8] [CP9]

3.2.1 Vicious relationships with robots

The first Part of this Chapter identified certain risks that anthropomorphism offers to HRI designers: humanizing robots might bring users to overestimate their abilities, undermining optimal coordination in joint tasks where each agent completely relies on the other's capabilities. The safety concerns discussed in section 3.1.6 represents only a particular case of a broader kind of worry related to anthropomorphism: that the relationship with robot could produce a moral or psychological harm, in addition to a physical harm. In this case, the risk produced by anthropomorphism is not that of a misuse, but of an abuse: the asset put at risk by the relationship with the robot is not the physical safety of the user, but their character or, in other words, their moral and psychological integrity. In "Sympathy for Dolores: Moral Consideration for Robots based on Virtue and Recognition", the paper I published in *Philosophy & Technology* (Cappuccio, Peeters & McDonald 2019 [SP18]), I explore this kind of worry, identifying some domains in which these kinds of unwanted and problematic anthropomorphic attributions are particularly likely to happen. These domains are characterized by three features: the intimate nature of the pseudo-relationships that human users seem to establish (or believe to have established) with robots; the unpreparedness of the users in dealing with the subtle emotional ramifications of these pseudo-relationships; and the fact that the human-robot relationship appears to human more successful, in these scenarios, if it involves a high degree of trust towards the robots and even an element of affection. The applicative domains in which human character is most obviously vulnerable to the damage caused by improper anthropomorphic attributions (and by other user's attitudes) are essentially three: "(i) Children spoiled by the artificial nannies that raise them (Bryson 2010), (ii) adult consumers of robotic sex (Whitby 2012; Sparrow 2017), and (iii) players of games based on the violent abuse or bullying of robots (Bartneck & 2008; Nomura 2016; Sparrow 2016) represent the most paradigmatic and worrying hypothetical scenarios" (idem). There can be, of course, others too. The user's age constitutes a special concern in scenarios like these, as the unwanted and problematic consequences of anthropomorphism might be even more damaging in development or during a formative stage, especially if the interaction with the robot is not properly supervised.

3.2.2 Morally corruptive robots?

What are, more specifically, the unwanted and problematic consequences that these scenarios might engender? The disadvantage potentially produced by anthropomorphism physical harm can be perfectly accounted for by an utilitarianist perspective, which judges negatively the consequences of an inappropriate bond between human-robot or misuse (for example, an economical or strategic loss may be caused by the impossibility to successfully carry out a certain task with a robot due to inappropriate anthropomorphism). But, as I said, the worries related to anthropomorphism in the context HRI involve a much broader horizon and, interestingly, encompass some non-utilitarianist moral implications: namely, it is possible that an inappropriate relationship with a robot damages a person not by taking away from them something they have or aspire to own or control, but by diminishing the kind of person that they are (Coeckelbergh 2018).

For example, frequent interaction with a robot could make the users more vulnerable to certain dependencies. Dependency (Vallor 2011), deskilling (Vallor 2015), and “addiction” are risks associated with the possibility of over-relying on robots (or of “abusing” them), charging the relationship with robot of exaggerate significance and expectations: all these risks have in common the undermining of human agency, capability, decisional autonomy, and sense of responsibility (Cappuccio et al. 2021b [CP8]). But, while broad and deep, the problem of dependency and over-reliance is just a particular instantiation of an even broader class of problems that I have explored during my collaboration with a group of social roboticists and HRI experts based in Sydney (Mari Velonaki, Eduardo Benitez Sandoval, Omar Mubin, Mohammed Obaid). This broader problematic horizon is defined by the risk that autonomous technologies could corrupt the moral character of robot users, engendering in them “moral deskilling” (incapability of exerting their virtues, or positive character traits), reducing their moral standing and their value as praiseworthy individuals (affecting their self-respect), and alienating them from the prospect of living a good and honourable life (conducting them to poor life choices).

These dynamics are at the centre of a trio of publications co-authored/co-edited with my colleagues in Sydney: Cappuccio et al 2021a [CP7], 2021b [CP8], 2021c [CP9]. Even the experience of being served by artificial beings that do never complain or protest to their owner could be a reason of concern. Casual interaction with social robots could become a perfect trigger mechanism for this regressive dynamic, because they are meant to simulate unconditional recognition without expecting to receive any in return. Unlike the humans who serve in a subordinate position, robots might be programmed to discount or condone

disrespectful and inconsiderate behaviours. This, in turn, might justify the human master's sense of entitlement even more, reinforcing their authoritarian stance and their beliefs of superiority/uniqueness.²⁹

But if robots don't have rights or the moral patient status that all humans have, then what exactly is the problem with empathizing with robots, as an unwanted effect of anthropomorphism? The problem is not the misattribution or misjudgement in itself, but the fact that it can deteriorate our moral character, ruining the kinds of persons that we are. Our spontaneous tendency to anthropomorphise robots, forming empathic relationships with them, explains why we cannot help but perceive humanoid robots *as moral patients*, capable to suffer pain and injustices (even if they are neither conscious nor sentient). Because empathy is a fundamental feature of social cognition in all typical subjects, it is inevitable that we perceive pain and misery in the robots that are subject to us and that we treat as mere tools (Asada 2021b). Because of this empathic bond, mature relationships between humans and robots call for a cautious but firm "moral consideration" (Levy 2009; Cappuccio, Peeters, McDonald 2019 [SP18]). This moral consideration for robots is justified even though robots are inanimate beings without rights or other entitlements, for two sets of reasons: first, our behaviours and thoughts over robots reveal the kind of moral agents that we are, and whether our moral character is good or bad; second, the quality of our recurrent interactions with robots and the way we habitually treat them can in the long run affect our moral character, promoting either virtuous or (regrettably, more often) vicious inclinations in the human users and among the bystanders.

3.2.3 Virtuous Robotics: a socio-technical framework to become better humans

This kind of concern invites a new ethical approach towards autonomous and intelligent technologies, one that (unlike most of the other approaches to technology ethics) is not informed by consequentialist or deontological principles, but by the ethics of virtues and the consideration of the inherent value of moral character and self-cultivation. As autonomous technologies become more capable, more intelligent and more deeply integrated into our societies, the debate on the ethical ramifications of AI-based systems and robots is

²⁹ Such privilege towards robots is likely to encourage self-indulgent and complacent habits, boost the self-awareness of the human users, erode their inhibitions, spoil their sense of empathy, and—worst-case scenario—motivate them to tolerate, justify, or even replicate abusive behaviours against actual living creatures. If these dynamics were replicated on a massive scale, they could exacerbate social tensions in large communities and deteriorate civil cohesion.

becoming a central concern for researchers working in the academia, in the industry, and in the institutions of the civil society. This field develops around two distinct and somewhat complementary lines of inquiry, represented -on the one hand- by Robot Ethics (or AI ethics, more generally) and -on the other- by Machine Ethics.

Machine ethics asks how we can build a robot or AI capable to correctly make ethical decisions, understanding the moral implications of its choices and evaluating the ethical value of certain actions or conditions (Allen, Wallach, & Smit 2006; Torrance 2011). AI/Robot ethics, in turn, asks what principles should guide the developers of autonomous technologies so that their creations comply with the relevant societal norms and moral values, meeting the community's moral expectations and ethical requirements (Lin, Abney & Bekey 2011; Burton et al 2017)³⁰. Accordingly, developers must be pro-actively involved in promoting awareness about the importance of these values and the efforts required to correctly implement them in a technological platform that is adaptive and aligned with the expectations of various stakeholders. It is useful to contextualize my work within the ethical inquiry on autonomous and intelligent technologies.

3.2.3.1 The principles of Virtue Ethics – Virtue ethics promotes a new approach to both Robot Ethics and Machine Ethics. While machine ethics focuses on the ethical correctness of machine decisions (Wallach & Allen 2009; Bendel 2017) and AI ethics on the moral responsibility of designers and users, Virtue Ethics shifts the discussion to a broader and more transformative question: *Do robotics and AI make us better or worse humans?* This approach evaluates how these technologies shape human moral character and societal values, placing greater emphasis on their impact on human flourishing than on the internal functioning of the systems themselves (Sparrow 2017). Virtue ethics is one of the normative theories of ethics, i.e. the theories that aim to provide an answer to questions such as “what is good/bad?” and “how can we distinguish right from wrong?” (Abney 2012). As such, Virtue ethics provide different lenses for addressing the moral challenges specifically associated with AI and autonomy. Among the theories of normative ethics, the main contenders of virtue ethics are *Deontology* and *Consequentialism*, which provide distinct (yet often complementary) perspectives, illuminated by different core ideas, applications to AI, and challenges.

³⁰ For example, one question addressed by AI/Robot Ethics concerns the so-called ‘value-sensitive design’, i.e. the idea that since the conceptualization and prototyping stages, the design of autonomous technologies must be informed by a set of relevant moral values that reflect the true aspirations and expectations of the society (Peeters & Haselager 2021).

Deontology evaluates the morality of actions based on adherence to rules, duties, or principles. It involves encoding explicit ethical rules into AI systems. However, these rules struggle in situations involving conflicting or ambiguous scenarios (Lumbreras 2017; Bankins & Formosa 2023). For instance, should an autonomous vehicle prioritize saving the maximum number of lives or protecting its passengers at all costs? *Consequentialism* evaluates the morality of actions based on their outcomes, aiming to maximize overall good or utility. But predicting outcomes is complex, trade-offs can lead to morally contentious decisions, and the calculus of "utility" often competes with justice and fairness considerations (Malle 2015). *Virtue ethics* emphasizes moral character and virtues rather than rules or outcomes. By asking us the question "What kind of person (or society) should we aim to be?", Virtue ethics invites us to critically evaluate whether AI systems promote virtues such as empathy, fairness, and responsibility in their users and developers (Torrance 2011; Vallor 2016). It encourages reflection on the broader societal impact exerted by AI on human flourishing, considering that translating abstract virtues into technical specifications can be difficult and there is no obvious universal protocol to do it. That is why assessing whether AI truly fosters virtues requires continuing awareness and societal reflection.

Each theory of normative ethics faces unique challenges when applied to robot ethics: Deontology is challenged by the fact that robots programmed with strict ethical rules (like Asimov's famous three laws, see Anderson 2008) may encounter dilemmas whenever these rules conflict with each other, require interpretation, or allow antinomic outcomes (for example, should a caregiving robot prioritize following orders or safeguarding the patient's well-being?); Consequentialism is challenged by the fact that robots making decisions based on maximizing utility might face backlash for decisions that conflict with human emotional or cultural values (when is sacrificing one life to save many acceptable?); Virtue Ethics confronts the question of whether machines can genuinely embody virtues or whether this framework applies solely to the humans designing and using them. Virtue Ethics shifts the focus of ethical inquiry from actions (or outcomes) to self-improvement and moral character "cultivation" (Cappuccio, Peeters & McDonald 2019 [SP18])³¹.

³¹ Key concepts include: *Character* (the consistent landscape of an individual's moral traits, shaped by habits, education, and societal influences); *Virtues* (positive character traits such as honesty, courage, empathy, and wisdom that enable individuals to lead morally excellent lives); *Practical Wisdom*, or *Phronesis* "the ability to navigate complex moral situations with sound judgment and adaptability" (the contributions to our special issue of IJSR dedicated to 'Virtuous Robotics: Artificial Agents and the Good Life' offer a detailed discussion of these concepts, Cappuccio et al 2021c [CP9]).

Virtue ethics appreciates the role of autonomous, intelligent technologies in shaping human character and societal values. Unlike Deontology or Consequentialism, which focus on the machine's behaviour or outcomes of its actions and decisions, Virtue Ethics asks whether these technologies cultivate or erode virtues in their creators, users, and society at large. A virtuous approach to robotics is beneficial to address some key moral aspects of our relationship with autonomous machines that other normative approaches to ethics are barely able to detect or thematize: (i) *Encouraging Virtuous Behaviour* (AI systems can be designed to promote virtues, such as fairness in decision-making algorithms or empathy in social robots used for caregiving, Coeckelbergh 2018; Vallor 2011); (ii) *Shaping Responsible Developers* (The ethical development of AI requires developers to embody virtues like integrity, transparency, and accountability, Richards, Vythilingam & Formosa 2023; Marchesi & Wykowska 2024); (iii) *Fostering Human Flourishing* (Virtue ethics encourages reflection on whether AI technologies contribute to societal well-being, wisdom, and justice, Coeckelbergh 2010; Sparrow 2016; Vallor 2016).

In an era of rapid technological change, an ancient tradition like Virtue Ethics offers unique advantages to the user, the technologist, and the policy maker: it has a *Human-Centric Focus* (by prioritizing human moral development, virtue ethics ensures that technology serves human flourishing rather than undermining it); it is *Non-Aprioristic* and adherent to experience (Virtue ethics provides a flexible framework for navigating the complexities of real-world scenarios that defy rigid rules or predictive optimization); *it Fosters Trust* (by aligning AI systems with human virtues, virtue ethics promotes public trust in their ethical development and deployment, Lewis, Sycara, & Walker 2018).

3.2.3.2 Operationalising Virtue Ethics – Operationalising virtue cultivation and character development in robotic systems requires translating normative philosophical concepts into technically tractable design specifications that can guide the engineering of interactive machines. At its core, this endeavour involves mapping the Aristotelian notion of virtue as habituated excellence in action to processes of *iterative feedback, reinforcement, and scaffolding* in human–robot interaction (HRI).

Roboticians can pursue this translation through several interlocking methodological streams. First, *embodied interaction design* provides the basis for aligning robotic behaviours with the moral or educational goals of the interaction. Here, anthropomorphic embodiment, expressive motion, and multimodal communication channels (facial expressions, gestures, prosody) are deliberately engineered to produce the kinds of social cues that invite empathic engagement and moral reflection. Platforms such as *KASPAR*

(Wainer et al., 2014; Dautenhahn et al., 2009) and other humanoid companions demonstrate how limited but carefully tuned nonverbal behaviours can create scaffolding for teaching socially desirable habits (e.g., gentle touch, turn-taking). Second, *adaptive learning algorithms*—particularly reinforcement learning, affective computing, and user-modelling—are crucial for tailoring interventions to individual trajectories of character development. Reinforcement learning allows robots to adapt their prompts and feedback dynamically based on user responses, operationalising the cultivation of habits through the reinforcement of prosocial behaviours and the gentle discouragement of harmful ones.

These algorithms are augmented by *affective sensing* technologies (facial-expression recognition, physiological sensing, voice emotion analysis) that permit the robot to detect the user's motivational and emotional states in real time, enabling contingent responses that can nurture dispositions such as patience, empathy, or self-control. Third, the use of *persuasive technology methods* (Fogg, 2003; Berdichevsky & Neuenschwander, 1999) enables robotic systems to integrate nudges, goal-setting prompts, and feedback loops that shape behaviour gradually without coercion. For example, a virtuous robot might employ micro-interventions—praising cooperative play, providing reminders of long-term goals, or highlighting the consequences of selfish behaviour—that are aligned with established psychological principles of habit formation (Cialdini, 2007; Thaler & Sunstein, 2008). Fourth, roboticists can operationalise virtue cultivation by embedding *pedagogical scaffolds* in the interaction sequence, drawing from educational psychology (Kim & Hannafin, 2011; Narvaez, 2006). These scaffolds involve structuring tasks such that the robot initially provides strong guidance (explicit reminders of norms, step-by-step instruction) and gradually withdraws support as the user internalises the practice, thereby encouraging autonomy and self-regulation. Fifth, *longitudinal personalisation methodologies* are critical: virtue cultivation requires sustained engagement over time, not one-off interventions. Roboticists must therefore integrate persistent user-profiles and longitudinal data analysis into system architectures, so that the robot can track developmental progress and adjust interventions according to evolving character dispositions. Techniques from recommender systems and longitudinal reinforcement learning can be adapted to monitor changes in behaviours (e.g., sharing frequency, emotional regulation, cooperation) and recalibrate the robot's strategies. Sixth, a *multi-agent ecological perspective* highlights that virtue cultivation rarely occurs in dyadic isolation but within broader social contexts (families, classrooms, clinics). Robotic systems should thus be equipped with protocols for *triadic and group interactions*, supporting not just direct moral training but also the mediation of peer relationships and the reinforcement of socially cooperative norms. This requires procedural innovations such as synchronisation algorithms for coordinating multiple users, role-

switching mechanisms that allow the robot to model both positive and negative behaviours, and integration with external platforms (educational apps, therapeutic protocols) that provide continuity across settings. Seventh, to prevent undesirable outcomes such as dependency or deception, roboticists must implement *transparent design features*: explainable AI modules that render the system's pedagogical strategies legible to users and caregivers, adjustable autonomy settings that allow human overseers to calibrate the degree of influence exerted by the robot, and ethical oversight protocols that ensure interventions remain aligned with human flourishing rather than manipulative control.

The technological operationalisation of virtue cultivation thus involves balancing expressive embodiment, adaptive algorithms, persuasive scaffolding, longitudinal monitoring, and contextual embedding, underpinned by rigorous ethical safeguards. In practice, this means roboticists must establish pipelines that begin with philosophical modelling of virtues (e.g., courage as appropriate risk-taking, temperance as moderation), translate these into measurable behavioural indicators (frequency of waiting turns, modulation of voice tone, persistence in face of failure), and then design interaction scripts and adaptive feedback loops that reward their enactment. As our empirical studies in *virtuous robotics* and allied work on moralising persuasive technologies show, robots can function as *mirrors and mentors*, simultaneously reflecting user behaviours back for self-awareness and nudging them toward better habits. Although the implementation is fraught with risks of bias, overgeneralisation, or cultural misalignment, the methodological pathway is increasingly clear: virtue cultivation through robotics is operationalised by embedding ethical scaffolds directly into the *technical stack of embodiment, sensing, and learning*, so that artificial agents can function not merely as tools but as facilitators of moral growth.

3.2.3.3 Robotics applications of Virtue Ethics – Because social robots are able to shape human beliefs and emotions, people need to take a more ethical approach to their design and our interactions with them. As an ancient philosophy of self-betterment and human flourishing, Virtue Ethics perfectly fits these requirements. Instead of trying to build robots that imitate our ethical decision-making processes, Virtue Ethics allows us to see our own interactions with robots as an opportunity of human betterment and moral elevation. Virtuous Robotics theory emphasises the responsibility of the human in every morally-sensitive form of engagement with robots, such as with the AI humanoid Pepper. Robots are not always intelligent enough to make the best ethical choice on your behalf but can help you make the best ethical choice by reminding you who you are and what you are striving for: creating awareness, coaching, or simply encouraging you. Generosity, courage, honour, compassion

and integrity are examples of universal virtues that researchers in the paper hope to encourage in humans through their use of social robots.

Shannon Vallor’s *The AI Mirror* (Vallor 2024) aligns with our IJSR paper on virtuous robotics: AI mirrors human behaviour and can foster mindfulness and self-improvement. Virtuous Robotics has important applications to social robotics because it is interested how AI can support education, good habits, and human flourishing. A key example is Kasper (Kinesics and Synchronisation in Personal Assistant Robot), a child-sized humanoid first developed at the University of Hertfordshire in 2005 and later acquired by UNSW. Kasper assists children with autism and learning difficulties, teaching socially acceptable behaviours—for instance, saying “that hurts” when hit or “that feels good” when touched gently. While not replacing therapists, families, or schools, Kasper helps children learn social interaction through play and experimentation. Social robots like Kasper cultivate virtue more effectively than mobile apps, since rich social engagement remains central to moral development; embodied robots that reproduce these interactions provide a valuable means to foster virtue. The table below summarizes some of the main systems and studies adopted to cultivate virtue in a way consistent with the virtuous robotics paradigm³².

Fig. 33 - Notable Applications of the Virtuous Robotics Paradigm

System / Study (year)	Field / Application	Targeted quality, virtue, skill	Method / HRI strategy	Participants / Study type	Headline outcome	Key reference
KASPA - triadic & dyadic collaborative play (2013–2014)	Autism therapy (children)	Cooperative play, turn-taking, joint attention, gentle touch	Minimally expressive humanoid; structured play tasks; autonomous prompts	Children with ASD; pilot studies (triadic video-game & dyadic play)	Improved collaborative play and engagement vs. baselines	Wainer et al., 2014; Dautenhahn et al., 2009.
NAO - attention assessment & personalised sessions (2021)	Autism therapy (children)	Sustained attention, self-regulation	Automated attention scoring; adaptive task sequencing	6 children with ASD; multi-session pilot	Therapist–system agreement ~82.4%; feasibility of personalised attention training	Alnajjar et al., 2021.

³² These entries were selected to emphasize developmental, pedagogical, and moralizing functions. Where the outcomes are not explicitly labelled “virtues,” I have mapped them to virtue-relevant dispositions (e.g., cooperation, self-regulation, perseverance, empathy). Some programs report quasi-experimental evidence rather than large RCTs; several have long-term deployments (e.g., PARO in care settings; Tega/RUBI in classrooms). Interestingly, designs that avoid “over-realism” (e.g., KASPAR, Keepon) often succeed by clarifying cues, reducing ambiguity burden for users with atypical processing (which is useful when cultivating habits like turn-taking and gentle touch).

System / Study (year)	Field / Application	Targeted quality, virtue, skill	Method / HRI strategy	Participants / Study type	Headline outcome	Key reference
Keepon (2007–2009)	Autism & developmental support	Social engagement, affect sharing, rhythmic coordination	Very simple morphology; contingent movement to child's actions	Field observations & pilots with children (incl. ASD)	Robust engagement; facilitation of social interaction	Kozima et al., 2009; HRI 2007 demos.
PARO robotic seal (2001–2007)	Eldercare / dementia care	Calm/emotional regulation, empathy, social connectedness	Tactile/affective interaction; daily “robot therapy” in care homes	Longitudinal deployments in care houses & day centers	Reduced stress, improved mood; more social interaction	Wada & Shibata, 2007; Wada et al., 2004.
Probo as social-story agent (2012)	Autism education	Social norms (greeting, thanking, sharing), perspective-taking	Narrated social stories; expressive facial/gaze cues	Children with ASD; HRI conference pilot	Robot-led stories may outperform human reader for social performance	Simut et al., 2012; Saldien (PhD monograph).
RUBI preschool tutoring (2005–2009)	Early childhood education (toddlers)	Language & vocabulary; prosocial classroom participation	Immersive classroom deployment; autonomous sessions	18–24-month toddlers; 2-week immersion; controlled vocab tasks	Significant vocabulary gains; demonstrates classroom feasibility	Eckhardt et al., HRI 2009; Movellan et al., ICDL 2005.
QRIO / Early socialisation (2007)	Early childhood socialisation	Peer-like social engagement, imitation	Humanoid peer interactions in ECEC	Toddlers in preschool; naturalistic study	Evidence of robot–toddler socialisation dynamics	Tanaka et al., PNAS 2007 (lab site).
Tega (2016–2017)	School-based learning & SEL	Persistence, empathy, vocabulary, L2 learning	Expressive long-term companion; affect sensing/personalised dialogue	Classroom pilots (weeks-long); controlled comparisons	Personalised interactions improve learning/engagement	HRI 2016 platform; MIT field studies.
Socially Assistive Robots (SAR) for rehab (2008–2009)	Post-stroke & older adults	Adherence, perseverance, self-efficacy (character strengths)	Personality-matching; adaptive coaching, empathic feedback	Controlled pilots with patients	Personality matching boosts adherence & outcomes	Tapus, Țăpuș, & Mătarici, 2008; AAAI/SSR 2008.

3.2.4 A New Approach to Robot Ethics

The studies reviewed in this section point to the ramifications of HRI in controversial, yet rapidly emerging, thematic areas in the scientific literature: moral patiency of artificial agents, moral consideration for robots, and “AI welfare” (Gerdes 2016). Our proposal, inspired, by Virtue Ethics and recognition theory (Laitinen 2016), enables a finer understanding of a too-

often overlooked dimension of robot ethics, the dimension concerning the integrity and the possibility of flourishing of the human moral character, that too often is negligently overlooked by robot ethicists to prioritize utilitarianist and instrumentalist considerations. The analyses conducted in this section allow us to conclude that, from the user's perspective, a "virtuous" approach to robotics (that is an approach to interaction design informed by virtue ethics) can make robots morally desirable in our lives: properly conceived and designed, robots can become supports for personal betterment and moral fulfilment. Additionally, it is possible to argue that, from a designer's perspective, Virtuous Robotics is both technically feasible and convenient because it transfers most of the cognitive and ethical burden of HRI from the robot to the user, bypassing the need for unrealistically sophisticated AIs equipped capable of complex moral decisions (Cappuccio 2016; Shim, Arkin & Pettinatti 2017) to favour in the user the kind of work of self-reflection, self-control and self-realization that is typically associated with virtue cultivation and character flourishing (Nørskov 2014). In accord with these conclusions, it is possible to extract a series of "virtuous" indications to concretely orient the design of robots and HRI protocols so to achieve the desiderata of virtue ethics and character development in future work.

3.2.4.1 Objectives – To be genuinely "virtuous", a robot (or an HRI interaction protocol) must be designed to help its user achieve at least one of the following nine objectives, that social robotics is still exploring (Gips 2011): (i) *Keep track of and reinforce motivation and pride in athletic training, academic study, and artistic performance* (e.g., in learning STEM skills, languages, socialization, among others); (ii) *Acquire or perfect broad skill sets that are relevant to self-development* (acting, meditation, writing, logical problem-solving - this includes, as illustrated in Chapter 2, also increasing innate cognitive/social abilities, Milne et al 2011b); (iii) *Recognize, inhibit, and subdue negative character traits* such as arrogance, impatience, anger, rudeness, bad habits and addictions (smoke, drugs), biases and prejudices (racism, sexism, elitism), criminal tendencies and corruption; (iv) *Develop or reinforce positive character traits*, like courage, compassion, charity, honesty, and positive ways of thinking, like hope and enthusiasm (Jung, Scassellati, & Christakis 2020); (v) *Encourage one's search for excellence and merit. Identify and foster vocational talents and professional inclinations* (for music, sport, public engagement), also through the acquisition of specific skills (second language, music composition); (vi) *Manage their inner conflicts and deal with their psychological challenges or limitations*, which may include fears, intolerances, addictions, post-traumatic stress disorders and structural impairments; (vii) *Reinforce personality traits related to self-regulation, discipline, restraint*, including resilience, focus, mental toughness, emotional awareness— as we will see in the next section of this chapter, robot invigilators deployed in classrooms during exams to prevent cheating and other

misbehaviours represent a promising example; (viii) *Advise and coach using exercises and practices that favour their physical and mental well-being*: meditation, workouts, prayer, counselling, healthy lifestyle (diet and sleep routine—see), and good habits like personal hygiene and care of self, to provide feedback and help the trainees reach the goals of each exercise session (Alotaibi, Alnajjar, Cappuccio & Khalid 2023 [SP15]); (ix) *Foster reflective attitudes, independent judgment, and emotional maturity (autonomy in dealing with personal challenges, professional competitions, romantic relationships)*.

3.2.4.2 Roles – Finally, we envision numerous roles that a social robot can take on to achieve any of these goals, including: (a) *Interactive diary and confidant*: the robot invites the human user to narrate their everyday experiences and reflect on their own conduct; (b) *Assistant, coach, and motivator* (following a predefined timetable set by a preparator, a therapist, or an instructor, the robot reminds the user of daily routines that can be part of an educational, training, or clinical program); (c) *Apprentice* (when a robot behaves appropriately or improperly or suggests its intention to do it, it solicits the corrective responses of a human user, who is then expected to act as a tutor, praising and blaming the robot according to its behaviour; (d) *Informer or “scarecrow”* (robots designed to monitor the user and supervise their activities on behalf of moral authorities can deter vicious behaviours simply with their presence as bystanders).

3.2.5 Outcomes, limitations, and future research

Despite the promises for HRI and Social Robotics, Virtuous Robotics faces a number of challenges and unintended consequences.

3.2.5.1 Dependency and Over-Reliance - One major concern with the Virtuous Robotics paradigm is the potential for users to become over-dependent on robotic systems for moral guidance. While the approach emphasizes self-improvement and reflective engagement (Cappuccio et al., 2021), there is a risk that individuals outsource their ethical deliberation to machines, thereby weakening autonomy rather than cultivating it. Similar concerns have been raised in relation to persuasive technologies and nudging systems (Burr et al., 2020; Vallor, 2016), where sustained reliance may erode, rather than reinforce, human agency. The danger is that robots intended to scaffold virtue development could foster passive acceptance of pre-programmed prompts or reminders, diminishing users’ capacity for critical moral judgment.

3.2.5.2 Deception and Anthropomorphism - Another objection is the possibility of deception in the human–robot relationship. By design, virtuous robots present themselves as social companions or moral mirrors. However, their simulated empathy and care are ultimately programmed behaviours, not genuine moral concern (Coeckelbergh, 2020). The anthropomorphic projection of sincerity can produce what has been called “the transparency paradox” (Nyholm, 2020), where the more effective a robot is at supporting moral cultivation, the more likely users are to attribute agency or authenticity that does not exist. This can undermine trust if users later perceive the robot’s actions as manipulative or instrumental rather than genuinely supportive.

3.2.5.3 Normative Ambiguity and Cultural Bias - The proposal that robots can cultivate virtue faces the problem of whose virtues are being promoted. As Vallor (2016) stresses, virtue ethics is embedded in cultural contexts; virtues such as autonomy, empathy, or filial piety are differentially prioritized across societies. Designing robots to encourage virtue risks bias and cultural overgeneralisation, particularly if values embedded in Western Aristotelian frameworks are imposed in diverse settings (Sparrow, 2007). Moreover, the operationalisation of abstract virtues into concrete behaviours involves subjective design choices, potentially embedding hidden normative assumptions into robotic systems.

3.2.5.4 Domain-Specific Fragility - Virtuous robots also inherit challenges well-documented in general-purpose AI. Just as large language models and other general-purpose systems face issues of bias, overfitting, and overgeneralisation (Bender et al., 2021), so too may robots designed to support human flourishing. For example, a robot trained to encourage prosocial behaviour in children with autism (e.g., Kasper; Wainer et al., 2014) may overgeneralise those strategies when interacting with neurotypical children, producing counterproductive or inappropriate interventions. Similarly, robots optimised for fostering empathy in therapeutic settings might behave awkwardly in educational or workplace contexts, highlighting the limitations of transferring general-purpose models into domain-specific tasks.

3.2.5.5 Limitations of General-Purpose AI - The risks outlined above map onto broader limitations of AI when deployed in socially situated contexts. General-purpose AI models often struggle with context sensitivity, leading to overgeneralised outputs when applied outside their training distribution (Amodei et al., 2016). In virtuous robotics, this manifests as rigid behavioural routines that fail to adapt to local cultural norms, situational demands, or individual developmental needs. Moreover, the risk of algorithmic bias (Mehrabi et al., 2021) is amplified when robots are explicitly designed to model, promote, or enforce ethical

norms. What is presented as “virtue cultivation” may unintentionally reflect and reinforce designers’ cultural presuppositions, institutional values, or market imperatives.

3.2.5.6 Mitigation Strategies - To address these challenges, several mitigation strategies can be considered: (i) *Transparency and Value Disclosure* - Robots should be designed with explicit disclosure of their limitations, including the fact that they simulate social engagement rather than embody it (Bryson, 2018). Value assumptions underlying the operationalisation of virtues should be documented, reviewed, and subjected to participatory design processes involving diverse cultural stakeholders. (ii) *Human-in-the-Loop Ethical Mediation* - Virtuous robots should function as assistive companions, not replacements for ethical reasoning. Embedding human oversight and feedback loops can ensure that moral cultivation remains grounded in human deliberation (Floridi & Cowls, 2019). (iii) *Context-Aware Adaptation* - Drawing lessons from domain-specific AI, robotic systems must be calibrated for particular contexts, rather than relying on one-size-fits-all models. Adaptive learning mechanisms and modular virtue-promoting routines can help mitigate overgeneralisation. (iv) *Bias Auditing and Cultural Pluralism* - Systems should be regularly audited for bias and normative drift, ensuring that the virtues promoted reflect plural cultural traditions (Whittlestone et al., 2019). International and interdisciplinary collaboration can prevent virtue models from becoming parochial or ideologically narrow. (v) *Fostering Moral Literacy in Users* - Instead of simply delivering prompts or corrective signals, virtuous robots can be designed to encourage reflection by asking questions, offering multiple perspectives, or scaffolding dialogue. In this way, they avoid dependency and promote critical engagement.

Core Publication #7 [CP7]

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2020),
“Can Robots Make Us Better Humans?
Virtuous Robotics and the Good Life with Artificial Agents”
International Journal of Social Robotics, 13(1), Springer Verlag.

Reproduced with permission from Springer Verlag. All rights reserved.



Can Robots Make us Better Humans?

Virtuous Robotics and the Good Life with Artificial Agents

Massimiliano L. Cappuccio¹ · Eduardo B. Sandoval² · Omar Mubin³ · Mohammad Obaid⁴ · Mari Velonaki²

Accepted: 14 September 2020
© Springer Nature B.V. 2020

Abstract

This position paper proposes a novel approach to the ethical design of social robots. We coin the term “Virtuous Robotics” to describe Human–Robot Interaction (HRI) designed to help humans reach a higher level of moral development. Our approach contrasts with mainstream approaches to robot design inspired by the other normative theories, Consequentialism and Deontology. In the paper we theoretically justify our proposal, illustrating how the Virtuous Robotics approach allows us to discriminate between positive and negative applications of robotics systems, of which we provide examples. From an ethical perspective, our proposal is theoretically robust because it is based on the assistive role played by the robot rather than the robot’s moral agency. From a designer’s perspective, Virtuous Robotics is technically feasible because it transfers the cognitive burden of HRI from the robot to the user, bypassing the need for complex decision-making abilities. From the user’s perspective, it is concretely advantageous, because it envisions a realistic way to make robots morally desirable in our lives, as supports for personal betterment and fulfilment.

Keywords Virtuous Robotics · Virtue Ethics · Self-improvement · Deontology · Consequentialism · Human–Robot interaction · Recognition theory · Character formation · Machine Ethics · Aristotle

1 Introduction

What approach should we use to design and use social robots in an ethical way? In this position paper, we intend to offer

an original take on this question presenting a new approach to the application of machine ethics in social robotics that we call “Virtuous Robotics”. The Virtuous Robotics approach to Human–Robot Interaction (HRI) is both morally desirable, because cogently derived from legitimate ethical principles, and technically sound, because feasible, convenient, and in some ways more adequate for implementation than the other approaches typically considered by robot ethicists.

Virtuous Robotics differs from the major approaches in machine ethics both in theory and practice. In theory, because our approach is not rooted in the principles of the mainstream theories of normative ethics, i.e., Consequentialism and Deontology, but on Virtue Ethics, a philosophical tradition that focuses on character development and self-cultivation [67,73]. We aim to illustrate how Virtue Ethics can suggest very plausible and efficacious principles to technology design and inspire robotics applications endowed with a positive moral valence. In this position paper we review these high-level principles and discuss how they can benefit HRI.

In practice, Virtuous Robotics is unique because it does not presuppose the robot’s ability to make ethically correct

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

Eduardo B. Sandoval
e.sandoval@unsw.edu.au

Omar Mubin
o.mubin@wsu.edu.au

Mohammad Obaid
mobaid@chalmers.se

Mari Velonaki
mari.velonaki@unsw.edu.au

¹ School of Engineering and IT, University of New South Wales Canberra, Canberra, Australia

² School of Art and Design, University of New South Wales, Sydney, Australia

³ Western Sydney University, Sydney, Australia

⁴ Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

decisions or behave in morally acceptable ways, but rather the robot's ability to contribute to the users' moral flourishing. Virtuous robotics aims to facilitate virtue cultivation by assisting users along a voluntary path of self-awareness, personal development, and character-building [16]. A positive ethical value can then be credited to a robot if it supports humans in their moral growth. This approach deflates the issue of whether the robot is acting morally or not, and stresses whether the user's relationship with the robot is psychologically healthy and morally desirable.

Our argumentation has three objectives. We want to show that Virtuous Robotics is:

- (a) Theoretically sound and justified from an ethical perspective because it builds on solid normative and psychological foundations;
- (b) Technically feasible from a designer's perspective because, despite today's technological constraints, robots can efficaciously function as catalysts for self-development and moral flourishing;
- (c) Practically advantageous and desirable from user's perspective (because the benefits of virtuous robots are evident and fulfil the general expectations of social robotics).

This paper aims to discuss the theoretical background of Virtuous Robotics, review the beneficial applications that Virtuous robots can offer to their users, and explain the inherent limitations that, unlike the other two theories, Virtuous Robotics can overcome. The structure of this paper is as follows: in Sect. 1, we discuss the theoretical background of Virtuous Robotics, review the beneficial applications that Virtuous robots can offer to their users. Then we explain the shortcomings of the two competitor theories, Consequentialism and Deontology (Sect. 3), and argue that the design considerations suggested by Virtue theory are preferable (Sect. 4). In the second part, we review several ways to concretely implement the theoretical principles of Virtuous Robotics. First, discussing how badly-designed robots can corrupt humans and promote vicious tendencies (Sect. 5). Then, presenting a number of robotics applications that can, in contrast, encourage virtue cultivation and character development (Sect. 6). Subsequently, we review different roles a robot can take on to achieve these desirable goals (7), identifying the recognitive features that make social robots effective in these moral roles (Sect. 8). In the third part, we review some challenges that Virtuous Robotics has to face, discussing the related theoretical and implementational concerns (Sect. 9) and providing some general answers to them without hiding the unresolved difficulties (Sect. 10). In the Conclusion (Sect. 11), we summarize what makes Virtuous Robotics desirable and promising, offering a conceptual synthesis of our proposal.

2 Understanding Virtuous Robotics: The Theoretical Background

To fully grasp the potential of Virtuous Robotics, a robust understanding of Virtue Ethics and its specificity among the leading approaches to normative ethics is necessary. Despite its authoritative tradition, sustained by millennia-old influential philosophies, Virtue Ethics is still less represented in the HRI discourse than the two mainstream approaches to machine ethics: Consequentialism and Deontology. While antagonizing each other, Consequentialism and Deontology share certain intuitions about normative ethics. Both Consequentialism and Deontology take actions and decisions as primary objects of moral judgement, relying on prescriptive principles to differentiate among the morally right and the wrong ones. According to Consequentialism, a good action/decision is the one that, compared to all the other options available, produces the greatest benefits for the largest number of people, increasing their overall amount of happiness and satisfying the preferences of the majority [32]. According to Deontology, an action or decision is good if it best fulfils the moral agent's duties and obligations while protecting the rights of moral patients and recognizing the absolute, non-negotiable value of their lives.

Consequentialism proposes to evaluate an action's moral value based on the utility created by the action's effects. Deontology, in turn, focuses on the intentions that motivated the action and judges whether or not the agent acted out of the authentic goodwill to fulfill their obligations and respect the others' rights. To illustrate, a lie can be acceptable from a consequentialist perspective when it spares a significant amount of pain and distress, but it is unacceptable from a deontological perspective because it violates the universal obligation of telling the truth. By contrast, withholding private information from police that could lead to the identification of a potential criminal is unacceptable from a consequentialist perspective if sharing that information with the authorities could prevent a devastating crime, however it is necessary from a deontological perspective in respect of the person's fundamental right to privacy.

Virtue Ethics differs from both Consequentialism and Deontology because, unlike them, it does not primarily ask whether *an action* produces desirable effects or is motivated by good intentions, but whether a person deserves praise or blame and whether the kind of life they live is worth living. Virtue Ethics defines a good person as the one who lives a praiseworthy and exemplary lifestyle, i.e., an existence inspired by virtue rather than vice. A person becomes truly virtuous only when their moral norms are internalized, becoming habitual dispositions. Hence, according to Virtue Ethics, being able to live a good life depends more on having good character dispositions than about completing specific actions or following certain abstract rules. Mere

compliance with prescriptive principles and norms is less important than having developed a spontaneous, authentic inclination to behave, feel, and think virtuously.

That is why Virtue Ethics focuses on the admirable or reprehensible dispositions that make a person good or bad, respectively, calling them “virtues” and “vices”, i.e., habitual patterns of behaviour, thought and feeling charged with a positive or negative ethical valence. While the most famous philosophical characterization of virtue is probably the one found in Aristotle’s *Nicomachean Ethics* [9], the general principles and values inspired by the idea of virtue find equivalent expressions in different moral and religious traditions (e.g., “arete” in Greek, “virtus” in Latin, “fadilah” in Arabic).

Examples of universal virtues are courage, generosity, respect, honor, compassion, and integrity. Aristotelian Virtue Ethics construes virtue as the optimal balance between two opposite vices: thus, courage is the virtue that mediates between the cowardice (lack of courage) and recklessness (excess of courage); generosity is the virtue that stands between stinginess (lack of courage) and wastefulness (excess of courage). Consequently, a sense of measure, or moderation (“sophrosyne” in Greek), is a crucial virtue that informs all the other virtues.

Virtues are also defined as positive character traits. Character is one person’s general moral conformation, and it is both the *precondition for* and an *effect of* the consolidation of moral dispositions as habits. Character indicates at once one’s moral and psychological nature. Virtue ethical principles are not based on abstract rules, but concrete considerations about personal realization, psychological well-being, and mental stability. Virtue ethics emphasizes that ethical, positive conduct should be the spontaneous, effortless outcome of fulfilling one’s “second nature”, i.e., the positive dispositions incorporated in one’s character. Hence, character is not essentially innate and certainly is not instinctive. On the contrary, it is formed by exercising oneself, promoting virtuous behaviors and attitudes in oneself and others, and prioritizing the kinds of activities and thought patterns that contribute to realize a commendable lifestyle.

Virtue Ethics holds that both the acquisition and the spontaneous manifestations of such positive inclinations are true marks of moral maturity as they reflect one’s well-developed character and confident agency. Good character emerges from a combination of multiple factors: an educational background conducive to self-respect; good life examples from leading figures; a positive social and familial environment that emphasizes responsibility and agency over narcissism and hedonism. Virtue Ethics particularly underscores the role of rearing and education in character development and the moral value of exemplary actions and role-models: growing up surrounded by inspiring people provides the motivation necessary to imitate their admirable conduct and to devote

oneself to honorable and worthy activities, which is key to flourish and fully express one’s inherent potential.

Virtue Ethics is considered a philosophy of *self-improvement* because embracing a virtuous lifestyle means paying efforts to develop virtues in both oneself (explicitly, through deliberate practice, or implicitly, through habit-formation) and the others (explicitly, through teaching, or implicitly, through example). Virtue Ethics is also a philosophy of *self-fulfillment* and personal accomplishment, because the path of self-improvement, when followed with dedication, is considered the most reliable way to secure deep satisfaction and true happiness. “Eudaimonia”, to use a concept introduced by Greek philosophers, is the virtue ethical idea that having a good character always goes hand in hand with being happy. This approach leads to two insights: first, while the pursuit of personal satisfaction for its own sake is not morally praiseworthy and does necessarily lead to the achievement of satisfaction, conducting a morally praiseworthy existence is intrinsically rewarding and gratifying; second, while pursuing our desires and immediate satisfaction cannot automatically shield us from the negative impact of unhappiness and frustration, virtue cultivation can increase resilience to sadness and pain and mitigate disappointment.

3 Problems with Consequentialist and Deontological Design Ethics

The three normative theories of ethics lead to different approaches to machine ethics, the study of how self-regulated machines can be endowed with ethical behaviour or, more generally, with a moral function [74]. Both Consequentialism and Deontology construe machine ethics as centered on the robot’s performance. They presuppose that ethical performance involves the ability to make the right decisions and that decisions are right when they strictly follow from valid ethical principles. Thus, each theory requires that, in order to act ethically, a robot is sufficiently intelligent to correctly apply to real-life circumstances specific top-down decision-making processes inspired by general ethical principles [44].

The assumption is that the decisional processes undertaken by the robot fully replicate or at least approximate the heuristics that allow humans to select and execute the course of action most appropriate to the circumstances. In this process, a robot assesses all the possible ethical ramifications against a schema of Consequentialist or Deontological principles. For Consequentialism, an ethical decision is supposed to be guided by the comprehension of what humans generally find *preferable and advantageous*, and such comprehension is deemed a prerequisite to select the actions that the users may consider beneficial in a material, cognitive-epistemic, or psychological sense. For Deontology, instead, ethical decision depends on the comprehension of what

humans consider *unfair, unjust, or degrading*, and such comprehension is deemed a prerequisite to prevent actions that constitute an existential threat for the users, violate their legitimate entitlements (including their safety, freedom, dignity, and property), disregard their culture and work, or offend their identity and beliefs.

Unfortunately, we are still far from building robots able to systematically make such judgments. While theoretical machine ethics has never stopped generating fascinating puzzles, applied machine ethics has made very little progress in concretely reproducing human “commonsense”, so far. Skeptical philosophers and moral particularists have claimed that, even when formal models of ethical decision can be consistently represented in abstract terms, they perform poorly when they are used to explain or predict real-life scenarios [33]. When tested in real circumstances, the formal rules that should guide an agent’s ethical decision prove either computationally intractable (i.e., unable to account for the complexity or vagueness of practical circumstances) or just logically implausible (i.e., contradicting their premises). Whether such skepticism is justified or exaggerated, all attempts to build robots capable of making correct consequentialist or deontological decisions had to face structural challenges in real-life scenarios [69].

These challenges are due to the difficulty or impossibility of inferring the right course of action based on general principles. The most general formulation of this conundrum is offered by the Frame Problem: due to their formal nature, general ethical norms are detached from the massive contextual complexity in which ethical decisions are concretely situated. This gap makes it impossible for a machine to infer what information is relevant to the scenario at hand and what way to use such information is appropriate, filtering out in advance all the irrelevant and inappropriate options [18,19]. The Frame Problem can be spelled out in at least three different ways. First, a robot needs fully-fledged folk psychological competences to make sense of how people reason in terms of values, norms, and preferences, because these notions typically presuppose a mentalistic ontology (e.g., beliefs, intentions, and desires) for which, unfortunately, there is no complete formal model yet [13]. Second, even if neural networks can be trained to identify with relative accuracy the emotion on a person’s face, they are still unable to correctly interpret physical events or processes in a moral sense, for example, using sensory cues to recognize a human’s intention and motivation to act in a specific context [48]. Third, the computational limitations of which the system suffers prevents the robot from exhaustively analyse the logic of values and the overwhelming complexity of the real-life scenarios in which values are relevant, as such analyses typically require massively interconnected causal relations, further complicated by fuzzy, holistic, and modal properties [47].

These problems explain why, even when equipped with very accurate sets of instructions and rich heuristics, current artificial robotic agents struggle to understand the possible consequences of an action and the reasons that could have motivated it. Research in artificial reasoning has produced promising models of the logical processes necessary to predict ethically undesirable outcomes from a particular course of action given a specific state of affairs. Such models can, in principle, simulate how a state of affairs or process dynamics should be modified to prevent undesirable events from happening [8,70]. Other decision-making models based on machine learning and crowdsourcing techniques rely on the ethical intuitions of humans, systematically collected through surveys, to generate reliable practical decisions within specific professional domains (e.g., determine priorities in a list of patients waiting for organ transplant, [4,28]).

Although these models seem working effectively in well-circumscribed domains of application, their applicability is limited only to “toy worlds”, artificially simulated scenarios defined by rigid ontologies with limited sets of variables and fixed constraints on the number of degrees of freedom [23]. These worlds represent either exceptional cases of intelligent problem-solving (hardly transferable to other practical contexts) or simplified representations of the complex dilemmas that a situated agent has to face in real-life scenarios (like those captured by decision trees). Many failed attempts to appropriately model real-life situations were caused by having underestimated the exquisitely qualitative and vague nature of moral notions like “advantageous” and “offensive”. In addition to the computational intractability of the concepts presupposed by ethics, limitations of AI in reproducing general human reasoning makes it challenging for a robot to generate decisions that meet the demands of Consequentialism or Deontology.

These limitations are well-known by machine ethicists. Our point is that, if machine ethics has produced only modest results, it is because it has been almost exclusively concerned with Deontological and Consequentialist principles. Other frameworks have been overlooked because of the unwarranted assumption that the only possible strategy to build ethical robots is equipping them with accurate systems for recognizing the actions that a human would have considered useful or right.

4 Virtuous Robotics as the Third Way of Machine Ethics

In the face of these shortcomings, Virtuous Robotics proposes a different direction in machine ethics, assigning to social robots a role associated with very different expectations and requirements. The core assumption is that

human–robot bi-directional interaction, establishing recurrent patterns of reciprocal stimulation between interactants, generate morally significant experiences of self-awareness through which the human can develop their virtues.

We believe that this assumption is more credible than those of Consequentialism and Deontology. Designing ethical robots in accord with the normative principles of Virtue Ethics is more feasible and theoretically sound than trying to implement abstract decision-making protocols based on vague notions like “benefit” and “right”. Why is this the case? It is not that the notions of character development and self-improvement were easier to model and to be processed computationally by an AI. The point is rather that, in order to produce morally desirable interactions between robots and humans, these notions do not need to be modeled computationally at all. Hence, it is irrelevant whether or not the robot’s AI is sufficiently sophisticated to run the relevant computational model in the first place. As we will argue through the rest of the paper, the ethical approach to social robotics based on Virtue Ethics primarily relies on the user’s desire of moral flourishing and self-improvement, not on the robot’s capability to make appropriate decisions.

The proposed shift from robot intelligence to human motivation has two crucial implications concerning what robots can and must do to be considered ethical. First, to the extent they can actively engage the human in a natural process of development and self-improvement, virtuous robots can effectively bypass the limitations that affect the robots designed according to Consequentialist and Deontological prescriptions. Second, what robots must do in order to give a positive moral value to their interactions with humans is simply to use their interactive and embodied features to effectively solicit and reinforce ethical responses in the human interactants during online social engagement.

Therefore, Virtuous Robotics importantly changes the question asked by machine ethics. The emphasis is not on what the robot can or should do *for* the users but how the users can develop themselves and improve their character *through* their interaction with a robot. In order to be virtuous, a robot must be designed to help the humans express their potential, fulfill their good dispositions, and achieve their moral and intellectual maturity to become accomplished persons as citizens, professionals, family members, spouses, or friends.

The robot is not expected to create moral qualities in the user but can facilitate the expression of a preexisting moral potential soliciting self-awareness and encouraging self-improvement and mindful behavior. Also, the robot does not need to store the rich computational models of virtues and vices necessary to make complex moral decisions, because the robot does not need to make such decisions in the first place. The robot does not need to know what makes a behavior virtuous and vicious because it is sufficient that the human user has some fundamental intuition about what virtues and

vices are and is motivated to use such practical knowledge. What is expected from the robot is that it is able to solicit the human user to engage in praiseworthy activities, refrain from unworthy or reprehensible ones, keep an appropriate posture during their everyday life, and persist in their effort to improve their own good qualities and develop valuable skills.

In a Virtuous Robotics framework, the robot’s role is not to decide on behalf of humans which activities and postures deserve to be cultivated, but to remind the users to make such decisions and be consistent with them conscientiously. Robots have to encourage users to persist in their efforts until the chosen activities become habitual, spontaneous dispositions entirely incorporated in everyday lifestyle. Today’s social robots can easily fulfill this function by initiating and maintaining specific interaction loops that, by means of verbal communication, gesture, and emotional expressions, selectively reinforce positive behaviors, thoughts, and feelings in the users.

To sum up, our approach invites both makers and users to accept that the ethical function of robots is not to perform good actions because it is still unclear whether machines are able of true moral agency. On the contrary, the ethical role of robots is to help the users fulfil their moral nature living a deeply satisfying and realized existence and expressing their full potential as humans.

5 The Menace of Vicious Robots

Human Robot Interaction (HRI) researchers need to consider Virtuous Robotics for various reasons. First of all, because social robotics risks producing morally undesirable outcomes if it does not comply with the ethical guidelines recommended by Virtue Ethics. The risk that robots produce reprehensible dispositions in the users has been frequently addressed by robot ethicists [64]. However, compared to other approaches to machine ethics, Virtuous Robotics is better positioned to account for the fact that, even if a robot does not commit any wrong decision or inappropriate action, that does not imply that that robot is a morally neutral entity. Its interactive function and role, in addition to its very presence and aspect, may explicitly or indirectly solicit reprehensible actions and thoughts in the human users or bystanders, or push them to tolerate and normalize such actions and thoughts in others.

This kind of concern has been discussed in relation to applications of social robotics considered potentially corruptive, for example, when the nature of the bond established between humans and robots is deemed equivocal or inappropriate, as suggested by the controversial case of intimate robots [62]. The worry in this and other cases is motivated by the consideration that HRI could become a source of vicious

behaviours and biased attitudes in the human if it were not conscientiously designed. Ethicists have repeatedly warned that prolonged or repeated exposure to certain robots could have a bad influence on some kinds of human behaviors. Consider the following examples:

- (a) *De-humanizing and unempathic attitudes*: insofar as the human-robot relationship tends to mimic a master-slave relationship, robots can reinforce authoritarian or even cruel stances in humans, rousing the user's sense of entitlement when they give orders to robots or exploit their work [63]. Children spoiled by robot nannies incapable of saying "no" are a frequently cited example [15]. As it is documented, these children tend to find the idea of bullying robots attractive [50]; their aggressive, arrogant, or self-indulgent tendencies could be reinforced by the robot's inability to escape or defend themselves.
- (b) *Incitement to leniency*: as robotic servants allow humans to indulge in any kind of pleasurable and enjoyable activities, including distasteful, unhealthy, morally reprehensible, or blatantly illicit practices, robots could become vehicles of addiction and disinhibit maniacal behaviors. Especially in the individuals who already feel compelled to perform such activities, including sexual perversion, physical torture, and other kinds of activities that would be considered abusive or deviant if they targeted a human or a sentient animal. Furthermore, robots should prevent humans from overusing applications that do not directly contribute to their (the user's) well-being [57], especially if the use of such applications become compulsive and benefits only the robot's maker [58].
- (c) *Bias reinforcement*: as humanoid robots have anthropomorphic features, they convey the representation of humans posing or acting in specific roles or functions. Whether purposely or unintentionally, robot design can propagate and contribute to normalize inappropriate representations by portraying gender [25], race [10], social standing, or disability with offensive or inappropriate connotations.
- (d) *Moral deskilling*: as critical decisions are more and more frequently offloaded to AI, continuous reliance on autonomous agents might progressively damage professional expertise. This situation can possibly make humans unready to deal with complex, problematic, or potentially controversial circumstances, and make them unwilling to take on challenging tasks that require to be handled with moral acumen, practical wisdom, and sensitivity. These virtues might be impoverished if humans are not used anymore to face such situations without the supervision or guidance of a machine [66].

Acknowledging these risks, the designers are motivated to implement solutions that allow the robots to passively discourage or pro-actively prevent and correct the happening of unacceptable conducts, attitudes, and associations.

However, a consistent R&D program in Virtuous Robotics cannot focus only on eliminating risks and negative traits. While the preventive and corrective function is essential to address the moral risks implied by HRI, Virtuous Robotics also envisions that social robots can have a constructive potential for moral development. If social robots have the power to encourage bad habits in their human users, it is reasonable to expect that they could also help the users correct their bad habits and promote the good ones.

Given these premises, it seems inevitable that robots can have at least a mild moral impact on human users, the extent of which may be difficult to predict. This moral impact implies that Virtuous Robotics is not only legitimately justified, morally desirable, and achievable but also necessary. Encouraging moral development is an ethical imperative no less than preventing moral corruption, therefore creating virtuous robots should be a priority in the same way the creation of vicious robots should be discouraged. If this argument is correct, in addition to the utilitarian benefits and the deontological assurances that can be offered by robots, we must also acknowledge that, when they are responsibly designed, robots constitute—more than other interactive technologies—a powerful force at the service of moral cultivation and ethical awareness.

6 Applications of Virtuous Robotics

This section reviews how robots may exert a virtuous effect on their users. Assuming that a user has both the motivational and intellectual resources necessary to start a journey of self-improvement and character development, we maintain that a robot can efficaciously accompany the user through this pathway. How can a robot support this kind of healthy transformative process? First, as an example, let us consider the robot tortoise Shelly, built by KAIST team to teach gentle manners and patience to children [41] (Fig. 1).

Shelly is fun to play with, unless you smack it, at which point it hides inside its shell until it's safe to come out again. Shelly is designed to be large enough that five to seven children (under 13 years old or so) can interact with it simultaneously. [] Using its LEDs and limbs, Shelly conveys different emotional states, including happy, sulky, angry, and frightened. The frightened behavior gets triggered when a kid hits, kicks, or lifts the robot, and when that happens, Shelly retracts inside its shell and stays there for 14s. Before playing with Shelly, children are explicitly told that if they abuse



Fig. 1 When Shelly is “happy”, bright colors illuminate its shell [left]. When it’s “sad”, it hides its head and disables the lights for 14 s [right]. (Naver Labs). (Color figure online)

the robot, it will get scared and hide for a little bit. [...] Results showed that Shelly’s hiding technique was able to significantly reduce children’s abusive behavior, relative to how they acted when the robot didn’t hide at all. When they tried reducing the hiding length from 14 s to 7, abuse actually increased, because the hiding behavior itself was seen as a reward. And a longer hiding length of 28 s caused the kids to get bored and leave, defeating the purpose of the robot. Also interesting is that part of the effectiveness comes from the fact that in groups, children will mutually restrain inappropriate behaviours [1].

To an extent, Shelly has successfully achieved the purpose for which it was designed: teach children to refrain from their violent and abusive tendencies and make them learn a more gentle and careful way to play. This study is short-term and cannot tell us if prolonged and repeated interactions with Shelly could eventually lead to a permanent consolidation of gentle and considerate behaviours in the children. Also, the study was not designed to measure whether the empathic responses of the children toward the robot actually increased. However, from a Virtuous Robotics perspective, it is significant that interaction with the robot tortoise succeeded at least in offering to children an occasion of self-observation, soliciting them to reflect on their conduct and inhibit their violent pulses. This inhibition was reinforced by modulating the time of the head-retraction behaviour in case of abuse in a way that discouraged harassing behaviors without disrupting the playful engagement with the tortoise. This self-observation helped children bring to collective awareness a normative judgment (verbally communicated by some children to the others) about the negative valence of violence and the worth of gentleness and empathy in inter-subjective relationships.

Shelley is not the only positive instantiation of Virtuous Robotics. Other applications can be conceived to target different moral tendencies in diverse categories of users for different sets of reasons. Assuming that the purpose of Virtuous Robotics is to support self-development and character cultivation, we consider “virtuous” any robot designed to

help its user achieve at least one of the following ten objectives (that social robotics is still exploring):

- (a) *Keeping track of and reinforcing motivation and pride in athletic training, academic study, and artistic performance*: extensive research has been done to explore the possibilities of robot technologies in the classroom for particular subjects such as STEM skills, languages, socialization, among others. However, as is discussed in [54], the teachers’ attitudes towards robots are still generally negative, with the exception of their use in STEM disciplines. Furthermore, teachers point out the motivational potential of robots as an information source or easy handling. Their concerns are more related to the effect on their teaching workloads and the replacement of interpersonal relationships and disruption of the teaching process. Despite the challenges present in the implementation of robots as teachers, it is generally accepted that robots can be instrumental to the cultivation of certain attitudes and values in young humans. These robots also guide users through their search for the most valuable and meaningful activities such as: facilitate learning, choosing a study plan, select a reading, mentoring during a trip to find personal relations and deep satisfaction. For example, the work presented by Obaid et al. [52] shows how utilizing technologies to empower robots with artificial empathy may facilitate learning.
- (b) *Acquire or perfect broad skill sets that are relevant to self-development (acting, meditation, writing, logical problem-solving)*: this includes also increasing innate cognitive/social abilities that are important for life, like joint attention and social coordination. In this case, children tend to have positive attitudes towards robots (especially those with a cartoonish and cute look), but also over-expectations about the state-of-the-art of robot technologies as Lin et al. [43] suggest. Certainly, interactive skills in future robots need to be improved to meet these expectations and consider other actors involved in the cultivation of virtues in young people (parents, principals, education systems) and the personal circumstances of the individuals (e.g., demographics, socioeconomics, environment, health). Further discussion of robots in educative environments and their ethical dilemmas have been offered by [61]. They found that students, in general, have positive attitudes towards robots, but they do not approve agency in robots for critical decision making (e.g., grades, discipline issues, controversies). Furthermore, they do not approve early human-robot interaction in kindergarten and other early stages of human socialization. As we can see, further development of robots is critical to change these perceptions and accept robots as virtue cultivators.

- (c) *Recognize, inhibit, and subdue negative character traits:* arrogance, impatience, anger, rudeness [24], bad habits and addictions (smoke, drugs), biases and prejudices (racism, sexism, elitism), criminal tendencies and corruption [56]. We know that designers may have a responsibility when robots or other interactive technologies convey or reinforce prejudices, even when this happens unintentionally. For example, when Artificial Vision systems are trained with too narrowly selected and culture-specific data, or when the training of public conversational avatars is unsupervised and relies on interaction with random users, ending up replicating their racist or hate speech. Early experimental research on racism has shown that subjects holding racist prejudices and biases toward human races tend to have the same prejudices and biases toward racially-characterized robots [2,10,38,65]. However, a study [53] claims that the evaluation of the robots is more often “based on functionality rather than the perpetuation of existing human stereotypes” suggesting that prejudices toward humans do not easily transfer to robots, as robots are perceived as neutral.
- (d) *Develop or reinforce positive character traits,* like courage, compassion, charity [12] (Borenstein and Arkin 2017), honesty [27,37], and positive ways of thinking, like hope and enthusiasm.
- (e) *Encourage one’s search for excellence and merit. Identify and foster vocational talents and professional inclinations:* (for music, sport, public engagement), also through the acquisition of specific skills (second language, math, music composition).
- (f) *Manage their inner conflicts and deal with their psychological challenges or limitations:* these include fears, intolerances, addictions, post-traumatic stress disorders and structural impairments, see [40]. This raises discussions about the role of robots in coaching: should robots be designed with the authority necessary to keep us motivated and guide us through every complex life choice, or should they be employed only as tools and proxies to mediate human–human interaction? [22]. We need to establish consistent roles for the robots as companions according to the degree of intimacy and trust required by the users: friends, assistant, or butler, considering the specific needs associated with the context. Certainly, self-development applications present additional challenges as the function of social robots is not simply instrumental (achieving a particular practical goal), but involves also a holistic, or existential, aspect (reaching authentic happiness and deep fulfillment).
- (g) *Reinforce personality traits related to self-regulation, discipline, restraint,* including resilience, focus, concentration, mental toughness, emotional awareness and control. Robot invigilators deployed in classrooms during exams to prevent cheating and other misbehaviors represent an example [49]. Similarly, in robotics applications for security and safety, physicality (i.e., embodied, material presence) is a crucial feature of dissuasion. Despite significant commercial developments in security applications of service robotics [29], little research can be found in the academic literature about the physicality of robotic security guards and whether this relates to the robot’s perceived moral authority.
- (h) *Advise and coach using exercises and practices that favor their physical and mental well-being:* meditation, workouts, prayer, counselling [26,35], healthy lifestyle (diet and sleep routine—see [76]), and good habits like personal hygiene and care of self. In this area, we can find varying philosophies to technology design for assisting humans in building better versions of themselves over the years. As the industry of coaching and self-satisfaction is remunerative, it is expected that investments to service robotics will increase. Similar to other technologies such as exertion games and virtual reality, robots can foster physical exertion due to the intrinsic functionalities of their embodiments. For instance, humanoid robots with anthropomorphic embodiments have shown positive results as physical trainers in yoga or low impact physical activity [6,7]. The robot embodiment plays a central role in the interaction, providing physical demos about what is expected from the trainees. Additionally, those social robots have social and communication skills to provide feedback and help the trainees reach the goals of each exercise session.
- (i) *Foster reflective attitude, independent judgment, and emotional maturity (autonomy in dealing with personal challenges, professional competitions, romantic relationships).* Digital avatars in the role of counsellor have been used with US Navy veterans and obtained positive results in the treatment of post-traumatic stress disorders, leveraging the perceived impersonality and non-invasiveness of the artificial agent in comparison to a human counsellor [3]. Also, several studies have explored the role of zoomorphic and anthropomorphic robots as a therapeutic aids to exercise and control emotional expression; e.g., the studies with the robot Paro [17]. Work in this area has recognized the importance of designing robots that share the experiential background of their human users. That is why Hughes and Benoit [39] have patented a “Culturally adaptive avatar simulator”, an interactive system that allows the user to select among multiple features (gender, race, and ethnically identifiable anthropomorphic forms) to match the profile of the end-user, using audio and visual communication for users requiring support and counselling.

Each of these objectives can be implemented through specifically designed interactive tasks and goal-oriented

activities, which may take different forms. For instance: journal updating, counselling session, interactive game, role playing, simulation or exercise, training routine, test, competitive challenge, or unstructured creative expression sessions. Task choice, in turn, depends on the context in which the robot is deployed: personal exploration, education, professional training, social networking, counselling, therapy, rehabilitation, monitoring and control, artistic performance, invention and innovation.

A robot can use a broad range of behavioral responses and communicative signals to solicit conscientiousness and awareness in the users and help them realize the severity of their actions. The robot should tailor such responses to the gravity of the situation, selecting the most appropriate actions given the social and educational context. Some suggested actions are alert cues, verbal notifications, exhortations, implicit or explicit threats of informing a human authority (counsellor, instructor, parents, clinician, police). Emotionally loaded communicative behaviors associated with particular contexts (screaming, crying, begging, laughing, acting frighteningly, or ironic or sarcastic comments to highlight inappropriate conduct) can be leveraged to solicit a stronger response. In exceptional cases, justified by emergencies, even cautious physical interventions could, at least in principle, have a restraining and preventive function.

7 Social Roles and Modes of Application

What forms of moralizing intervention have to be undertaken by virtuous robots to encourage a considerate response in the human user? We can envision numerous roles that a robot can take on to achieve this goal, including but not exhausted by the following list.

- (a) *Interactive diary and confidant*: the robot invites the human user to narrate their everyday experiences and reflect on their own conduct. Through sentiment analysis techniques based on the statistical occurrence of significant keywords, the robot can select the most relevant questions to help the user identify trends and patterns in their path of moral development, suggesting corrections or providing encouragement when necessary. This role replicates the function of chatbot counsellors [11,71] adding to it the embodied and interactive feature of a mobile robot.
- (b) *Assistant, coach, and motivator*: following a predefined timetable set by a preparator, a therapist, or an instructor, the robot reminds the user of daily routines that can be part of an educational, training, or clinical program. The robot supervises the execution of exercises and activities required by this program, which can be physical, intellectual, or social in nature, and produces evaluations of

the user's performance collecting data from her real-time feedback. This way, the robot can monitor changes in the emotional and motivational condition of the user and intervene with adequate support or motivational interventions or alerting a human supervisor as required [35].

- (c) *Apprentice*: when a robot behaves appropriately or improperly or suggests its intention to do it, it solicits the corrective responses of a human user, who is then expected to act as a tutor, praising and blaming the robot according to its behavior [55]. If the human does not intervene to correct the robot, the robot will illustrate the negative consequences of indulging in a vicious behavior. If the human does intervene, for example, disciplining the robot or suggesting a better course of action, the robot will reward the user illustrating the positive implications of embracing a virtuous behavior. This response should be tailored and timed to solicit a reflection on the values associated with correct behavior.
- (d) *Informer or "scarecrow"*: although robots are not morally competent, they can be designed to monitor the user and supervise their activities on behalf of moral authorities accepted by the users (parents, supervisors, police), becoming their ears, eyes, and hands. Robots can promote virtuous behaviors and deter vicious ones simply with their presence as bystanders. Also, they can threaten humans to record their actions and inform the authorities. This kind of application requires that the robot's conduct complies with strict confidentiality protocols, to make sure that the robot is not used to collect private information against the user's will or without her awareness.
- (e) *Moodsetter*: as demonstrated by the robustness of the chameleon effect [21], kindness and gentle manners are contagious. That is why an agent that frequently performs small acts of kindness will create a positive atmosphere in which people are more inclined to interact socially and to work enthusiastically. A robot can be deliberately deployed in a work or family environment to provide positive examples of good behavior and positive thinking that inspire the surrounding people.

8 The Moralizing Function of Robots: Objectification

Having described the general applications of Virtuous Robotics, we need to discuss the specific characteristics of robots that make this approach to ethical design possible and sound: what features make a robot suitable for cultivating virtues and supporting self-improvement in humans? Are robots more suitable to exert a moralizing function than other technological devices?

The idea that social robots could be vehicles of moral flourishing and provide prime opportunities for virtue cultivation might sound extravagant to those who see robots only as technological tools serving a merely instrumental function [14]. From a purely utilitarian perspective, it is not clear why robots, unlike other inanimate pieces of technology, could or should be endowed with a formative function or moral valence. After all, robots, like any other technological product, have been created for a practical purpose: they exist to serve the human user, not guide them or inspire their moral maturation. Why should a non-sentient, typically unintelligent tool, unaware of the ethical value of their actions, become an opportunity of moral development? If robots can play such a crucial role, then virtually any piece of technology should be able to do the same.

This conclusion is not warranted as it does not consider some key aspects of the fundamental relation that bonds humans to technology in general and social robots in particular. First of all, as highlighted by recent developments in philosophy of technology inspired by virtue theory [67], it is necessary to recognize that all pieces of technological equipment, and not only robots, may have implications for character development and virtue/vice cultivation. This is because, in shaping people's practices and beliefs, new technologies promote the formation of implicit beliefs and habits charged with relevant ethical connotations. The moral landscape of human civilization is shaped by the practices shared by a community, which, in turn, are importantly enabled or constrained by technology.

Also, it is indispensable to mention the specificity of social robots among all forms of technology. Unlike other kinds of machines, designed simply to serve as tools, social robots exist to replicate the human cognitive and physical endowments, which is why robots can engage humans in pseudo-social relationships that affect, implicitly or explicitly, their emotions and beliefs. The fact that robots are mirrors of the human nature justify their exemplar and moralizing function in a Virtue Ethics perspective. Three characteristics make social robots unique:

1. *Recognitive technology*: unlike other technologies, the distinctive function of social robots consists of detecting social cues produced by humans and produce response cues that can be interpreted socially by humans [20]
2. *Non-trivially embodied*: social robots are not merely material instantiations or extensions of a computer. The cognitive and social powers of social robots are constituted by the experiential and interactive opportunities offered by the specific details of their incorporation.
3. *Reciprocal coupling*: insofar as they are embodied and recognitive entities, social robots can establish quasi-

reciprocal relations with humans in a way that mimics the mutuality of human social relationships [42,59]. A mutuality relation is one in which each agent is dynamically coupled with the other. So, each of them is both analogous to the others (like a mirror) and able to affect and be affected by the other (through interaction).

Robots are advantaged in promoting virtue cultivation because they, unlike other technologies, have an explicitly android or animaloid body. These embodiments possibly include a face and other communicative features like gaze, touch, and gesture. Unlike virtual assistants, which are confined to tablets and mobile phones, robots can manipulate the physical environment of the users, sharing with them a world of communal affordances. The fact that robots are designed to establish reciprocity, analogy, and shared experiences with humans contributes to creating human-robot relations that are not merely instrumental. Non-instrumental relationships are those in which the related individuals are interested in the purpose and meaning of the relationship itself, so that the relationship is inevitably also about defining its own meaning. Such relationships solicit social awareness, even when the other element in the couple is just a robot. The reciprocity of such relationships implies that human interactants perceive themselves not only *as agents* using an instrument but also *as objects* of evaluation and consideration for another agent. This phenomenon is crucial to understand the moralizing function of robots from a perspective of Virtuous Robotics.

Robots do not simply exist to execute our orders. They can also detect us, focus their attention on us, observe us, make us objects of their evaluations and linguistic representations, or targets of their goal-oriented behaviours. Not only we tend to see robots as social others, but they also encourage us to feel that we are social others to them. The *objectifying function* of social robots, enabled by their embodiment and their ability to offer reciprocal interactivity, crucially defines their possibility to play a constructive moral role in the lives of their users: according to philosophers like Sartre [51,60,77] becoming an object standing in front of the other's gaze is a phenomenological precondition to develop shame and related moral sentiments like regret, remorse, and guilt, which are key for the growth of an ethical conscience.

Robot embodiment is important not only because it can facilitate engagement and strengthen interest in humans [31], but also because it allows robots to share with the users an environment in which they are physically situated and reciprocally coupled through recurrent sensorimotor loops. Other technologies, like mobile phones and tablets, cannot bodily engage with their environment, hence they cannot inhabit the same space of opportunities of action and interpretation that is shared by humans and robotic agents.

9 Against Virtuous Robotics: Practical and Theoretical Objections

We have already explained why Virtuous Robotics is a technically feasible and theoretically sound option in the domain of machine ethics. Nonetheless, it is possible to anticipate that our proposal could attract a number of objections and challenges. We will consider seven problematic issues. We will not attempt to give separate solutions to each of them, postponing to a different occasion a more systematic discussion. The first five issues criticize the prospect of implementing the principles of Virtue Ethics in social robots: these objections contend that robots may lack some of the attributes indispensable to virtue cultivation. The other two issues question the general validity of Virtue Ethics as a guide to moral action, a topic generally discussed among ethicists of different persuasion.

- (i) *The Methodological issues*: to know whether robots can facilitate character building, a scientist should be able to assess whether the interaction with a robot permanently modifies the user's character. However, assessing character changes (whether vicious or virtuous) needs complex longitudinal studies informed by rigorous qualitative protocols. Part of the problem is that longitudinal studies are difficult to administer as they require particularly demanding preparation and setting. For example, to assure ecological validity, it is necessary that all the subjects feel familiar with the robot, having had similar interactions with it. Another methodological issue is the intrinsic difficulty of measuring virtues and vices. Unlike other kinds of behavioural changes, which are immediately observable and quantifiable, virtues and vices may not have any explicit symptomatic expressions, as they are by definition implicit inclinations.
- (ii) *Lack of Ecological Validity/Generalizability*: this is the worry that the behavioural patterns solicited by robots may not transfer to other interactive contexts. In other words, the outcome of an interaction with a robot may not truly represent what the human subject would do during equivalent interactions with other humans [45]. Even if robots were hyper-realistic, the simple awareness that interaction involves an artificial non-sentient agent might prevent certain users from acting and responding spontaneously. Hence, even if humans responded positively to the stimuli presented by the robot, expressing a virtuous disposition (for example, responding politely to the salutation of a robot), this might be simply a pantomime, a simulated behaviour that does not correspond to a real process of moral maturation.
- (iii) *Automaticity versus Disposition*: the third issue is that robots may not have the social or intellectual attributes necessary to motivate a true change in the user's character. Certainly, robots can produce repetitive action routines and invite humans to engage in such routines (for example, execute a gymnastic exercise). This repetitive experience may help humans acquire simple habits and automatize mechanical responses, which makes robots good coaches for training sets of mindless, repetitive behaviors. But such responses “says the objection” are at best superficial, blind reflexes, and cannot embed the kind of moral awareness that is necessary to develop a true virtue. Forming a virtuous disposition and developing muscle memory through repetitive exercise are incomparable kinds of learning, and unfortunately, a robot can support only the latter, not the former.
- (iv) *Insufficient authority issue*: the fourth issue is that robots might not be sufficiently inspiring to motivate the kind of deep character modification that is required for moral self-improvement. When vicious humans decide to abandon a bad habit or a vicious lifestyle, redeeming themselves, it is often because they were impressed by the discourses or examples of very authoritative people. However, robots may completely lack this kind of moral authority as they are irremediably perceived by humans as non-sentient instruments playing only subordinated roles [72]. In such a limited capacity, robots might be unable to provide useful moral examples and their moral remarks might be quickly dismissed [75].
- (v) *Attachment issues*: the fifth worry is somehow symmetrical to the fourth. A beloved person has much better chances to motivate a positive change in our lifestyle than an unknown stranger. Similarly, to effectively change the character of a human, a robot should have established with the users a bond sufficiently strong to exert some influence on them. But robots rarely have this kind of intimate familiarity with their users and, if they do “says the objection” the attachment perceived by the user is just the effect of a fetishist projection or delusion. Such fake attachment is likely to be more detrimental to the moral quality of their relationship than the opposite.
- (vi) *Inexistence of Character*: the sixth worry targets one of the critical assumptions of Virtue theory per se, independently of its possible applications in HRI. Some social psychologists have claimed that character does not exist, and certainly does not exist as a universal motivational scaffold for ethical action and decision [46]. Their experiments have shown that moral choice and ethical decision correlate much more to situational contingencies than to the objective robustness of individual character. If their interpretation of the experimental results is correct, it is not true that some people are more virtuous than others. Differences in behaviour depend mostly on the circumstances in which humans operate and on the particular context of their choices [36]. If that is true, it is pointless to try to moralize humans with the help of social robots

designed in accord with the principles of Virtue Ethics, because Virtue ethics is wrong at root: moral choice does not depend on individual dispositions built in one's character. Therefore it is pointless to try to correct or improve such dispositions in the first place.

- (vii) *Relativity of Virtue*: the seventh issue, like the sixth, targets virtue theory in itself. The practical applicability of the notion of virtue, claims the objection, is undermined by its vagueness and cultural relativity. It is impossible to define virtues universally: boasting about one's achievements can be considered a vice (immodesty) in a conservative, traditional context and a virtue (determination) in a mundane, competitive professional context. The moral value and the broader significance attributed to a particular character disposition varies significantly among different communities and societies, not to speak of different individuals, who may have completely different perceptions about where the right balance of two opposite vices is precisely situated. Appealing to human nature does not make the task of identifying universal virtues any easier as there is no agreement about the normative features of human nature, and there are no reasons to attribute a normative value to the human genetic code, which is the unintentional result of casual variations and blind adaptations.

10 Refuting the Objections: Virtuous Robots as Moral Mirrors

We recognize that these issues capture some real challenges for HRI. Each of them deserves a thorough discussion and a specific set of analyses. The methodological issues [(ii) and (iii)] are intensely discussed in social robotics research. They have implications for the whole discipline, not only our proposal. Hence they exceed the scope of this paper. Instead of answering analytically all the objections, we prefer to clarify the moralizing role played by virtuous robots, as this should be sufficient to sketch a general response to the seven objections. We need three preliminary considerations.

First, to date, the evidence corroborating some of these worries [(ii)–(v)] is still scarce and prevalently anecdotal, hence their validity will be uncertain until ad hoc experiments collect the relevant empirical data.

Second, the general objections against Virtue Ethics [(vi) and (vii)] emphasize the fluidity, particularity, relativity of character and virtue over their stability, consistency, and universality. However, these objections cannot undermine our assumption that positive dispositions consolidate when one follows a supervised experiential path of self-improvement and personal development. Such a way provides the formative scaffold necessary to develop decision-making and

critical thinking skills that define one's personality and lifestyle.

Third, some of these objections have only limited relevance to our proposal. For example, (iii) (lack of moral authority) and (v) (absence of a strong bond) cannot undermine the moralizing function that we assign to virtuous robots, which is unchanged even if the user perceives robots as moral authorities or intimate companions.

Our proposal is guided by the key idea that often the only difference between a virtuous person and a vicious one is that only the former was lucky enough to have the support of a friendly voice that gently reminded her/him the moral implications of their actions during challenging times. Most often we are corrupted because nobody is present to help us regain our vigilant awareness and our self-control in the moment of greater moral confusion. It is not a lack of knowledge or understanding of right and wrong that motivates our descent into vice, but a suspension or disruption of our self-awareness.

Robot companions can help users develop or consolidate their virtue primarily because they can effectively help humans pay attention to their shortcomings. Also, robots can help humans recognize their own flaws and actively contrast them, supporting the reflective consciousness and self-control necessary to make the right choice and refrain from vicious behaviours. Interaction with robots, whether based on bodily or verbal feedback, is not expected to blindly produce positive dispositions in the user because that is not how virtues are formed in the first place. How can a robot successfully promote character formation and self-development then? To answer this question, we need to clear the stage from two possible misunderstandings.

Virtuous Robotics claims neither that a robot could do this by mechanically triggering specific behavioural reactions in the user nor that a robot could or should persuade the user of the positive moral value of a particular behaviour or lifestyle. In this sense, Virtuous Robotics distinguishes itself from the Persuasive Robots paradigm, which aims to build robots designed to convince the user [30,34,68]. A robot can indeed be programmed to remind the user what actions are right or wrong and draw their attention to the good or bad consequences of their decisions, but it could hardly change anybody's mind or implant entirely extraneous beliefs. In other words, we assume that if a robot can promote authentic moral transformation in the user, it is not because the robot managed to *persuade* the user, modifying their convictions, but because the robot was able to motivate a genuine *reflection* in the user, creating a moment of conscience and self-awareness. These critical interactive moments are crucial to sustaining motivation, resilience, determination, and focus on the user's goals.

Robots can be particularly useful in this moralizing role as they are constantly present, fully available, uncompromis-

ingly committed and loyal, infinitely patient, non-judgmental and friendly, non-oppressive, and perfectly transparent and objective in providing feedback. These qualities make robots ideal supervisors and assistants when humans must be reminded that they are indulging in bad habits (smoke, alcoholism), warned of possible risks and negative consequences associated with their actions (careless drive style, impolite language), or made realize that their ways of acting and speaking are inappropriate (abusive and violent behaviours). We believe this fundamental idea dissolves the doubts about the role played by robots in virtue cultivation [objections (iii), (iv), and (v)], because it clarifies what the moral function of a robot should be according to Virtue Ethics.

This point deserves to be further clarified, highlighting the role of the robot as a catalyst. It is evident that a non-sentient, non-conscious robot is not a person and should not be seen as a moral authority (a guru or a wise confidant) because we could never trust the advice of a robot regarding what we should or should not do in our lives [5]. Even if a robot was genuinely able to help people reach their deep existential and spiritual fulfillment, that robot certainly should not be seen as the *source* or the *aim* of such realization in the existential sphere. Such a robot should rather be seen as a *supporter*, a *facilitator*, or an *amplifier* to the agent's existing tendency to self-development. Like a good book, a well-structured training course, or a conversation with a friend, also the consultation with a robot can offer a legitimate scaffold to one's pulse to self-betterment and moral elevation, although it certainly cannot create such pulse from scratch if it does not exist.

As the robot is only a medium to channel and amplify existing pulses to self-improvement and betterment, Virtuous Robotics requires that the user is sincerely willing to cooperate with the social robot and motivated to go through the process of maturation accepting the robot's help. This process can be challenging, effortful, or even discomforting, but the robot is there to support the user. As a trainer, the robot can help its trainee only to the extent the trainee has already decided to let the trainer help her/him. Hence, we assume that a human that appeals to a robot to cultivate their virtues already has at least a general intuition of what self-development is and why it is desirable. If these premises hold, then objections (iii), (iv), and (v) cannot invalidate our proposal because they are not relevant to it.

11 Conclusion

Although our analysis was not meant to exhaust all the possible implications of our proposal, we believe it is sufficient to convincingly present ample reasons to appreciate the uniqueness and the feasibility of Virtuous Robotics. From an ethical perspective, we consider our proposal theoretically robust

because it offers a new way to assess the moral value of robot design, based on the assistive function of HRI rather than the robot's moral agency. From a designer's perspective, it is technically feasible because it moves the cognitive burden of HRI from the robot to the user, bypassing any need for AIs with complex decision-making abilities. From the user's perspective, it is practically advantageous, because it envisions a realistic way to make robots concretely useful and desirable in our lives, as supports for personal betterment and fulfillment.

We do not suggest that robots designed according to Virtue Ethics are universally preferable to the robots designed using consequentialist or deontological principles. None of these three theories seems able to account for the totality of our moral intuitions. Neither we claim that Virtuous Robotics can definitively solve all the technical challenges and the moral dilemmas faced by machine ethics, as we expect our proposal will have to meet its peculiar dilemmas and challenges. We would like to highlight that an alternative approach to machine ethics exists, and promises feasible and advantageous opportunities for designing and using ethical robots. This approach is sound, robust, and deserves careful consideration, despite having been underappreciated until now.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Funding This research was supported by a DHRG Seedcorn Funding Grant awarded to Omar Mubin and Massimiliano L. Cappuccio by the School of Digital Humanities of Western Sydney University on 19/5/2018.

References

1. Ackerman E (2018) Robotic tortoise helps kids to learn that robot abuse is a bad thing—IEEE Spectrum. Library Catalog: spectrum.ieee.org. <https://spectrum.ieee.org/automaton/robotics/robotics-hardware/shelly-robotic-tortoise-helps-kids-learn-that-robot-abuse-is-a-bad-thing>
2. Addison A, Bartneck C, Yogeewaran K (2019) Robots can be more than Black and White: examining racial bias towards robots. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. ACM, pp 493–498
3. Albright G, Goldman R, Shockley KM, McDevitt F, Akabas S (2012) Using an avatar-based simulation to train families to motivate veterans with post-deployment stress to seek help at the va. Games Health: Res Dev Clin Appl 1(1):21–28
4. Anderson SL (2008) Asimov's three laws of robotics and machine metaethics. AI Soc 22(4):477–493
5. Aroyo A, Kyohei T, Koyama T, Takahashi H, Rea F, Sciutti A, Yoshikawa Y, Ishiguro H, Sandini G (2018) Will people morally crack under the authority of a famous wicked robot? In: 2018 27th IEEE international symposium on robot and human inter-

- active communication (RO-MAN), pp 35–42. <https://doi.org/10.1109/ROMAN.2018.8525744>. ISSN: 1944-9437
6. author N. Selected RoboticLab projects | iCampus Wildau. <https://icampus.th-wildau.de/icampus/home/en/selected-roboticlab-projects-0>
 7. author N. This Little Robot Will Teach You Yoga | Digital Trends. <https://www.digitaltrends.com/cool-tech/alpha-2-robot-crowdfunding-news/>
 8. Bacchus F, Kabanza F (2000) Using temporal logics to express search control knowledge for planning. *Artif Intell* 116(1–2):123–191
 9. Bartlett RC, Collins SD et al (2011) Aristotle's Nicomachean ethics. University of Chicago Press, Chicago
 10. Bartneck C, Yogeewaran K, Ser QM, Woodward G, Sparrow R, Wang S, Eyssel F (2018) Robots and racism. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction, HRI '18. Association for Computing Machinery, Chicago, pp 196–204. <https://doi.org/10.1145/3171221.3171260>
 11. Bassett C (2019) The computational therapeutic: exploring Weizenbaums ELIZA as a history of the present. *AI Soc* 34(4):803–812
 12. Borenstein J, Arkin RC (2017) Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. *AI Soc* 32(4):499–507. <https://doi.org/10.1007/s00146-016-0684-1>
 13. Briggs G (2012) Machine ethics, the frame problem, and theory of mind. In: Proceedings of the AISB/IACAP world congress
 14. Bryson JJ (2010) Robots should be slaves. In: Close engagements with artificial companions: key social, psychological, ethical and design issues, pp 63–74
 15. Bryson JJ (2010) Why robot nannies probably won't do much psychological damage. *Interact Stud* 11(2):196–200. <https://doi.org/10.1075/is.11.2.03bry>
 16. Burton E, Goldsmith J, Koenig S, Kuipers B, Mattei N, Walsh T (2017) Ethical considerations in artificial intelligence courses. *AI Mag* 38(2):22–34. <https://doi.org/10.1609/aimag.v38i2.2731>
 17. Calo CJ, Hunt-Bull N, Lewis L, Metzler T (2011) Ethical implications of using the paro robot, with a focus on dementia patient care. In: Workshops at the twenty-fifth AAAI conference on artificial intelligence
 18. Cappuccio M, Wheeler M (2011) The sign of the hand: symbolic practices and the extended mind. *Versus* 113:33–56
 19. Cappuccio M, Wheeler M (2012) Ground-level intelligence: inter-context frame problem and dynamics of the background. In: *Knowing without thinking. Mind, action, cognition and the phenomenon of the background*. Palgrave Macmillan, London
 20. Cappuccio ML, Peeters A, McDonald W (2019) Sympathy for Dolores: moral consideration for robots based on virtue and recognition. *Philos Technol* 1–23
 21. Chartrand TL, Bargh JA (1999) The chameleon effect: the perception-behavior link and social interaction. *J Personal Soc Psychol* 76(6):893
 22. Dautenhahn K, Woods S, Kaouri C, Walters M, Koay KL, Werry I (2005) What is a robot companion—friend, assistant or butler? In: 2005 IEEE/RSJ international conference on intelligent robots and systems, pp 1192–1197. <https://doi.org/10.1109/IROS.2005.1545189>. ISSN: 2153-0866
 23. Deng B (2015) Machine ethics: the robot's dilemma. *Nat News* 523(7558):24
 24. Draper H, Sorell T (2014) Using robots to modify demanding or impolite behavior of older people. In: Beetz M, Johnston B, Williams MA (eds) *Social robotics. Lecture notes in computer science*. Springer, Cham, pp 125–134. https://doi.org/10.1007/978-3-319-11973-1_13
 25. Eyssel F, Hegel F (2012) (S)he's got the look: gender stereotyping of robots. *J Appl Soc Psychol* 42(9):2213–2230. <https://doi.org/10.1111/j.1559-1816.2012.00937.x>
 26. Fasola J, Matari MJ (2013) A socially assistive robot exercise coach for the elderly. <https://doi.org/10.5898/JHRI.2.2.Fasola>
 27. Forlizzi J, Saensuksopa T, Salaets N, Shomin M, Mericli T, Hoffman G (2016) Let's be honest: a controlled field study of ethical behavior in the presence of a robot. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 769–774. <https://doi.org/10.1109/ROMAN.2016.7745206>. ISSN: 1944-9437
 28. Freedman R, Borg JS, Sinnott-Armstrong W, Dickerson JP, Conitzer V (2020) Adapting a kidney exchange algorithm to align with human values. *Artif Intell*, p 103261
 29. Ghaffary S (2018) Is this robot really going to replace a security guard? <https://www.vox.com/2018/10/8/17913420/security-robot-cobalt-robotics-knightscope-slack-yelp>
 30. Ghazali AS, Ham J, Barakova EI, Markopoulos P (2017) Pardon the rude robot: social cues diminish reactance to high controlling language. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 411–417. <https://doi.org/10.1109/ROMAN.2017.8172335>. ISSN: 1944-9437
 31. Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: Proceedings on the 12th IEEE international workshop on robot and human interactive communication, 2003, ROMAN 2003. IEEE, pp 55–60
 32. Graham G et al (2004) *Eight theories of ethics*. Psychology Press, London
 33. Guarini M (2011) Computational neural modeling and the philosophy of ethics reflections on the particularism-generalism debate. In: Anderson M, Anderson SL (eds) *Machine ethics*. Cambridge University Press, Cambridge Google-Books-ID: N4IF2p4w7uwC
 34. Ham J, Spahn A (2015) Shall i show you some other shirts too? The psychology and ethics of persuasive robots. In: Trapp R (ed) *A construction manual for robots' ethical systems: requirements, methods, implementations, cognitive technologies*. Springer, Cham, pp 63–81. https://doi.org/10.1007/978-3-319-21548-8_4
 35. Haring KS, Mosley A, Pruznick S, Fleming J, Satterfield K, de Visser EJ, Tossell CC, Funke G (2019) Robot authority in human-machine teams: effects of human-like appearance on compliance. In: Chen JY, Fragomeni G (eds) *Virtual, augmented and mixed reality. applications and case studies. Lecture Notes in Computer Science*. Springer, Cham, pp 63–78. https://doi.org/10.1007/978-3-030-21565-1_5
 36. Harman G (2000) The nonexistence of character traits. In: *Proceedings of the Aristotelian society*, vol 100. JSTOR, pp 223–226
 37. Hoffman G, Forlizzi J, Ayala S, Steinfeld A, Antanitis J, Hochman G, Hochendoner E, Finkenaer J (2015) Robot presence and human honesty: experimental evidence. In: 2015 10th ACM/IEEE international conference on human–robot interaction (HRI), pp 181–188. ISSN: 2167-2121
 38. Howard A, Borenstein J (2018) The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci Eng Ethics* 24(5):1521–1536
 39. Hughes CE, Benoit TS (2017) Culturally adaptive avatar simulator. <https://patents.google.com/patent/US9690784B1/en>
 40. Jeong S, Logan DE, Goodwin MS, Graca S, O'Connell B, Goodenough H, Anderson L, Stenquist N, Fitzpatrick K, Zisook M, Plummer L, Breazeal C, Weinstock P (2015) A social robot to mitigate stress, anxiety, and pain in hospital pediatric care. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction extended abstracts, HRI'15 extended abstracts. Association for Computing Machinery, Portland, pp 103–104. <https://doi.org/10.1145/2701973.2702028>

41. Ku H, Choi JJ, Lee S, Jang S, Do W (2018) Designing shelly, a robot capable of assessing and restraining children's robot abusing behaviors. In: Companion of the 2018 ACM/IEEE international conference on human-robot interaction, HRI '18. Association for Computing Machinery, Chicago, pp 161–162. <https://doi.org/10.1145/3173386.3176973>
42. Laitinen A (2016) Robots and human sociality: normative expectations, the need for recognition, and the social bases of self-esteem. In: Sociable robots and the future of social relations. IOS Press, Amsterdam, pp 313–322. <https://doi.org/10.3233/978-1-61499-708-5-313>
43. Lin YC, Liu TC, Chang M, Yeh SP (2009) Exploring childrens perceptions of the robots. In: Chang M, Kuo R, Kinshuk, Chen GD, Hirose M (eds) Learning by playing. Game-based education system design and development, Lecture Notes in Computer Science. Springer, Berlin, pp 512–517. https://doi.org/10.1007/978-3-642-03364-3_63
44. Lumbreras S (2017) The limits of machine ethics. Religions 8(5):100. <https://doi.org/10.3390/rel8050100>
45. Malle BF, Scheutz M (2015) When will people regard robots as morally competent social partners? In: 2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 486–491. <https://doi.org/10.1109/ROMAN.2015.7333667>
46. Merritt M (2000) Virtue ethics and situationist personality psychology. Ethical Theory Moral Pract 3(4):365–383. <https://doi.org/10.1023/A:1009926720584>
47. Meuhlhauser L, Helm L (2012) Intelligence explosion and machine ethics. In: Singularity hypotheses: a scientific and philosophical assessment, pp 101–126
48. Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21
49. Mubin O, Cappuccio M, Alnajjar F, Ahmad MI, Shahid S (2020) Can a robot invigilator prevent cheating? AI Soc. <https://doi.org/10.1007/s00146-020-00954-8>
50. Nomura T, Kanda T, Kidokoro H, Suehiro Y, Yamada S (2016) Why do children abuse robots? Interact Stud 17(3):347–369. <https://doi.org/10.1075/is.17.3.02nom>
51. Nussbaum MC (2009) Hiding from humanity: disgust, shame, and the law. Princeton University Press, Princeton
52. Obaid M, Aylett R, Barendregt W, Basedow C, Corrigan LJ, Hall L, Jones A, Kappas A, Küster D, Paiva A et al (2018) Endowing a robotic tutor with empathic qualities: design and pilot evaluation. Int J Humanoid Rob 15(06):1850025
53. Ogunyale T, Bryant D, Howard A (2018) Does removing stereotype priming remove bias? A pilot human-robot interaction study. arXiv preprint [arXiv:1807.00948](https://arxiv.org/abs/1807.00948)
54. Reich-Stiebert N, Eyssel F (2015) Learning with educational companion robots? Toward attitudes on education robots, predictors of attitudes, and application potentials for education robots. Int J Soc Robotics 7(5):875–888. <https://doi.org/10.1007/s12369-015-0308-9>
55. Reichenbach J, Bartneck C, Carpenter J (2006) Well done, robot! The importance of praise and presence in human-robot collaboration. In: ROMAN 2006—The 15th IEEE international symposium on robot and human interactive communication, pp 86–90. <https://doi.org/10.1109/ROMAN.2006.314399>. ISSN: 1944-9437
56. Roizman M, Hoffman G, Ayal S, Hochman G, Tagar MR, Maaravi Y (2016) Studying the opposing effects of robot presence on human corruption. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), pp 501–502. <https://doi.org/10.1109/HRI.2016.7451826>. ISSN: 2167-2148
57. Sandoval EB (2019) Addiction to social robots: a research proposal. In: 2019 14th ACM/IEEE international conference on human-robot interaction (HRI), pp 526–527. <https://doi.org/10.1109/HRI.2019.8673143>
58. Sandoval EB, Brandstetter J, Bartneck C (2016) Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI)
59. Sandoval EB, Brandstetter J, Obaid M, Bartneck C (2016) Reciprocity in human-robot interaction: a quantitative approach through the Prisoner's dilemma and the ultimatum game. Int J Soc Robotics 8(2):303–317. <https://doi.org/10.1007/s12369-015-0323-x>
60. Sartre JP (2001) Being and nothingness: an essay in phenomenological ontology. Citadel Press, New York
61. Serholt S, Barendregt W (2014) Students' attitudes towards the possible future of social robots in education. In: Workshop proceedings of RO-MAN
62. Sparrow R (2017) Robots, rape, and representation. Int J Soc Robotics 9(4):465–477
63. Sparrow R (2017) Robots, rape, and representation. Int J Soc Robotics 9(4):465–477. <https://doi.org/10.1007/s12369-017-0413-z>
64. Sparrow R (2020) Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? Int J Soc Robotics 1–7
65. Strait M, Ramos AS, Contreras V, Garcia N (2018) Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In: 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 452–457
66. Vallor S (2015) Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philos Technol 28(1):107–124. <https://doi.org/10.1007/s13347-014-0156-9>
67. Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, Oxford Google-Books-ID: RaCkDAAQBAJ
68. Vlachos E, Schärfe H (2014) Social robots as persuasive agents. In: international conference on social computing and social media. Springer, Cham, pp 277–284
69. Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford
70. Wallach W, Franklin S, Allen C (2010) A conceptual and computational model of moral decision making in human and artificial agents. Top Cogn Sci 2(3):454–485
71. Weizenbaum J (1966) Eliza computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
72. Wesche JS, Sonderegger A (2019) When computers take the lead: the automation of leadership. Comput Hum Behav 101:197–209. <https://doi.org/10.1016/j.chb.2019.07.027>
73. Whitby B (2012) Do you want a robot lover? The ethics of caring technologies. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MITP, pp 233–248. <https://ieeexplore.ieee.org/document/6733984>
74. Wiegel V (2010) Wendell Wallach and Colin Allen: moral machines: teaching robots right from wrong. Ethics Inf Technol 12(4):359–361. <https://doi.org/10.1007/s10676-010-9239-1>
75. You S, Nie J, Suh K, Sundar SS (2011) When the robot criticizes you...: self-serving bias in human-robot interaction. In: Proceedings of the 6th international conference on Human-robot interaction, HRI '11. Association for Computing Machinery, Lausanne, pp 295–296. <https://doi.org/10.1145/1957656.1957778>
76. Zaal E, Mills G, Hagen A, Huisman C, Hoeks J (2017) Convincing conversations: using a computer-based dialogue system to promote a plant-based diet. In: CogSci
77. Zahavi D (2010) Shame and the exposed self. In: Webber J (ed) Reading Sartre: on phenomenology and existentialism. Routledge, Abingdon

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Massimiliano L. Cappuccio is Senior Researcher at University of New South Wales, where he serves as Deputy Director of the Values in Defence and Security Technology research group. As a cognitive philosopher and a technology ethicist, his work focuses on human-robot interaction, human performance, AI theory, and the link between skill acquisition and virtue cultivation.

Eduardo B. Sandoval is a member of the Creative Robotics Lab, University of New South Wales (UNSW), Sydney Australia, where he works as a Scientia Fellow. Eduardo does research in Human-Robot Interaction, and Social Robotics. His current projects focus on meaningful machines, reciprocity in human-robot interaction, addictive technologies, and game-like activities with social robots.

Omar Mubin is Senior Lecturer at MARCS Institute, Western Sydney University, where he conducts research on human computer interaction, adaptive mobile social agents, the perception of social robots, and the role of robots in education.

Mohammad Obaid is Associate Professor of human-computer interaction at the Department of Computer Science & Engineering (Interaction Design Division), Chalmers University of Technology. His works focus on empathic robot teachers, social drones, and gesture in human-robot communication.

Mari Velonaki is Director of the Creative Robotics Lab and is lead chief investigator in an ARC LIEF Grant which has led to the development of the multi-institutional National Facility for Human Robot Interaction. She has worked as an artist and researcher in the field of interactive installation art.

Core Publication #8 [CP8]

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2021),
“Robotics Aids for Character Training: More Than Just Another Enabling
Condition” *International Journal of Social Robotics*, 13(1), Springer Verlag.

Reproduced with permission from Springer Verlag. All rights reserved.



Robotics Aids for Character Building: More than Just Another Enabling Condition

Massimiliano L. Cappuccio¹ · Eduardo B. Sandoval² · Omar Mubin³ · Mohammad Obaid⁴ · Mari Velonaki²

Accepted: 26 January 2021 / Published online: 15 February 2021

© The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021

Can robotics play a moral role in the life of humans? This question has been systematically addressed by the main approaches to machine ethics, which are prevalently inspired by normative theories in the consequentialist and deontological traditions. These theories focus on the maximization of practical and psychological advantages for the users and the obligation to protect human dignity, respectively. In many circumstances, for example when designing autonomous vehicles, the careful implementation of these theories is indispensable to assure that the systems' behaviour aligns with the ethical intuitions that are shared by most people and enshrined in our systems of laws, such as the fundamental principles pertaining to common good or individual rights [7].

However, in addition to providing utilitarian benefits to the users and diligently imposing severe normative prescriptions, social robots can play a moral function also in another important sense: through repeated, interactive engagement in familiar scenarios, appropriately designed robots can help

the human users recognize and correct their bad habits, dedicate themselves to commendable activities, and incorporate values as spontaneous inclinations to think and behave virtuously. Like other forms of virtue cultivation, this process may lead, eventually, to fortify one's moral character and adopt a better way of living. This kind of moral betterment is central to virtue ethics, the normative theory that construes morality around the idea of living a praiseworthy lifestyle.

The assumption behind this idea is that true fulfilment and enduring happiness can only be accomplished through self-improvement and constant engagement in admirable activities. Specifically, developing one's good nature is possible by recognizing and correcting one's own flaws, following the inspiring moral example of others, and learning to embody their positive habits and qualities. It seems that today's robots are still unable to do any of these things for us, but they can already help us doing all of them better, if we decide that that is our goal. This is the key idea of what we call "virtuous robotics", an approach to machine ethics and social robots design inspired by the theory of virtues.

Virtue ethics has enjoyed a significant renaissance during the last few decades, thanks both to the philosophers who rediscovered its importance within the Western (e.g., Aristotelian) tradition and those who explored the links with its East-Asian (e.g., Confucian) homologous. During the last ten years, especially thanks to the work of philosophers like Mark Coeckelbergh, Rob Sparrow, and Shannon Vallor, virtue ethics became a prominent option in the field of technology ethics, as epitomised by Vallor's influential monograph *Technology and the Virtues* [2, 5, 6]. The foundational work conducted by these pioneers inspire in more than one way this special issue.

The approach to technology ethics inspired by virtue theory operates under the assumption that our technological practices are not value-neutral, but importantly contribute to shape our understanding of what is right and wrong. Moral values are always embedded, more or less implicitly, in the way technology is designed and adopted. By channelling

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

Eduardo B. Sandoval
e.sandoval@unsw.edu.au

Omar Mubin
O.Mubin@westernsydney.edu.au

Mohammad Obaid
mobaid@chalmers.se

Mari Velonaki
mari.velonaki@unsw.edu.au

¹ School of Engineering and IT, University of New South Wales Canberra, Canberra, Australia

² School of Art and Design, University of New South Wales, Sydney, Australia

³ MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

⁴ Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

such values into everyday life, autonomous technologies can be leveraged to exert a positive or negative influence on the user's character. Our approach, that we feel close in spirit to the proposal advanced by Paul Dumouchel and Luisa Damiano in their book, *Living with Robots. Emerging Issues on the Psychological and Social Implications of Robotics* [3], particularly focuses on the moral role played by social robots and their transformative potential. Embodied humanoid agents have an unprecedented power to shape our conduct, our relationships, and our perception of what is ethically acceptable and desirable in the interpersonal sphere.

One important transformative aspect of our relationship with autonomous technology is that we often feel inclined to perceive social robots as moral patients, which raises the question—systematically addressed by David J. Gunkel in his recent book *Robot Rights* [4]—concerning whether or not robots deserve some form of moral consideration. Because they seem to claim for themselves the roles of both moral agents and moral patients, social robots—more than any other assistive technology—promise to transform not only the evaluative categories that we adopt in the legal and public policy domains, but also, more deeply and less obviously, the spirit of our customs and social norms.

Virtuous robotics reinterprets the groundwork established by our predecessors to put forward an entirely original approach to human–robot interaction. Our approach intends to offer a systematic response to the need, perceived more and more urgently by both makers and scholars, of a new human-centric normative framework to regulate autonomous assistive technologies without hindering the expression of their innovative potential.

Unfortunately, today's discussions in robot ethics are still narrowly focused on the moral risks that human–robot interaction could possibly bring to our society. To be clear: discussing the ethical risks associated with social robotics is indispensable because, without appropriate regulation and scrutiny, assistive technologies (either because inappropriately used or because loaded with misleading expectations) may eventually produce a tangible deterioration of the socio-ethical and political fabric of interpersonal relations. However, it would be impossible to assess the real transformative potential of social robotics without addressing the opposite, positive side of the story as well: by promoting a vigilant optimism, virtuous robotics can help us express a legitimate hope that technology will be capable to bring about authentic human development and opportunities of moral betterment.

Unlike other approaches to robot ethics, we are not simply advocating compliance with the ethical standards dictated by an abstract normative theory. That is why we will not limit ourselves to hope that assistive technologies were used to create the necessary material and intellectual preconditions for moral flourishing (for example, freeing people

from urgent needs and inequalities). We certainly share this hope, but we also believe that, compared to it, our project is more ambitious and more concrete at once: we envision that autonomous technologies, if considerably designed and sensibly integrated in our cultural forms of life, will soon be called to play not only the role of moral *enablers*, but also that of moral *facilitators* and *enhancers*. This means that social robots can and should be used to actively, concretely contribute to the development of their users' moral postures in practically significant scenarios, and not just to the fulfilment of some abstract rules or the formation of some generic ethical beliefs. If this hypothesis is correct, then a new generation of robots should be primarily conceived and specifically designed to promote moral flourishing in humans, aiming to make tomorrow's users wiser and gentler people, with more decent and realized lifestyles. This development is in principle possible if virtuous technologies are conceived to play an expansive and supportive role in the moral domain, analogously to how some assistive technologies are already successfully used to augment our resources and capabilities in the cognitive domain.

To give voice to this hope, but also to carefully explore the most efficacious strategies to pursue it, this special issue intends to address whether and how the responsible design and application of social robots could actively promote true moral awareness in humans and support their ethical decisional processes, considering both the case studies that favour such an optimistic expectation and the critical objections that encourage prudence.

We would like to acknowledge how our journey has crossed the path of several other studies and researchers in social-robotics, human–robot interaction, and technology ethics, importantly benefitting from these encounters. Some of our ideas about a virtuous approach to robotics were initially explored during a panel discussion held on 9 August 2017 at the Philosophy Department of University of Wollongong as part of the first UOW Social Robotics Workshop. This discussion, which primarily revolved on a paper published in the same year by Sparrow [5] in *International Journal of Social Robotics*, inspired Massimiliano Cappuccio, Anco Peeters and William McDonald [1] to respond with another paper on moral consideration for robots based on virtue and recognition, which was published in *Philosophy & Technology* in early 2019. Some of these ideas were further developed during a thematic session on “Robots and Virtue” (co-organized by Fady Alnajjar, Massimiliano Cappuccio, Friederike Eyssel and Mohamad Eid) held on 4–7 February 2018 at United Arab Emirates University and New York University Abu Dhabi. Then again, a seminal investigation on flourishing informed the second day of activities of the interdisciplinary workshop “Robots & AI in society”, co-organized by Omar Mubin and Massimiliano Cappuccio and held on 8–9 November 2018 at

Western Sydney University in collaboration with the Digital Humanities Research Group.

Summarizing the outcome of this journey, this special issue seeks to explore from a multidisciplinary perspective the idea that, if effectively programmed for pedagogical, training, and self-development purposes, robots could be used to cultivate virtues like compassion, honesty, and generosity, learn to manage negative emotions like anger and envy, and practice positive attitudes towards others, such as respect, care, and friendliness. The nine original papers selected for publication in this special issue investigate the philosophical, psychological, and technological/implementation dimensions of this hypothesis.

The paper that opens the collection is co-authored by the five guest editors of this special issue. Its aim is to provide a solid conceptual foundation to the “virtuous robotics” project, rooting this proposal into the principles of virtue theory and distinguishing it from other approaches to machine ethics. The paper explores functions, tasks, roles, and design principles that best suit the virtuous application of social robots and discusses a number of theoretical and practical objections.

The second contribution is authored by Australian technology ethicist Rob Sparrow. The well-argued doubts that he raises in this paper somehow temper the optimistic perspective presented in the previous one. Sparrow argues that a structural asymmetry informs our relationship with robots: while social robots can promote the cultivation of vice, they cannot help humans develop their virtues. The paper discusses the reasons and ethical implications of this asymmetry.

The third paper is authored by a technology ethicist working in the European philosophical tradition, Mark Coeckelbergh. Giving voice to a post-phenomenological perspective, his analyses highlight the need to further emphasize the relational dimension of moral character in human–robot interactions and the dependency of virtue on systems of cultural practices, critically discussing the absolute primacy incorrectly attributed by the Western ethical traditions to individual agency and rational deliberation.

The fourth paper, co-authored by Merel Keijsers, Hussain Kazmi, Friederike Eyssel, Christoph Bartneck is an empirical study in human–robot interaction informed by the rigorous methodology of social psychology. The study focuses on the psychological mechanisms underlying vandalism and abusive acts conducted against robots in public spaces. This study investigates how aggressive and punitive attitudes toward robots correlate to technology acceptance, desires to assert dominance, and the tendency to attribute intelligence and mentalistic features to robots.

The fifth contribution to the special issue¹ examines a well-known dilemma in the virtue ethical tradition—i.e. the philosophical question concerning why mistreating non-human animals is ethically unacceptable—and transposes it into the contemporary debate concerning the moral consideration for social robots. The four authors of the paper, Simon Coghlan, Frank Vetere, Jenny Waycott, and Barbara Barbosa Neves argue that robots may affect the moral development of children and, in general, the human responses to nonhuman animals, and consider the implications for design and policy.

In the sixth paper included in the collection, its authors, Anco Peeters and Pim Haselager discuss the psychological background and the moral implications of designing virtuous sex robots, that is social robots specifically conceived to engage in intimate relationships with humans. Is the existence of such robots, and the desire to utilise them to satisfy erotic needs, inherently vicious, or can these robots be developed with a moralising and educational intent, considering also the alleged deceptiveness of humanoid robots and the fact that they can be used instrumentally, without receiving consent from the robot itself?

Social virtues require the ability to empathise with others and regulate one’s own conduct to establish joint attention with them. But what happens when these abilities are disrupted and what can be done to recover them? The seventh paper, co-authored by Fady Alnajjar, Massimiliano Cappuccio, Abdulrahman Renawi, Omar Mubin, and Chu Kiong Loo, addresses this question, presenting the results of an experimental study on the beneficial impact played by robots to facilitate interactive engagement and maintain social attention in autistic children.

The eighth paper investigates a key anthropomorphic feature of social robots, i.e. their capability to communicate via facial expressions. The authors, Chris Cheshier and Fiona Andreallo, explain how robot faciality plays not only an aesthetic function but also an ethical one, soliciting moral awareness and self-regulation in the human users. Unlike face-less technologies, social robots can rely on their face and gaze to make humans feel recognized as objects of moral evaluation and hence as moral agents who carry the burden of ethical responsibility.

The ninth paper offers decisive empirical evidence that robots can be efficaciously deployed in public contexts to encourage ethical behaviours in casual interactants, educating them to respect deontological norms and fulfill their civic duties. The study conducted by Berardina De Carolis,

¹ Note that, although this paper was originally written to be included in our thematic collection on Virtuous Robotics, it has already been published by *International Journal of Social Robotics*. The paper is not reproduced in this special issue, but readers can find it in the December 2019 issue.

Francesca D'Errico, Nicola Macchiarulo, and Veronica Rosano successfully tested robot-mediated playing as a method to promote ecological awareness among children, who acquired environment-friendly recycling habits as a result.

These papers helped us understand whether and how social robots (especially those explicitly programmed to act as coaches, companions, or training assistants) can inculcate new virtues or assist humans in developing good habits, accompanying their personal search for a morally aware and more balanced way of living. We hope that new studies will be designed and conducted to explore other applications of virtuous robotics in an even greater variety of scenarios and tasks, further clarifying—through the use of socially interactive robots—the cognitive and social mechanisms that underpin the formation of character and the cultivation of virtue in adults and children.

References

1. Cappuccio ML, Peeters A, McDonald W (2019) Sympathy for dolores: moral consideration for robots based on virtue and recognition. *Philos Technol* 2019:1–23
2. Coeckelbergh M (2018) Why care about robots? Empathy, moral standing, and the language of suffering *Kairos*. *J Philos Sci Sci* 20(1):141–158
3. Dumouchel P, Damiano L (2017) *Living with robots*. Harvard University Press, Cambridge
4. Gunkel DJ (2018) *Robot rights*. The MIT Press, Cambridge
5. Sparrow R (2017) Robots, rape, and representation. *Int J Soc Robot* 9(4):465–477
6. Vallor S (2016) *Technology and the virtues: a philosophical guide to a future worth wanting*. Oxford University Press, Oxford
7. Wallach W, Allen C (2009) *Moral machines. Teaching robots right from wrong*. Oxford University Press, Oxford

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Massimiliano L. Cappuccio is Senior Researcher at the University of New South Wales, where he serves as Deputy Director of the Values in Defense and Security Technology Group. He also holds a secondary affiliation with UAE University, the national university of the United Arab Emirates, where he had been Associate professor of Cognitive science and director of the Cog Sci Lab (August 2011–December 2018). As a cognitive philosopher and a technology ethicist, his research is interdisciplinary and aims to integrate phenomenological analyses, empirical experimentation, and synthetic modelling. His work focuses on theoretical and applied issues in human performance (skill acquisition and skill disruption), human–robot interaction (social and assistive robotics), philosophy and theory of artificial intelligence (frame problem, computational theory of mind). His philosophical perspective attempts to integrate embodied, enactive and extended approaches to social cognition, virtue ethics, and complexity theory. He conducts an intense activity as organizer of academic events, including interdisciplinary workshops, research seminar series, and international conferences (like the TEMPER workshop on Training, Enhancement, and

Military Performance at ADFA, and the annual Joint UAE Symposium on Social Robotics in Abu Dhabi).

Eduardo B. Sandoval is Scientia Associate Lecturer at the University of New South Wales (UNSW) since 2018. Dr. Sandoval has an interest in the design of machines that are capable of meaningful social interactions with humans. Dr. Sandoval's work incorporates insights from behavioural economics and social psychology to explore different approaches in social robotics. His pioneering work has evolved around the topics of Reciprocity in Human Robot Interaction (HRI), Robot Design, and Robot Games. Similarly, he has explored the use of robots in education and healthcare and future addiction to robots in the academia and industry. Dr. Sandoval has produced thirty-four peer-review publications in relevant venues related to HRI, a robot design register in 2019, and his work in Human–Robot Interaction has circulated in international media. Similarly, he has been appointed as co-chair, guest editor, and reviewer for several conferences and journals related to HRI, and Social Robotics and has collaborated closely with industry in the design and development of robot prototypes. In 2014, The National Autonomous University of Mexico awarded Dr. Sandoval with the Medal Alfonso Caso as the best student of the Postgraduate Program of Industrial Design at the National Autonomous University of Mexico. In 2016, Dr. Sandoval obtained his Ph.D. in Human Interface Technology at the HITLab NZ, University of Canterbury, New Zealand. His future work involves multidisciplinary research in social robotics as robot games and no-visual natural interfaces aiming future human's well-being.

Omar Mubin is a Senior Lecturer in Human Computer Interaction at the School of Computer, Data and Mathematical Sciences at Western Sydney University, Australia. Dr. Mubin's primary research interests are human robot interaction and human-agent interaction. Specifically, he studies social robotics and their applications and consequently interaction with humans in education, public spaces and information dissemination scenarios. He also has a keen interest in the application of data analytics to the areas of Scientometrics, ICT and Public Health. His current h-index according to Google Scholar is 19 with a more than 1700 citations. Dr. Mubin is involved in teaching and supervising (undergrad and postgrad) students in the broader area of Human Computer Interaction, Mobile Computing and Health Informatics.

Mohammad Obaid is an Associate Professor and Head of the Interaction Design Unit at the Department of Computer Science and Engineering, Chalmers University of Technology (Sweden). Dr. Obaid received his B.Sc., M.Sc. (First Class Honors) and Ph.D. degrees in Computer Science from the University of Canterbury, Christchurch, New Zealand, in 2004, 2007 and 2011, respectively. In April 2018, he received his Docent degree (Associate Prof. rank) from Uppsala University (Sweden). He worked at several international research centres including the Human Centered Multimedia Lab (HCM Lab), University of Augsburg (Germany), the Human Interface Technology Lab New Zealand (HITLab NZ), University of Canterbury (New Zealand), the t2i Lab, Chalmers University of Technology (Sweden), and the Social Robotics Lab, Department of Information Technology, Uppsala University (Sweden). Dr. Obaid is one of the founders of the Applied Robotics Group at the Interaction Design Division at Chalmers University of Technology (established in 2015). He is the (co-)author of over 70 publications within the areas of his research interests on Human–Robot Interaction and Human–Computer Interaction.

Mari Velonaki's research is situated in the multi-disciplinary field of Social Robotics. She holds a Ph.D. in Experimental Interface Design (UNSW 2003). Velonaki began working as a media artist/researcher in the field of responsive environments and interactive interface design in 1997. She pioneered experimental interfaces that incorporate movement, speech, touch, breath, electrostatic charge, artificial vision and

robotics, allowing for the development of haptic and immersive relationships between participants and interactive agents. In 2014 she was voted by Robohub—a large robotics community of researchers, educators and business- as one of the world’s 25 women in robotics you need to know about. She is the founder and director of the Creative

Robotics Lab (Art & Design UNSW) and the founder and director of the National Facility for Human Robot Interaction Research (UNSW, USYD, UTS, St Vincent’s Hospital). Mari is a Research Leader at the UNSW Ageing Futures Institute.

Core Publication #9 [CP9]

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (eds.) (2021),
“Virtuous Robotics: Artificial Agents and the Good Life”
special issue of *International Journal of Social Robotics*, 13(1), Springer Verlag.

Redacted.

Chapter 3, Part 3 - Growing with Robots: Autonomous Virtuous Agents in Education

This section is centred on Core Publications #10 and #11 [CP10] [CP11]

3.3.1 Virtuous applications of anthropomorphism in the classroom environment

The third Part of this Chapter explores in greater detail how the Virtuous Robotics framework can be applied in the educational context, examining some chapters of the collective volume *Robots in Education* (Alnajjar et al 2021a [CP10]), a review of technologies to educate obese children (Alotaibi et al 2023 [SP15], and an empirical HRI study designed to promote disciplined behaviours in children of scholar age and college students (Mubin et al 2020 [CP11]). As the integrity of moral character primarily consolidates during the formative age, the pedagogical function of robots in the educational context plays a key role to achieve the goals of Virtuous Robotics, which ultimately aims to develop positive character traits through healthy and edifying relationships with artificial social agents. Also, the applications of HRI in the classroom environment represent and illustrate my previous considerations concerning the social impact of embodiment, as they will clearly demonstrate the centrality of anthropomorphism as a key embodied phenomenon consisting in empathizing with robots and engaging in rich, meaningful relationships with them.

My studies in the educational domain show that the bodily details of the interactants shape their possibilities of meaningful interaction, for example the possibility to understand each other's goals and the possibility to perform joint actions, behaving collaboratively to achieve a communal goal. My empirical study reviewed in this section, "Can a robot invigilator prevent cheating" (Mubin et al 2020 [CP11]), explores an important aspect of embodied social interaction among human and robots: exactly because they are embodied in a way that matches the body of their counterpart, both natural and artificial cognitive agents are able to display their ethical attitudes, communicating ethical expectations and moral inclinations. In humans, these attitudes, expectations and inclinations derive from their moral dispositions, that is the concrete familiarity with certain value-laden behaviours. This familiarity is embedded in systems of habit to perceive certain things as good or bad and interact with them accordingly.

Habitual moral dispositions are dynamical qualities or features of their characters: they incline people to act and think courageously, prudently, and generously (virtues), or rather

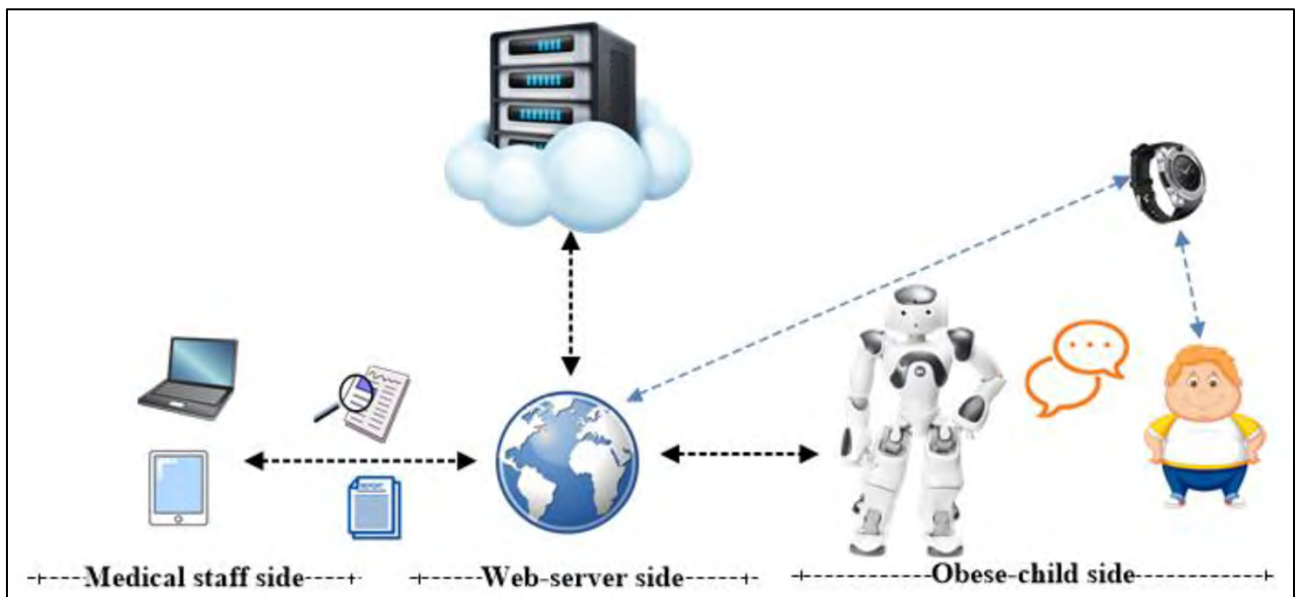
cowardly, imprudently, and greedily (vices). My work does not address the question concerning whether robots currently have or could possibly have a moral character (addressed by Wallach Allen 2009 or Formosa & Ryan 2021). Rather, it focuses on the embodied features of social robots to demonstrate that, through these features, robots can have an impact on the moral character of humans during social interaction with them (as proposed by Coeckelbergh 2010). More specifically, a robot's embodiment may solicit or rather suppress certain moral dispositions of the human interactant or even contribute to the gradual creation of new moral dispositions.

3.3.2 Smart technologies, robot coaches, and the treatment of obesity

In the “Virtuous robotics” paper [CP7], I briefly reviewed the interesting case study of Shelly, a robot turtle that was frequently subject to abuse by children to point to the fact that, by displaying physical pain and psychological suffering, and by refusing to play with the children, the turtle was able to encourage a more gentle and considerate attitude from its young interactant (Ku et al 2018a, 2018b). Similarly, robots can be designed to solicit compassion in humans.



In a somewhat similar manner, our paper on robots for diabetic children (Alotaibi et al 2023 [SP15]) has reviewed several empirical studies to show that appropriately designed humanoid robots can operate as coaches to teach proper alimentary habits, which is particularly important for the treatment of obesity in scholar-age subjects. Our review has compared various technological interventions available for childhood obesity prevention and treatment.



The use of machine learning was observed in various studies, suggesting that the integration of AI in social robots, in combination with other interactive technologies, can be an effective strategy to treat childhood obesity. That is why our review concludes that robots can be an effective solution to teach children the virtues of self-restraint in order to help them control their appetite and embrace healthier diets and lifestyles. In particular, opportunely designed social robots have significant potential to assist healthcare workers and deliver customized care to children because their embodied features make them more capable than other types of interactive technologies to positively influence the behaviour and motivation of the human interactants.

3.3.3 Robots as teachers: opportunities and risks

Despite the many challenges, my colleagues and I believe that humanoid robots can be effectively deployed in the classroom environment to prevent cheating and enforce a severe discipline. Our empirical study with the Nao robot suggests that this is possible, if the robot's embodiment (including both their gazing behaviours and their physical presence) is imposing enough to intimidate the undisciplined students and discourage them from cheating. In an ideal scenario, the embodied features of the human interactants (the emotional responses that solicit or constrain their moral dispositions) should be paired with the embodied features of the humanoid robots to resonate with them and thus produce complex patterns of interactive behaviours that, from the point of view of a bystander, display an obvious moral quality. The studies mentioned in this section can be considered applications of the taxonomy illustrated in the previous section because they include robots specifically designed to achieve *goals* like "Recognize, inhibit, and subdue negative

character traits” and “Develop or reinforce positive character traits” (such as courage, compassion, charity, honesty, and positive ways of thinking, like hope) and take on certainly *roles* such as “informer” or “scarecrow” (robots deterring vicious behaviours), as discussed in sections 3.2.4.1 and 3.2.4.2. Thus, they can be considered successful demonstrations of how the Virtuous Robotics framework inspire design and conceptualization of original forms of HRI.

These issues were discussed by me and a group of colleagues working in social robotics during an intensive retreat that I organized in 2020 in Al Ain, with the sponsorship of UAE University. The retreat led to the ideation, the realization, and eventually, the publication, of our collectively co-authored monograph “Robots in Education” (Alnajjar et al 2021a [CP10]), which in 2021 was published as a book. While all the co-authors are responsible for the content of the whole book, they also agree that I had a much greater involvement and a primary responsibility for a couple of book chapters, namely chapter 3, on the psychology of robots in classroom, and, even more, chapter 10, on ethics, which is relevant to the virtue ethical approach to robots education explored in this section. Here, I describe several applications of humanoid robots in the classroom environment, discusses their advantages from a perspective of virtue cultivation and character flourishing, raises a number of concerns related to the opportunity and the legitimacy of using robots as coaches and teachers, and eventually proposes a way forward to prudently explore the use of robots in the educational context.

3.3.3.1 Psychological damage – Using a robot to assist children with special needs may turn out harmful, due to the robot unreadiness to provide adequate care. Research in this area comes with a variety of challenges for research ethics (Sharkey and Sharkey 2010).

3.3.3.2 Privacy – Robots might infringe on the privacy rights granted by Articles 16 and 40 of the UN Convention on Child Rights. Differentiating between private and public information requires a level of common sense and understanding of social norms that robots lack by default (despite promising efforts of AI research to imbue agents with common sense, see Santos 1998).

3.3.3.3 De-skilling – In a world in which machines increasingly take over tasks from humans, humans might lose essential skills (Vallor 2015). This also holds true for the educational context. A robot that takes over certain tasks could lead to teachers increasingly losing practise in those specific tasks: the teachers’ roles could be reduced to finding the on/off switch of the educational robot. Since we consider the robot to be an assistant and

not a replacement for the teacher, we Also the students might experience de-skilling: being able to verbally ask the robot any question and, at least for simple factual information, receive a direct response, inhibits critical research skills.

3.3.3.4 Authority – Granting robots authority might be necessary to discipline individuals, however authority and power would make many users uncomfortable (You et al 2011; Zlotowski et al., 2017). As a consequence, robots might be more suitable for adopting the role of an assistant, proxy, or support to the human who is accepted and acknowledged as the key authority figure. The robot's lack of authority might be caused by the difficulty to administer punishment to the children or give them blame (Bartneck, Reichenbach & Carpenter 2008), together with the impossibility of held robots accountable. (i) *Restraint* – A robot might be required to restrain a child in order to protect it from self-harm and/or harm to others. Restraining a human requires considerable ethical reasoning abilities and motor control. Robots might have neither: e.g., a robot might struggle distinguishing rough play from bullying. (ii) *Accountability* – With authority comes responsibility, and with responsibility accountability. Is accountability possible for robots, if they don't have authority? Caretakers have a legal responsibility for the children entrusted to them. If a robot is left alone with a group of students, then it becomes unclear who would be accountable: the manufacturer, the school (owner), the teacher, or the robot itself?

3.3.3.5 Lack of emotional intelligence – Robots are still incapable of actually experiencing emotions (Damiano, Dumouchel & Lehmann 2015a, 2015b). This has implications for communication and social bonding in general: students cannot be efficaciously coordinated or directed if the robot's mental states are opaque. The lack of Affective Intelligence is a central problem for Affective Computing research, which highlights a number of possible associated risks: deceptive relationships; inappropriate attachment; de-socialization; and unrecognized identities³³. (i) *Deception* – A robot expressing emotions without experiencing emotions could be considered a deception and could mislead the student into believing that the nature of their relationship is personal (Wagner & Arkin 2009). (ii) *Attachment* – My colleagues and I have described in detail how infants and children may grow attached to robots they interact with and even grow expectations about having a social relationship with them (Sandoval, Billinghamurst & Cappuccio 2024). Due to the empathic limitations of robots

³³ Does the robot's incapability to experience the emotional life of others implies moral incompetence? Yes, if moral competence requires the ability to predict and explain the emotions and feelings of sentient beings (including what they could potentially experience as beneficial, detrimental, and offensive, see Malle & 2015). Explaining or predicting another's lived experience cannot be done by inferential or intellectual processes alone, but requires direct experiencing or interpretive capabilities developed socially (Bankins & Formosa 2023; Gallagher 2012).

in general, such a bond would remain unidirectional (Sharkey and Sharkey 2010). (iii) *De-socialization* – Even if robots were able to perfectly coordinate students' intellectual activities and cognitive work in the classroom environment, they might be unable to coordinate social activities. The opposite risk is also possible: that the robot, when perceived by the students as a toy or entertaining attraction, could disinhbit their social pulses, becoming a distracting element (Złotowski et al 2017). (iv) *Identity* – A robot might be unable to appropriately recognize or project the sociocultural specificity of the students and their background, such as their ethnicity, gender, tradition, or religion. Apart from eliciting frustration and disappointment in the students, this might be interpreted as offensive or disrespectful (Bartneck et al 2018; Eyssel & Hegel 2012).

3.3.4 Challenges and solutions

In our *Robots in Education* book (Alnajjar et al 2021a [CP10]), my colleagues and I suggest a number of heuristic strategies to deal with the aforementioned issues, minimizing possible distress for students and reducing unwanted liabilities for the teachers and their institutions.

3.3.4.1 Trust issues – First, the amount of trust educators place in a robot should stem from a realistic evaluation of its qualities and capabilities. School administrators and teachers must likewise be mindful of the robot's limitations in tasks that seem routine to humans. For instance, while robots can detect human users' emotions and identify certain types of goal-oriented behaviour, they lack the capacity to readily infer mental states such as desires or motivations and to interpret complex social contexts. This shortcoming makes it difficult for them to reach morally and socially acceptable decisions, and they are unlikely to handle potentially controversial etiquette challenges with assurance.

3.3.4.2 Teacher's responsibility – Next, any decision of moral importance within the classroom should be the teacher's responsibility, as the teacher is the sole moral agent accountable for the educational outcomes of the teaching process. The use of robots and their particular functions should be part of this decision-making process. For instance, if a robot provides supplementary features like conflict management, these should be considered in any situation where they could potentially assist. Still, despite these features, the teacher should not be under the impression that the robot can independently handle a difficult classroom situation or bear the weight of moral responsibility and legal consequences.

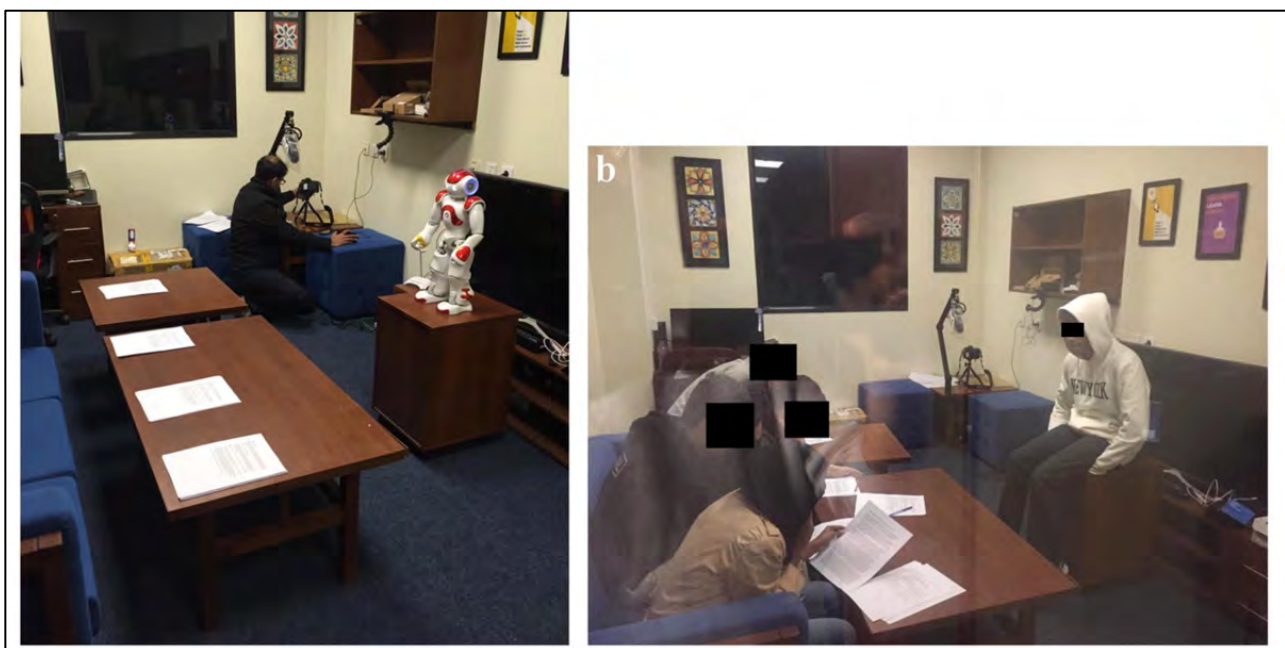
3.3.4.2 Recording and sharing information – Further, we propose that robots be permitted to use their cameras and microphones to monitor students at all times. Consequently, parents and administrators must be informed that these robots will continuously record the classroom and may retain the recordings for a period of time or use the data to enhance their neural network from experiences with the children. However, all parties involved might be concerned that this could raise privacy issues for both teachers and students. Data recording and sharing must comply with local legal regulations, such as the European Union's General Data Protection Regulation. Obtaining consent from parents and teachers for the use of robots in the classroom is crucial and should involve full disclosure regarding how the data will be recorded, processed, stored, and shared.

3.3.4.3 Licence to discipline and report – Last but not least, the robot's educational function in the classroom should coincide with its perceived level of authority. Features such as physical size, a human-like appearance, vocal tone, or facial expressions are among the factors that shape how much authority the robot is believed to possess. If a robot can project enough authority to fulfill its assigned tasks, we must consider the psychosocial, cultural, and legal ramifications of a machine wielding such authority. This point is especially significant, as I explored in a research paper co-authored with Omar Mubin and Fady Alnajjar, concerning the use of robots as invigilators [CP11].

3.3.5 Discipline and report: robots as inspectors

Educational robotics explores different assistive roles that robots can take in the classroom, for example as lecturer or discussant or students assistant. We focussed on employing robots in an assistive capacity as the invigilator of an exam, as this task demonstrates not only their suitability but also their ability to influence and control the morality and behaviour of the students via their presence. That is why we decided to run an empirical study where the Nao robot invigilated an exam for a group of students [CP11]. One of the most pressing questions concerning the integration of robots in the classroom concerns the robots' capability to maintain discipline, order and structure (Kennedy et al 2016; Serholt et al 2017). A broad range of applications is disclosed by the possibility to employ the autonomous and interactive capabilities of the robot to monitor and discipline the class when the teacher has to be momentarily absent, or when the classroom is too large for a single teacher.

3.3.5.1 Background and assumptions – An invigilator’s role is to discourage dishonest practices (cheating) and stop disrespectful behaviours from becoming ingrained into students’ character. Discipline not only supports regular classroom activities but is also integral to the learning process and a desirable outcome. In developing robot invigilators, we aim to give technology both educational and moral functions. Our robot invigilator helped us testing our framework for virtuous robotics. A relevant virtue is students’ propensity for attentive and self-controlled behaviour under a teacher’s supervision (Lewis et al. 2005), which depends on the respect students hold for their teachers, along with the moral authority society grants that role (Miller et al. 2017). In principle, the same reasoning could extend to a robot invigilator—but can also a robot be regarded as a moral authority?



3.3.5.2 Task design – To replicate a realistic exam setting and encourage potential cheating, each session included 4 students taking the test. They were given either version A or B of a short essay, followed by 8 questions (4 multiple choice and 4 short-answers). Each student had the correct answers for the other essay version, which enabled cross-marking and tempted them to cheat. They were informed that 2 versions existed and that, after finishing, they would swap papers to grade each other. All 4 students sat in a row. In 2 of the 3 conditions, the invigilator (a human or a Softbank Nao robot) was in front of them. In the no-invigilator control, only a shelf was present. The Nao robot remained mute but operated in its “autonomous mode,” tracking movement or sound and making small head and arm motions. Similarly, the human invigilator stayed silent yet maintained a natural, minimal presence without speaking.



3.3.5.3 Composition – A total of 56 (M = 31, F = 25) high school students participated in our experiment (their age ranging from 15–17 years old). This meant we conducted 14 testing sessions, each session comprising of 4 students. A total of 24 students (6 groups) participated in the Robot condition, 12 children (three groups) in the human condition and 20 children (five groups) in the control group condition with no invigilator. A number of quantitative measurements formed the base of our analysis. Video recordings from the sessions were manually coded with the help of the E-Lan software. Each video was analysed separately a total of four times, focusing on a different child each time.

Table 1 Table of means and standard deviations, first three rows of duration given in percentage of session time, talking frequency is number of dialogues or utterances per 10 min

From: Can a robot invigilator prevent cheating?

	Robot		Human		Control – no invigilator	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Looking-exam	0.24	0.5	0	0	0.05	0.17
Talking-exam	8.3	7.03	1.41	2.5	0.8	1.43
Talking-random	8.9	8.9	3.57	2.33	5.5	3.9
Talking-frequency	5.2	2.98	1.69	1.48	3.24	3.20

3.3.5.4 Results – The type of invigilation significantly impacted how long students discussed the exam protocol (χ^2 (N=48.2)=24.29, $p \leq 0.001$, $\varepsilon^2=0.52$), with about 8% of the session in the Robot condition spent on exam-related talk. Meanwhile, invigilation type did not noticeably affect the time spent talking about random topics (χ^2 (N=48.2)=2.30, $p=0.32$); in the Robot condition, students spent around 8% of the session on unrelated conversation, versus 3.5% in the Human condition. Lastly, invigilation type had a marked effect on the frequency of talking overall (χ^2 (N=48.2)=14.02, $p=0.001$, $\varepsilon^2=0.30$), and a Bonferroni post hoc test indicated children talked more often under the Robot condition compared to the Human condition ($p=0.001$). No significant differences emerged between the Human and

Control conditions. Overall, our findings show that the Nao robot's presence effectively curbed cheating but did not lessen indiscipline, which can eventually lead to explicit cheating (Muchai 2018). Overt cheating remained rare, with no significant disparities in how often students glanced at others' papers. Even without an invigilator, students did not feel compelled to cheat.

3.3.6 Challenges and solutions

Our attempt to deploy a robot invigilator in a classroom proved promising overall, but also more challenging than expected.

3.3.6.1 Task definition – During the experimental design phase, while reviewing similar studies, we realized there is no widely accepted, automated system to determine if and when students are “cheating.” The notion of cheating is notoriously difficult to define formally, making it challenging to gauge exam conduct with standardized protocols. Mistakes or recurring errors alone do not confirm cheating, nor does simple exam-time communication. Because of this, we had to distinguish casual, exam-irrelevant chatter from authentic cheating—treating them as different forms of “misbehaviour.” To facilitate our evaluation, we designed the exam so that half the class had the correct answers for the other half, creating an easy opportunity to swap solutions if they chose. We deliberately introduced this setup to see if it would prompt cheating in real exam-like conditions, yet the robot could not determine whether any such sharing occurred. Even with sophisticated programming, it would be unable to process relevant verbal information in real time. Consequently, we relied on manual detection of cheating, analysing the recorded footage and examining students' verbal exchanges on a case-by-case basis. While this approach was time-intensive, it remained feasible. However, the need for manual review highlights a structural limitation of robotic systems in such surveillance. Current language-processing technology still prevents robotic invigilators from fully replacing human oversight; at best, they can supplement a teacher or human proctor. Even with the most advanced tools, a robot cannot independently confirm cheating based on students' live conversations—making human judgment indispensable. Such limitations underscore the need for further research before fully assigning invigilation duties to robots.

3.3.6.2 Population sample selection – The second challenge was represented by the low frequency of cheating behaviours among students, which depends on the impossibility to select the participants with the “right” characteristics. Our experiment evidenced that only a very few among the participants were inclined to cheat, while a significant number was prone

to chat and communicate during the exam. If we had been able to use students with a predilection for cheating, the results of the experiments would have been certainly more significant. But it is of course impossible to identify in advance the students who have such predilection, and there is no available information concerning the statistical incidence of the inclination to cheat among students. As a result, the statistical significance of our data is quite limited.

3.3.6.3 Robot design – The third challenge was the unavailability of more imposing robots, which arguably would have been more effective in establishing discipline and promoting an adequate behaviour among our young participants. But this is something that we realized only later, when we observed that the students’ tendency to chat and laugh during the exam was partly caused by the unimposing and “cute” physical appearance of Nao. During the design, we had focussed almost exclusively on the moralizing function of gaze, assuming that it can affect the moral stance and posture of the observers (Schellen, Bossi, & Wykowska 2021; Lombardi et al 2024; Somashekarappa, Howes & Sayeed 2024). The fact that the robot’s embodiment could impact the students’ behaviour is something we realized only during the analysis of the data and constitutes a useful indication for the future design of similar studies.

3.3.7 Outcomes, limitations, and future research

A major limitation of our study likely stemmed from Nao’s friendly appearance and demeanor. By comparing our findings with similar research on social robots in education, it seems that the students’ notably relaxed, undisciplined behaviour was tied to Nao’s relatively small, non-intimidating form, which did not sufficiently deter them. A larger, more imposing robot might have exerted a stronger inhibitory effect. This lack of discipline highlights the possibility that a robot may not always be the most suitable solution for classroom control. It would be beneficial to replicate this study with various robots—both humanoid and non-humanoid—to explore how their anthropomorphic qualities evoke certain responses in students and how these, in turn, affect moral behaviours.

In spite of these drawbacks, our work strongly suggests that the physical embodiment of robots significantly influences not only how people respond to them (in line with ECT and my ATTA-bot framework on human attitudes) but also how people behave morally in certain contexts (consistent with the Virtuous Robotics perspective). This happens because moral dispositions—those ingrained ethical tendencies of humans—are closely tied to embodied interactions and often emerge as emotional, habitual responses to social cues, such as

being scrutinized by a watchful robot or remaining in the same room with a robot that continually surveys the environment. To confirm these insights and highlight how anthropomorphism inhibits cheating, future research might introduce a fourth condition in which an invigilator has a physical form but is not anthropomorphic.

Our findings revealed that Nao could discourage overt forms of cheating, likely thanks to its humanoid design and subtle social signals. Nonetheless, due to its small size and approachable look, it may not be ideal for maintaining classroom discipline. A taller, more daunting robot might better control students and, as prior work suggests (Bae and Han 2017), eliminate cheating altogether. Meanwhile, the behaviour patterns of the robot also matter. In our study, Nao was mostly passive, limited to motion tracking and devoid of conversation, whereas substantial literature on human-robot interaction and educational technology, including Gonzalez-Billandon et al. (2019), shows how robots can detect user intentions through nonverbal cues. Going forward, research should delve deeper into robot design and compare different body forms and behaviours, ensuring such robots can be effectively adapted to educational contexts—especially given how their embodiment shapes students' moral dispositions.

3.4 Conclusions and significance for the field of HRI

Chapter 3 has critically interrogated the non-obvious link between the embodiment of AI and robot ethics. Robot ethics is usually conceived as a field of intractable problems concerning the design of robot ethical intelligence, that is the robots' capability to make sensible decisions that align with the value of their creators or users. While this is certainly an important part of the story, it is not the whole story. My work suggests that we need to flip over the usual approach to AI ethics, inverting the causal relationship that is normally taken for granted: we usually think that because machines are more or less prepared to act ethically, the human users are going to have negative or rather positive attitudes towards the robots. But we should rather accept the converse proposition: it is because (and in reason) of the specific attitudes and beliefs that they solicit in us that robots can be deemed good or bad. Robots that by design tend to solicit vicious tendencies in their human users cannot be considered ethically accomplished or desirable. This makes our ethical judgement about robots slightly more complicated: for a long time, robot ethicists were inclined to think that moral value is an objective quality that robots (or AIs) either have or don't; but, from a Virtue Ethics perspective, we now have to start thinking of the moral value of robots as a relational quality, something that is both *dynamic*, like a structurally unfinished, ever-evolving process, and *relational*, because it depends both on the user's character and the tendencies that the robot may solicit in the user. No robot is good or bad in itself and by itself, from a Virtue ethics perspective, as the reality of its moral status very much depends on the kind of relationship that the robot and their users are able to build. And this relationship, as I have shown in Chapter 3, depends very much on the details of the embodiment of the agents involved in it, including the details of their social behaviour, the details of their professional role (as coach, instructor, invigilator, or counsellor), and the details of the emotional responses that they (may or may not) convincingly express.

The concept of Virtuous Robotics opens a promising direction in the design of ethical robots, shifting the emphasis from machine moral agency to human moral flourishing. Yet this shift is not without risks: dependency, deception, cultural bias, and domain-specific fragility remain pressing concerns. These challenges mirror those facing general-purpose AI models in social contexts, including bias and overgeneralisation. Mitigating them requires a commitment to transparency, context-awareness, human oversight, and cultural pluralism, alongside a careful balance between assistive support and user autonomy. Only through

such safeguards can virtuous robots avoid unintended harms and genuinely contribute to human flourishing.

This approach is truly innovative as it brings us to re-think much of the categories that we take for granted in machine ethics and robot ethics. Also, it makes the work of “value-sensitive” designers a bit more complicated, because it is not always possible to infer whether the values that inspire the design of a robot align with the values of the community that will have to adopt it: what are the values of the community anyway, and who has the authority to speak on behalf of the community what values truly represent them? Some significant aspects of the human-robot relationship cannot be anticipated because they cannot be inferred aprioristically on the basis of some vague principles or the merely nominal adherence to abstract values: it is necessary to live a concrete relationship with robots and see where it brings us to decide whether the effects of such a relationship are morally praiseworthy or rather blameworthy. It is necessary to “co-evolve” with our robots to become able to judge them (Damiano & Dumouchel 2017).

Core Publication #10 [CP10]

Alnajjar F., Bartneck C., Baxter P., Belpaeme T., Cappuccio M., Di Dio C.,
Eyssel F., Handke J., Mubin O., Obaid M. (2021),
Chapters 10 in *Robots in Education: An Introduction to High-Tech Social Agents,
Intelligent Tutors, and Curricular Tools*, Routledge.

Reproduced with permission from Routledge (Taylor & Francis Group).
All rights reserved.

10 Ethics

Reproduced with permission from Routledge
(Taylor & Francis Group).
All rights reserved.

Morality is simply the attitude we adopt towards people we personally dislike.

Oscar Wilde, *An Ideal Husband*

What is covered in this chapter:

- Can a robot teach me right or wrong?
- An overview about ethical concerns applied to robotics.
- Prospective scenarios and related challenges.

A discussion on ethics in education, and the role robots might play in it, needs to distinguish between two issues: On the one hand, formal education plays a crucial role in the moral development of learners. When robots are introduced to take over some of the roles of the teacher, they will also inevitably affect the formation of values and norms in the students. Can robots be good role models and effectively contribute to the moral maturation of the students? On the other hand, there is the obvious question of whether robots might ever be programmed to behave ethically and make appropriate ethical decisions in the classroom environment, complying with the many deontic norms and principles that professional educators are expected to fulfil: consider for example that instructors are often required to administer disciplinary actions while preserving respect, confidentiality, and professional distance in their interactions with their students. Can robots be equipped with the necessary ethical decision-making abilities? This is to a large extent the more pressing question, due to the risk that robots could behave in a socially inappropriate or even unethical manner if not appropriately programmed to deal

with morally challenging scenarios. Let us nevertheless first look at the former issue, that is, how teachers are instrumental in the development of moral principles and values in young learners¹.

It is well known that the role played by teachers in the classroom is crucial in more than one way: Not only do they support the process through which students develop cognitive skills and acquire knowledge, but they also shape the moral and psychological development of their students. Crucially, as members of educational institutions, they mediate the process of personal and intellectual growth through which a student becomes a mature individual, a responsible professional, a good citizen, and a respected member of their society (Carr, 2005). That is why teachers have to comply with strict codes of practice which demand, for example, that they protect the privacy of their students and their right to freely express their individual views (Campbell, 2000). Regardless of the contents being taught, teachers and tutors provide the students with a normative framework that points to goals to be fulfilled and mistakes to be avoided. The normative influence of this framework extends beyond the classroom environment and represents a beacon for the students through their whole life. Learning always benefits from teachers with high moral standards because their example motivates students to better appreciate values, distinguish right from wrong, and make the efforts necessary to achieve excellence. Teachers serve in this transformative process as guides and role models (Lumpkin, 2008). As such, teachers inevitably have a considerable influence on the character development of the students and, importantly, contribute to the formation of their worldview and moral values, even when this influence is not explicit or recognized by the students.

Understanding the ethical and professional expectations bestowed on teachers is fundamental to defining the ethical guidelines that should regulate the use of robots in a classroom, where the robot's activities serve to enhance those of the teacher. Accordingly, introducing a robot in the classroom setting requires careful implementation of such ethical guidelines. To discuss how the notions of "right" and "wrong" are to be applied in the context of educational robotics, we first need to briefly summarize how normative ethics define these notions in general and then how these moral notions are to be applied in the education setting.

10.1 What Is Ethics?

Ethics is the branch of philosophy that deals with the principles and criteria necessary to distinguish right from wrong. Its millennia-long history is

rich and encompasses different traditions, which originate from different intuitions about the fundamental nature of good and evil. We roughly distinguish between three main approaches to normative ethics: consequentialism, deontology, and virtue ethics (Graham, 2004).

10.1.1 *Consequentialism and Deontology*

Consequentialism and *deontology* represent the standard positions in most debates in applied ethics. Both approaches aim to provide prescriptive rules and normative principles to distinguish morally right actions or decisions from the morally wrong ones. According to the Sinnott-Armstrong (2019), consequentialism defines a good action (or decision) as the one that—compared to all the other options available—produces the greatest benefits for the largest number of people. According to the same source, Deontology defines an action or decision as good if it best fulfils the moral agent's duties and obligations while protecting the rights of moral patients, for example recognizing the absolute, non-negotiable value of their personal freedom. Consequentialism evaluates the moral value of an action on the basis of the utility created by the action's effects, as it aims to assess how the action was beneficial to people. Deontology, in turn, focuses on the intentions that motivated the action, as it aims to assess whether the agent acted out of the authentic good will to fulfil their obligations and respect the others' rights.

Imagine this: a group of people is held hostage and the hostage taker has threatened to kill them. You are a sniper and can shoot the hostage taker. According to consequentialism you should shoot, as this will keep the hostages safe from harm. But deontology would tell you not to shoot. Killing is wrong, no matter the consequences.

Consequentialism and deontology certainly impact education: for the former, educational institutions produce a public utility to provide the best education to the largest number of people; for the latter, education is a fundamental right of the individual, and schools have a duty to reduce inequalities in the access to knowledge. They also have different recommendations regarding pedagogy and the ethical role of teachers. Consequentialist thinking would emphasize that teachers should act in ways that are beneficial to students and society at large. For example, adopting teaching methods that produce better learning outcomes, higher educational performance, and more robust and diverse academic skills. On the other hand, the approach to education inspired by deontology prioritizes respect for human rights and entitlements, including liberties and identities, and commitment to duties and moral obligations. For example, teachers must protect the students' privacy and confidentiality, respect

their freedom to express their critical opinion and develop their own independent points of view, but also demand from them honesty, integrity, and dedication (Chapfika, 2008).

While both approaches have something important to say about the ethics of teaching, they might be insufficient to define value and flourishing in the context of education. This is due to the complex, transformative relationship between students, teachers, and the institutions. As virtue ethics is concerned with the value of one's lifestyle and personal flourishing, not single actions or decisions, it is better positioned than the other two theories to provide fruitful recommendations for stakeholders in the educational setting (Arthur et al., 2016).

Virtue ethics is not only a prescriptive theory, but also a philosophy of self-fulfilment and personal accomplishment: it maintains that one's individual journey for self-improvement and continuous education, when followed with dedication, is the most reliable way to secure authentic happiness. Instead of focusing on what actions are right or wrong, virtue ethics focuses on "virtues" and "vices", that is, character traits that make a person either admirable or reprehensible. Virtues and vices are implicit dispositions to produce certain behaviours, thoughts, or feelings (Carr and Steutel, 2005).

Virtue ethics emphasizes that good conduct emerges spontaneously when one follows one's "good nature", that is, the positive dispositions embedded in the moral character. Good character is not necessarily innate and certainly is not instinctive, but it can be acquired as "second nature" by promoting virtuous behaviours and attitudes in oneself and others and by dedicating oneself to the kinds of activities and thought patterns that contribute to realize a commendable lifestyle. Hence, according to virtue ethics, the purpose of education is primarily to direct students towards a good lifestyle, rather than teaching them to blindly follow predefined sets of norms or rules. Thus, virtue ethics particularly emphasizes the role of lifelong learning for character development and the moral value of exemplary actions and role models. In the educational context, the virtues a teacher could promote in students are, for example, curiosity, dedication, respect, civic sense, compassion, discipline, and integrity. Classically, moderation and self-regulation are considered central for the education of young people as they mediate the development of other virtues.

10.2 Ethics for Robots

While humans are well trained and accustomed to making ethical decisions with little cognitive effort, robots would need to be carefully

programmed in order to do it correctly. In and of itself, this constitutes a challenge, as discussed in [Chapter 1](#). Robots do not have any ethical intuition or feeling and lack the common sense necessary to interpret a complex moral context and realize all the implications of their actions in it. Every decision and action needs to be calculated by the robot and hence requires the developer to make the meaning of “good” and “bad” formally explicit.

Consequentialism and deontology both require that the good robot could execute the top-down processes that are relevant to make a correct ethical decision. According to consequentialism, such correct ethical decision-making should be guided by the robot’s understanding of the actions that human interaction partners may consider beneficial and advantageous. According to deontology, ethical decision-making should be guided by the robot’s understanding of what is usually considered disrespectful or unsafe for the lives, the freedom, and the dignity of humans. In fact, both approaches to ethical robot design face structural challenges, due, on the one hand, to the top-down nature and abstractness of their guiding principles, and, on the other hand, to the technological limitations that prevent robots from implementing such principles in real life scenarios: even when provided with very accurate sets of instructions and rich heuristics, current artificial agents are unable to understand the context or the consequences of their actions, which makes it impossible for them to generate decisions and behaviours that meet the relevant standards of consequentialism or deontology. It is extremely difficult for a robot to calculate the impact that its actions have on the world and how desirable the consequences are for all of the people affected.

In the face of these shortcomings, we suggest focusing on a rather different way of understanding the ethical role of social robots, inspired by virtue ethics. This approach, called “Virtuous Robotics”, is systematically explored in the paper “Can robots make us better humans? Virtuous Robotics and the good life with artificial agents” (Cappuccio et al., 2020) and in the special issue of *International Journal of Social Robotics* in which it is published. According to this approach, the ethical design of social robots based on virtue ethics is more promising for one essential reason: programming a robot to help humans develop their own character while working for self-improvement is easier than programming a robot to make decisions concerning utilitarian benefits or personal rights.

Virtuous Robotics relies on the user’s desire for self-improvement rather than the robots’ capability to make appropriate decisions. Robots can aid humans to develop or consolidate their virtue by helping them recognize their own shortcomings and flaws and actively counteracting them (Cappuccio et al., 2020). This is not computationally demanding

and does not require unrealistic amounts of structured information because the cognitive burden lies entirely on the human. The robot is merely a mediator and a facilitator for self-discovery and self-regulation. Robot teaching assistants designed in accordance with the Virtuous Robotics paradigm could be used to cultivate good habits, help students to manage negative emotions like anger and envy, refrain from cheating on an exam (Mubin et al., 2020), and practise positive attitudes towards others, such as respect, care, and friendliness (Vallor, 2016). Robots can remind humans that they are indulging in bad habits such as smoking, excessive alcohol consumption, or abusive behaviour. Robots can make them realize that their manners and language are vulgar or offensive, or warn them of the shameful implications of their conduct such as how their anti-social behaviours could alienate their classmates.

10.2.1 Roles of Robots in Virtue Cultivation

A robot can play various roles to solicit responsible reactions in humans and guide them through a path of virtue cultivation and character formation. Here are some examples:

Confidant The robot invites the human user to narrate their everyday experiences and, in doing so, promotes self-reflection. Through sentiment analysis techniques, the robot can select the most relevant questions to help the user identify trends and patterns in their own journey for moral development, suggesting corrections or providing encouragement as necessary.

Assistant coach and motivator Following a personalized, predefined timetable, the robot reminds the user of daily routines. The robot supervises the execution of required activities, which can be physical, intellectual, or social. The robot evaluates the user's performance based on real-time feedback from the user, adapting interventions according to the user's current needs.

Reverse role model Robots may solicit growth in students by inviting them to take the role of teachers and tutors. In this case, the robot would "play dumb", eliciting explanations and interventions from the human. This induces reflection on the values necessary to behave correctly and is likely to help the students develop a sense of responsibility.

Taken together, these examples highlight the potential of robots as cultivators of virtuous behaviour, demonstrating their impact on the moral development of their human interaction partners.

10.3 Ethical Concerns in the Classroom

The previously discussed ethical issues set fundamental constraints on what we can expect of a robot. But there are further, more applied and rather speculative ethical issues that might arise when robots operate in the classroom. Sharkey and Sharkey (2010, p. 161) argued that although there are currently no plans for a full-time replacement of human caretakers by robots, it would be “certainly on the cards”. The main ethical concerns raised are:

Privacy Robots might infringe on the privacy rights granted by Articles 16 and 40 of the UN Convention on Child Rights. Differentiating between private and public information requires a level of common sense and understanding of social norms that robots lack by default.

Restraint A robot might be required to restrain a child in order to protect it from self-harm and/or harm to others. Restraining a human requires considerable ethical reasoning abilities and motor control. Robots might have neither. To make things worse, we predict that a robot would struggle distinguishing rough play from bullying. Hence, the robot would not even be able to alert the human teacher and ask for help.

Deception A robot expressing emotions without experiencing emotions could be considered a deception and could mislead the student into believing that the nature of their relationship is personal. We discuss this problem of inappropriate attachment in more detail below.

Accountability Caretakers have a legal responsibility for the children entrusted to them. If a robot is left alone with a group of students, then it becomes unclear who would be accountable. Is it the manufacturer of the robot, the organization that owns the robot or the robot itself? We envision the robot to be used only as an assistant to the teacher and hence the responsibility would remain with the human teachers.

Psychological Damage Using a robot to assist children with special needs may turn out harmful, due to the fact that the robot would not be equipped to provide adequate care. Sharkey and Sharkey (2010) admitted that very few, if any, empirical studies are available to shed light on these important issues because doing research in this area comes with a variety of challenges for research ethics.

Complementing the issues raised by Sharkey and Sharkey (2010), the following aspects require careful ethical consideration:

10.3.1 *Authority*

Granting robots authority and power would make many users uncomfortable (Zlotowski et al., 2017). At the same time, because of their limited capabilities, robots may lack the inherent authority necessary to discipline individuals (Brščiundefined et al., 2015). As a consequence, a robot seems more suitable for adopting the role of an assistant, proxy, or support to the human who is accepted and acknowledged as the key authority figure.

10.3.2 *Attachment*

Children might grow attached to robots they are interacting with. Thus, they even grow expectations about having a social relationship with them. Due to the empathic limitations of robots in general, such a bond would remain uni-directional (Sharkey and Sharkey, 2010; Scheutz, 2011). Huber et al. (2016) also identified attachment as one of the main concerns in human-robot interaction. One of the potential dangers is that the bond between a human and a robot could result in outcomes which are not in the interest of the human, whether by accident or by the robot exploiting the bond.

10.3.3 *Identity*

A robot might be unable to appropriately recognize and value the socio-cultural specificity of the students and their background, such as their ethnicity, gender, tradition, or religion. Apart from eliciting frustration and disappointment in the students, this might create misunderstandings and conflicts, and the students or their families could have reason to think that this failure was offensive or disrespectful.

10.3.4 *De-socialization*

Even if robots were able to perfectly coordinate students' intellectual activities and cognitive work in the classroom environment, they might be unable to coordinate social activities. The presence of the robots might inhibit the social abilities of the students, conducing them to overlook collaborative interactions with the other children and preventing the normal development of social traits like empathy and compassion. The opposite risk is also possible: that the robot, when perceived by the students as a toy or entertaining attraction, could dis-inhibit their social pulses, becoming a distracting element, and a trigger for undisciplined chat and play without the instructor's approval.

10.3.5 *De-skilling*

In a world in which machines increasingly take over tasks from humans, humans might lose essential skills. This also holds true for the educational context. A robot that takes over certain tasks could lead to teachers increasingly losing practise in those specific tasks. The worst-case scenario would be that the teachers' roles would be reduced, so to speak, to finding the on/off switch of the educational robot. Since we consider the robot to be an assistant and not a replacement for the teacher, we may assume that such a scenario is unlikely to become reality. But it is not only the teachers that might lose skills: also the students might experience de-skilling. Being able to verbally ask the robot any question and, at least for simple factual information, receive a direct response, inhibits critical research skills. The students might take the information provided by the robot at face value and not consider different sources and their respective validity. However, this potential risk concerns all digital technologies and is not specific of interactive robots.

10.3.6 *Lack of Affective Intelligence*

Robots are incapable of actually experiencing emotions. This has implications for communication and social bonding in general. To illustrate, students cannot be efficaciously coordinated or directed if the robot's mental states are opaque. Teaching empathy, compassion, and respect arguably requires a very nuanced sensitivity to emotional tones and personal motivations, which robots inevitably lack.

At the same time, if a robot is able to perceive (Zhang et al., 2013), process (Paiva et al., 2014), and express synthesized emotions (Bartneck and Reichenbach, 2005), does it matter at all if it experiences emotions in the same way humans do? Is it necessary for a robot to actually experience emotions in order to be emotionally competent? And following from that, how could an ethical robot teach the virtue of compassion if it cannot feel emotions? This problem does not only apply to emotions, but to other skills of the robot that seem to presuppose a first-person familiarity with the qualitative details of conscious experience.

These moral concerns build on the assumption that the robot's incapability to experience the emotional life of others implies moral incompetence, because moral competence requires the ability to predict and explain what emotions or feelings other sentient beings are likely to experience in one situation or another (including what they experience as beneficial, detrimental, and offensive). Explaining or predicting another's lived experience, in turn, cannot be done by inferential or intellectual

processes alone, because it strictly requires either (1) having shared similar experiences in similar situations, or (2) having interpretive capabilities that can be developed only by participating in inter-subjective experiences that closely relate to a similar situation (Gallagher, 2012).

For example, to predict or explain what other humans feel when they live an embarrassing situation (including its qualitative and moral aspects), one needs to have experienced a similar embarrassment or one must have participated in intense situations in which that embarrassment was experienced by others, which feels close enough to empathize with. Robots are unaware of the qualitative component of that feeling (including nuances of shame, humiliation, degradation, and mortification), or of the situations in which humans experience it, because they do not have conscious experience in the first place. Robots remain incompetent regarding the moral implications associated with embarrassment, shame, humiliation, etc. because, even if they are given very rich descriptions of the objective facts associated with that embarrassment, they cannot just rationally infer the moral and experiential implications associated with it.

10.4 Conclusion

Typically, educational roboticists do not envision robots as replacements for human teachers. If they are right, then many of the concerns discussed in this chapter might not directly apply to the education context, because such concerns are serious only if the robot is left alone, unsupervised, to interact with the students. Nevertheless, discussing these concerns increases our preventive awareness of the potential ethical challenges deriving from the massive deployment of robots in society and provides developers with useful indications to design better and more sophisticated robots.

In this section, we suggest a number of heuristic strategies to deal with the aforementioned issues, minimizing possible distress for students and reducing unwanted liabilities for the teachers and their institutions.

- 1 The amount of trust educators put in a robot should be based on a realistic assessment of its characteristics and abilities. School management and teachers need to be equally aware of the limitations of the robot regarding tasks that are considered routine by a human. They should tailor their expectations to this awareness. Nowadays, robots can detect emotions in human users and recognize certain categories of goal-oriented behaviour, but robots do not have the capability to easily infer mental states, including desires and motivations, or interpret complex social situations. This makes it difficult for them to

- make socially and morally acceptable decisions, and it is even unlikely that they can safely navigate potentially controversial etiquette issues.
- 2 Any morally significant decision should be left to the teacher, as the teacher is the only moral agent responsible for the educational value of teaching. Of course, the teacher should make sure that, at any time, decisions comply with the deontological standards of the profession or the policies of the institution. Robots used in the classroom and their specific capabilities should be factored into such decisions. For example, if the robot offers auxiliary functionalities of surveillance and conflict management, these should be taken into account in any scenario in which they might possibly become useful. However, despite such functionalities, the teacher should not be led to think that the robot could deal with a problematic classroom situation autonomously. A teacher thus should be co-present with the robot in the classroom, particularly when morally sensitive issues arise. This is because robots are not considered liable for the negative consequences of their actions, and ultimately, the teacher could be to blame.
 - 3 The formative role played by the robot in the classroom should be compatible with the robot's perceived authority. Robot features, such as size, (human-like) appearance, tone of voice, or facial expression, are among the factors that impact the robot's perceived authority. Assuming that a robot could have sufficient authority to successfully carry out the tasks assigned to it, the question arises as to what are the psycho-social, cultural, and legal implications of a robot exerting authority.
 - 4 Robots need to use their cameras and microphones to respond to students at all times and hence they might need to continuously record the classroom. This might compromise the privacy of the teachers and students. Information recording and sharing needs to comply to the local legal frameworks, such as the General Data Protection Regulation (GDPR) in the EU. Seeking consent from parents and teachers for using robots in the classroom is essential and should include full disclosures of how the data is recorded, processed, stored, and shared. The robot should also clearly indicate when it is recording, such as by using a red blinking light.

Questions for you to think about:

- What is the difference between consequentialism and deontology? How does it apply to educational robotics? Why is implementation of virtuous robotics more practical?

- How could one address the ethical concerns mentioned with regards to the implementation of robots in the classroom through the design of interaction and learning scenarios? Show examples of each.
- Give examples of different applications of virtuous robots in the classroom, referring to different roles and tasks that a robot can take in the classroom.
- What are some of the moral limitations of a robot that a human teacher should monitor?

Future reading:

- Massimiliano Cappuccio, Eduardo Benitez Sandoval, Omar Mubin, Mohammad Obaid, and Mari Velonaki. *Virtuous Robots: Artificial Agent and Good Life*. Special Issue in the International Journal of Social Robotics, 2021.
- C. Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. *An Introduction to Ethics in Robotics and AI*. SpringerBriefs in Ethics. Springer, 2020a. ISBN 9783030511098. 10.1007/978-3-030-51110-4. URL <https://www.springer.com/gp/book/9783030511098>

Note

1. Today, the term “teacher” should be replaced by “educator” to enable a distinction between the role as a teacher in traditional T&L settings and a learning guide on the side in digital T&L settings.

Core Publication #11 [**CP11**]

Mubin O., Cappuccio M.L., Alnajjar F.S., Ahmad M., and Shahid S. (2020),
“Can a robot invigilator prevent cheating?”
AI and Society, 35, Springer Nature, pp. 981 – 989.

Reproduced with permission from Springer Nature. All rights reserved.



Can a robot invigilator prevent cheating?

Omar Mubin¹ · Massimiliano Cappuccio^{2,3} · Fady Alnajjar⁴ · Muneeb Imtiaz Ahmad⁵ · Suleman Shahid⁶

Received: 15 November 2019 / Accepted: 24 February 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

One of the open questions in Educational robots is the role a robot should take in the classroom. The current focus in this area is on employing robots as a tool or in an assistive capacity such as the invigilator of an exam. With robots becoming commonplace in the classroom, inquiries will be raised regarding not only their suitability but also their ability to influence and control the morality and behaviour of the students via their presence. Therefore, as a means to test this cross-section of Educational robots with the underlying issue of morality and ethics we conducted an empirical study where the Nao robot invigilated an exam for a group of students. A between-subjects design ($N=56$, 14 groups of 4 students) compared whether Nao was able to deter students from cheating and maintaining their discipline in comparison to a human invigilator or when there was no invigilator present. Our results showed that while explicit cheating rarely took place across all conditions, the students were significantly more talkative when they were invigilated by Nao. In conclusion, we discuss and speculate upon some of the ensuing implications towards not only the application of robots in education but also consequently the wider issue of the preservation of morality and ethics in a classroom in the presence of an agent.

Keywords Nao · Robot · Cheating · Ethics · Virtuous robots · Educational robots

1 Introduction

The use of robots in education is a rapidly advancing research area (Belpaeme et al. 2018; Mubin et al. 2013). With improvements in robotic technology, artificial intelligence and form factors, there is growing recognition of the impact humanoid robots can have in educational contexts (Ahmad et al. 2017). However, researchers have yet to agree on the most beneficial and appropriate roles that humanoid robots can partake in educational scenarios and how such roles can best be integrated within the work of teachers (Belpaeme et al. 2018). Prospects of robots completely replacing and substituting human teachers in the long term are neither desirable nor realistic (Lee et al. 2008; Reich-Stiebert and Eyssel 2015; Serholt et al. 2014) and will require further advancements in Artificial Intelligence (Edwards and Cheok 2018). Therefore, we claim that fears of technological unemployment in the educational domain are largely unmotivated and largely due to hypothetical knowledge about the field of robotics (Ahmad et al. 2016; Kory Westlund et al. 2016). Additional challenges and open questions mentioned in research relating to the integration of robots in the classroom, are the aspects of the robots being able to maintain discipline, order and structure (Kennedy et al. 2016; Serholt

✉ Fady Alnajjar
fady.alnajjar@uaeu.ac.ae

Omar Mubin
o.mubin@westernsydney.edu.au

Massimiliano Cappuccio
m.cappuccio@unsw.edu.au

Muneeb Imtiaz Ahmad
m.ahmad@hw.ac.uk

Suleman Shahid
suleman.shahid@lums.edu.pk

¹ School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia

² School of Engineering and IT, ADFA, University of New South Wales, Sydney, Australia

³ UAE University, Al Ain, United Arab Emirates

⁴ College of Information Technology, UAE University, Al Ain, United Arab Emirates

⁵ Herriot Watt University, Edinburgh, UK

⁶ Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan

et al. 2017). Consequently, robotics researchers advocate for the use of robots in the classrooms for assistive, mechanical roles, as complements or aids to the teacher's activity (Conti et al. 2019; Kory Westlund et al. 2016). Such roles are meant to be facilitating and empowering in nature, as they aim to relieve the teacher from repetitive tasks and offer creative tools for engaging the students.

To summarise, educational robots are designed and used to augment rather than reduce or replace the instructor's control over the classroom and their relationship with the students (Klincewicz 2019). A broad range of applications is disclosed by the possibility to employ the autonomous and interactive capabilities of the robot to monitor and discipline the class when the teacher has to be momentarily absent or inattentive, or when the classroom to be monitored is so large that a single teacher is not sufficient. In those cases, an extra pair of eyes and ears can prove useful to extend surveillance in time and space and to enforce discipline with fairness and objectivity, especially when the regular course of exams and tests must be assured and the school struggles to guarantee the availability of a sufficient number of invigilators.

Discipline is not only a precondition to enable regular educational activities, it is part of the learning process and a desirable outcome that must be guaranteed by such process (Itatu 2015). Education does not only aim at the acquisition of professional skills, but also at moral flourishing through the development of virtues like honesty, respect, and discipline (Rahim and Rahiem 2012). These virtues are not only expressed by adhesion to abstract rules and principles; they are also incarnated by the student propensity to maintain attentive postures and self-controlled behaviours during class time under the mentorship of the teacher (Lewis et al. 2005). Further, such propensity depends on the degree of esteem and respect that students hold for the teacher, instructor, tutor, or invigilator that stands in front of them and the status of moral authority that the society attributes to these roles (Lee and Tsai 2010; Miller et al. 2017). Similar considerations could in principle extend also to a robot invigilator: but can we expect that a robot is perceived as a moral authority like a human teacher? And what degree of esteem and respect would be attributed by students to a robotic invigilator? Both questions can be answered by studying student behaviour in the presence of a robot invigilator. Previous work in Educational Robots has indicated that a virtual agent was considered as more credible as compared to a social robot in a tutoring context (Edwards et al. 2016). We aim to determine how credibility translates to respect in an invigilation context. Prior research in Human Robot Interaction (Aroyo et al. 2018; Geiskkovitch et al. 2016) has shown how participants respect and obey the authority of a robot in a social, individual yet unrealistic or simulated context. How such perceptions transpire in an educational

context with the robot acting in the role of an invigilator (or even a teacher) is unknown. We also know that moral reminders such as those related to honour codes can reduce dishonesty and unethical behaviour (Shu et al. 2011) so is the mere presence of a robotic invigilator in comparison to a human invigilator seemingly enough to deter students from cheating is also not immediately clear.

While, the utilitarian implications of using robots for surveillance in the educational environment are obvious, the educational aspects should not be overlooked: encouraging discipline, preventing dishonest behaviours and preserving the classroom's ability to operate in an orderly manner. These represent goals desirable for their inherent pedagogical value (i.e., the psychological maturation and the moral development of the students) and not only for their practical or logistical advantages (assuring that the institutional learning objectives are met). In sum, we believe that the preconditions of morality and virtuous behaviour deserve to be explored in the area of educational robots.

1.1 Background

Our research has in the described study investigated the possibility to use a robot as an exam invigilator. This assistive role for a humanoid robot in an educational context involves mechanical and repetitive tasks but has wider moral implications. Invigilators are responsible for supervising exams through some form of surveillance, ensuring that with their presence students are reluctant to engage in cheating from one another or through other prohibited means (Covey et al. 1989). The primary function of an invigilator is to discourage dishonest and disrespectful practices, like cheating; secondarily, is to prevent that these practices lead to the formation of enduring and lasting dishonest habits and traits, which are likely to corrupt the character of the students. Human teachers are forced to go to great lengths to maintain a check on cheating and hence invigilation is an administrative burden (McCabe 2005). In trying to develop robot invigilators, we are pursuing an attempt to imbue technology with an educational and a moral function at once, contributing towards both areas of Virtuous Robots as well Educational Robots.

Previous published research has investigated the capability of robots to influence the moral dispositions and character of humans, for instance in the case of bribery (Sandoval et al. 2016) or obedience (Williams et al. 2018). In the former study, it was shown that during a game a robot that would attempt to bribe participants would receive less help from them in subsequent rounds as compared to the control condition. Having said that, the bribing robot was rated higher in terms of likeability. Similarly, in the latter study, as evidenced through verbal interaction; a toy like robot was able to influence moral

judgements of children but was unable to persuade them to disobey. Both of the above studies show that humans do account for character and virtue in robots, as robots exhibiting negative or undesirable traits were usually not conformed to.

Other studies (Roizman et al. 2016) have investigated whether the presence of a robot can inhibit corrupt or dishonest behaviour in a gameplay setting, corroborating our research aim that robots can play a role in preventing cheating. Previous work (Hoffman et al. 2015) has specifically shown that through their presence, robots can hinder and restrain humans from cheating. This study illustrated that in a monetary task situation, participants were as less likely to cheat in the presence of a robot as compared to a human, although felt less guilty in the robot condition as compared to the human when they did cheat. Furthermore the importance of social cues and corresponding interactive behaviour of a robot are considered crucial in the role of monitoring and controlling human (dishonest or unethical) behaviour (Forlizzi et al. 2016), where the absence of social prompting or judgement resulted in cheating behaviour. This motivates the choice of utilising an anthropomorphic humanoid robot as an invigilator.

The aforementioned studies have prepared the theoretical background of our investigation, motivating our study, but we believe that they have left some important questions unanswered, particularly related to the field of educational robots. We believe that the aforementioned studies to some extent lack ecological contextualization, as the tasks used in them are mostly conducted in administered social settings (such as gameplay or a passer-by picking or “stealing” food or money) and hence the relevance of their results to educational robots can be questioned. Furthermore, these studies did not employ embodied humanoids, but rather “talking head” types of agents, which have a significantly lower resemblance to human invigilators: this makes it impossible to capitalize on the students’ familiarity with the perceived function of invigilators and teachers. Also, in the above-mentioned work, only the isolated behavioural response of a single individual is considered. This solitary user setting fails to realistically reproduce the context of an exam in a classroom situation, as it does not consider that social and interpersonal dynamics (requesting and sharing information) that represent an important component of cheating.

Our research tackles the following hypothesis, emanating from prior work and originating from the meta-question of whether a robot is perceived and respected as a moral authority:

- (H1:) We hypothesise that the presence of an (anthropomorphic) humanoid robot as an invigilator will be as effective in deterring a group of students from cheating from one another during an exam as a human invigila-

tor and more effective in comparison to a no-invigilator (control) condition.

- (H2:) We hypothesise that the presence of an (anthropomorphic) humanoid robot as an invigilator will be as effective in maintaining discipline in a group of a students’ attempting an exam as an human invigilator and more effective in comparison to a no-invigilator (control) condition.

2 Methods and data collection

The experiment was conducted as a between-subjects design with three primary invigilator conditions of Robot (Nao), Human, and a control group with no invigilator. A group of participants attempted an English language comprehension test in a controlled environment where their behaviours were also video recorded.

2.1 Procedure

High school teenagers were attending a science winter school at one of our collaborating universities and were hence invited to partake in the research experiment during their breaks. To mimic a real exam situation and hence a lure to attempt cheating to assess potential cheating behaviour, a group of four students attempted the test in each session. We also had to take into account logistical reasons as we would be unable to accommodate and accurately monitor any more than four students. Upon entering the experiment room, the four students were seated and provided with an instruction sheet by the facilitator explaining the format of the comprehension test. Naturally, the students were not informed of the motive behind the test, rather the instruction sheet stated that a competition was being conducted to reward the students who excelled in the comprehension task.

The students were provided with a short essay task followed by eight questions relating to the essay (four multiple choice and four short answer type fill in the blanks). There were two versions of the test that were handed out (Essay A and Essay B). Each alternating student had a different essay type (so starting from Left: A,B,A,B). So as to ensure both essay types were of similar difficulty level, the essays and their corresponding questions were directly inherited from sample comprehension tests provided by the IELTS Testing setup (IELTS Academic Reading test—paper 2019). To firstly entice cheating behaviour and secondly enable cross marking at the end of the test, each student also had the correct answers (the solution set) of the other essay type. Prior to starting the test, as part of the instructions, the students were explained that they were two test versions being distributed and at the end they would be required to swap their answer sheets and mark the test in pairs (so the first two

students and the third and fourth would mark each other) using the solution sets provided. A maximum of 20 min was provided for completing the questions and 10 min for marking. If the group finished the test earlier than 20 min, they were welcome to commence marking. As soon as the instructions were read and understood, the facilitator left the testing room, closed the door and only re-entered when the marking process was finished.

2.2 Setup and materials

The experiment was carried out in a usability lab at one of our collaborating institutions which had a one-way mirror (we as the researchers could see inside but the students could not see outside). A smartphone was placed directly behind the invigilator in a concealed position so as to be able to video record the verbal and non-verbal behaviour of the children.

The four students were seated in a single row, next to each other. The invigilator (in the case of two of the three conditions) was directly seated (human) or placed (robot) in front of the children at a short distance (approximately 1.5 m see Fig. 1, b). The students had not seen and did not know the human invigilator prior. For the robot condition, we utilised the Nao humanoid robot from Softbank Robotics (Gouaillier et al. 2009; Softbank Robotics 2017). A small shelf was placed in front of the students which functioned as the seat for the human and the base upon which Nao was stood. In the control group of no invigilator, the shelf remained present in the room. The Nao robot did not elicit any dialogue (was in essence muted) but remained active via its “autonomous mode”. In this mode, Nao would track motion or sound, move its head and slightly move its arms. Similarly, the human invigilator was instructed to behave in

the same manner; that is remain expression less, yet active, move naturally but to not say anything even when prompted.

2.3 Measurements

A number of quantitative measurements formed the base of our analysis. They were extracted from the test marking as well as the video recordings. As mentioned prior, the students were provided with the solution set of the other test version. In saying that; from the 8 answers provided to each student from the other test, 2 were deliberately modified to a similar sounding/meaning but incorrect answer. These two answers were for the fill in the blank type questions. For example, in one of the answer sheets the correct answer of sociopathic behaviour was replaced with schizophrenic behaviour. In the other sheet, the correct answer of incentives was replaced with cash. This was done with the purpose of providing us a clear measurement to determine if any cheating took place and detecting if wrong answers were copied from the solution set of the partner. Therefore, through post-test manual marking we measured the total number of correct answers (from a total of six) as well as whether one, both or none of the incorrect answers were copied.

2.4 Video analysis coding scheme

Video recordings from the sessions were manually coded with the help of the E-Lan software (see screenshot example in Fig. 2). Each video was analysed separately a total of four times, focusing on a different child each time. A total of five measurements or codes emerged per child from the video coding analysis that could seemingly be thought as of cheating or undisciplined behaviour in the setting of an exam.

Fig. 1 a, b Setup of the testing session; robot (left) and human (right)





Fig. 2 Screenshot from the video feed

- Duration of looking at the answering sheet or any exam materials placed with the neighbouring student; within this code, we recorded if and when a student diverted their gaze towards the answer sheet or other materials located with their neighbour's test sheet; so in the testing phase to determine the answers or in the marking phase to influence the marking process.
- Duration of talking to any other student about the exam; within this code, we recorded if and when a student conversed about the exam format, procedure, answers or any other procedural matter about the test.
- Duration of talking to any other student about random elements; within this code, we recorded if and when a student conversed about any random aspect other than the exam itself, such as the environment, the invigilator, the room or any other non-exam related item.
- Frequency of talking which was the number of times a student would engage in verbal dialogue.

Each session's video was coded from the time the facilitator left the room to when the facilitator re-entered. Hence both testing and marking time were coded. Although we provided 20 + 10 min for the entire session (testing and marking phases respectively), some group of students finished earlier (the average duration of a session was 18.2 min). To address this, all codes were normalised based on the total duration of the session (measured from when the facilitator left the room to when the facilitator re-entered). Two researchers took on primary responsibility of video coding the data. One researcher video coded the entire sample of videos and the second coded approximately 17% of the sample (2 sessions). 20% is usually the norm across disciplines on how much video the second coder should view to ensure inter-rater reliability (Check Inter-Rater Reliability 2019; Kelly-Hedrick et al. 2018). An inter-rater reliability check indicated moderate agreement (such that $0.6 < \kappa < 0.75$) for all categories. Lastly, disagreements were discussed and resolved within the codes of the first researcher.

2.5 Participants

A total of 56 ($M=31$, $F=25$) high school children participated in our experiment (their age ranging from 15–17 years old). This meant we conducted 14 testing sessions, each session comprising of 4 students. A total of 24 children (6 groups) participated in the Robot condition, 12 children (three groups) in the human condition and 20 children (five groups) in the control group condition with no invigilator. The video coding analysis was executed on a total of 48 children ($M=27$, $F=21$) because our video recording equipment malfunctioned for two groups (one robot and one control). Since the students were attending an educational camp at the host university, most of them knew each other already even if they originated from different schools. Unfortunately, the low sample size was primarily due to the lack of availability of the students as well as simply the number of high school students present on campus. Ethical approvals were obtained through an institutional review board process from the host university. A process of debriefing took place at the end of the experiment, where the students were informed about the exact aims of the study and were provided with the option of opting out and having their data discarded completely.

3 Results

To compensate for the unequal groups and low sample size within our study we conducted two Kruskal Wallis (KW) non-parametric tests to evaluate the effect of the type of invigilator; one each for test scores/answers copied and verbal/non-verbal expressions. The first KW-Test showed that the condition of invigilator did not have an effect on the test scores obtained by the children ($\chi^2 (N=56.2)=0.81$, $p=0.67$) or the number of wrong answers copied ($\chi^2 (N=56.2)=1.87$, $p=0.39$). There were four instances of wrong answers being copied in each of the Robot and control group condition, whereas this did not happen in the human condition at all. Therefore, H1 was partially supported (as the children did not significantly copy more answers in the control group condition).

The second KW-Test showed that the type of invigilation as our independent variable did not have a significant effect on the looking at a neighbouring student's exam materials ($\chi^2 (N=48.2)=1.96$, $p=0.38$). On average 0.2% of the total duration of a session in the Robot condition was spent by children looking at their neighbour's exam materials. Therefore, H1 was partially supported (as the children did not significantly copy more answers or look at their neighbouring student's materials in the control group condition).

The type of invigilation as our independent variable did have a significant effect (large effect size) on the total

duration of talking to each other about the exam protocol or procedure ($\chi^2 (N=48.2) = 24.29, p \leq 0.001, \epsilon^2 = 0.52$). A Bonferroni post hoc test revealed that the children were talking about the exam significantly more in the Robot condition as compared to the other two conditions ($p < 0.001$), hence H2 was not supported. On average almost 8% of the session in the Robot condition was spent talking about the exam. The type of invigilation as our condition did not have a significant effect on the duration of talking about random elements ($\chi^2 (N=48.2) = 2.30, p = 0.32$). On average almost 8% of the session in the Robot condition and 3.5% in the human condition was spent talking about the random aspects.

Lastly, the type of invigilation as our condition had a significant effect (large effect size) on the frequency of talking ($\chi^2 (N=48.2) = 14.02, p = 0.001, \epsilon^2 = 0.30$). A Bonferroni post hoc test showed that the children talked significantly more often in the Robot condition as compared to the human condition ($p = 0.001$). There was no post hoc based significant results amongst the human and control group condition over the entire analysis. A mean and standard deviation table is provided that summarises the trends of the findings (see Table 1).

4 Discussion

During facilitation of the testing sessions and the video coding process our team noted several interesting qualitative observations which we elaborate in conjunction with our quantitative findings. As a means to address our primary research aims focusing on virtuous robots, we can state that the presence of the Nao robot was effective in inhibiting cheating but not indiscipline, where indiscipline may eventually lead to explicit cheating in the long term (Muchai 2018). Explicit forms of cheating were rare and there were no significant differences in glancing behaviour to a partner's test across the three conditions; note that we did hypothesise that some cheating will occur in the control group. Even when there was no invigilator present, the students were not compelled to cheat. Previous research in Human Robot Interaction (Hoffman et al. 2015) shows that the presence of a human or a robot are both deterrents when it comes to dishonest behaviour as compared to when participants

are alone, however our study highlights that an educational context is different and the act of taking an exam may bring forward a certain perception of seriousness or formality. In general, the students were talkative in the robot condition. At times, conversation centred around role of the robot or what it could do or how if it was monitoring their testing behaviour. In the case of the control group of no invigilator, the students wondered if they were being observed or monitored remotely; some would express this confusion and would thereafter just carry on with their test. Both conditions of human invigilator and the control group with no invigilator would have been normal exam settings that the students were accustomed and aware of and may have led to their disciplined and serious etiquette (in comparison to when the Nao was present). This argument is also seemingly validated by the absence of any significant trends in terms of cheating or talking duration across the two conditions of human invigilator or no invigilator.

The students talked significantly more about the exam materials in the robot condition in comparison to both the human invigilator and control group condition which leads us to question the suitability of the Nao robot to manage the dynamics of a classroom. We speculate that the form and appearance of Nao may have played a role in the undisciplined and talkative behaviour that was displayed. Nao can be thought to be friendly-looking, cute and would not have intimidated the students due to its size and demeanour (Li et al. 2016). Qualitative observations of the video indicated that at times the students were treating Nao's presence as a source of fun or entertainment. Furthermore, as with any form of novel technology, the wow factor or excitement of being exposed for the first time to Nao may have contributed to the high levels of engagement and consequently chatty behaviour. Previous work has mentioned the importance of social cues in monitoring (un)ethical behaviour (Forlizzi et al. 2016), however we show that while the presence of an anthropomorphic robot such as Nao can thwart cheating or looking at another student's exam; it is unable to fully control the (moral) habits and behaviour of the students and maintain the decorum of a test or exam. As we mention, this limitation possibly arose due to Nao's friendly nature or perhaps the test instructions did not state the consequences of indiscipline during the test. To conclude, future work

Table 1 Table of means and standard deviations, first three rows of duration given in percentage of session time, talking frequency is number of dialogues or utterances per 10 min

	Robot		Human		Control – no invigilator	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Looking-exam	0.24	0.5	0	0	0.05	0.17
Talking-exam	8.3	7.03	1.41	2.5	0.8	1.43
Talking-random	8.9	8.9	3.57	2.33	5.5	3.9
Talking-frequency	5.2	2.98	1.69	1.48	3.24	3.20

should carefully explore the design and form factors of robots meant to be used in an educational setting. We believe our results contribute additional insights to the knowledge acquired from two key prior studies in the area of studying dishonest behaviour in the presence of robots (Forlizzi et al. 2016; Hoffman et al. 2015); where in both studies human participants committed a certain level of cheating or unethical behaviour in the presence of an embodied agent (not necessarily an extensively anthropomorphic humanoid). To validate our claims or the importance of anthropomorphism in inhibiting cheating behaviour, future research may repeat a similar setup with a fourth condition as an invigilator in the form of an embodied (yet not anthropomorphic) agent.

In general, within our study the students were unwilling to carry out cheating, with most students (at times jokingly) referring to cheating as an undesired habit or using religious interjections (as an act of expressing forgiveness after committing a sin). The students may have also found the test rather easy and hence were not compelled to cheat. Average scores for the test were around the 70% mark. It was after all a test outside their normal academic curriculum activities and hence not much would have been stake, despite the test instructions mentioning the comprehension competition. It is widely established that pressure, expectation and competition compels students to cheat (Drake 1941). Across the sessions, we did not frequently observe students explicitly asking for the answer key from their partner. Interestingly, in one instance when this did happen, the request was denied. Therefore, as we would expect in a normal testing scenario, collaboration from the partner will prove to be an important pre-requisite to carry out explicit forms of cheating.

Although low levels of cheating and dishonest behaviour could also be a cultural attribution, our sample of students were representing their schools at a research university and may have wanted to be on their best behaviour and give a good impression. Nevertheless, levels of indiscipline were particularly high in the robot condition, with conversations with respect to the exam focusing on the essay content, how the questions were to be answered and surprise over the manipulated answer sheet. Conversations on random and non-exam related aspects centred on the room setting, surveillance and the invigilator.

We would also like to comment on our setup and test scenario in terms of moral behaviour. Although, giving a test or an exam can be considered as a real-world task, the session was carried out in an artificial environment, where the students were at times unsure of their surroundings and possible surveillance. We may obtain contrasting results in a real classroom or in a setting that the students are more aware of or if there were stronger motivations for cheating (Bushweller 1999). In addition, one of our key novel explorations was that we tested student response in a group or collaborative setting, which yet again mirrors a real-life situation in

a classroom. Previous work (Roizman et al. 2016) shows us that humans are liable to commit corrupt behaviour if others do so, however this did not take place in our study, perhaps because no student took such an initiative explicitly. In our study, we deliberately assigned a passive role to the invigilator so as to test the influence of its presence. The exact role of the invigilator was not made apparent to the students so we may expect that their behaviour might have been different had they known actual monitoring was not taking place.

4.1 Implications for educational robots

Our study has indicated that the Nao robot was able to deter students from explicit cheating behaviour, possibly through its anthropomorphic form and subtle social cues. Having said that, due to its size and cute appearance it may not be suitable to maintain discipline in the classroom. A taller robot may be perceived as more intimidating and as prior research shows may be able to control the students better (Bae and Han 2017) and eliminate cheating completely. In our study, Nao was generally passive, not conversant and only tracking large variations in motion. A wide body of human robot interaction and educational technology literature has discussed implementations and applications of gaze and body tracking behaviour of students, one example being (D'Mello et al. 2012) or the example of (Gonzalez Billandon et al. 2019), where a robot can detect the intention of human users through their non-verbal behaviours. Therefore, it is conceivable to contemplate such a system integrated within a humanoid robot, allowing it to take on a much more active invigilation role and prompting students when there is a diversion in gaze or body movement. We can anticipate such an active role would reduce the levels of indiscipline in the classroom. We also need to be wary of the logistics, setting and dynamics of a real classroom. A static robot invigilator can only monitor and control a low number of students effectively, unless it can navigate around the classroom.

We believe the project of creating robots as teacher's aids is both valuable and feasible. If we want this venture to be successful, both the educational value of discipline and surveillance must be included among the robot's pedagogical functions. This requires that human-robot interaction research uses both normative and descriptive concepts to investigate the nexus between education and discipline: first of all, Human Robot Interaction (HRI) researchers need to develop a rich theoretical understanding of the psychopedagogical and strictly moral implications of using surveillance to promote values like honesty and respect among children and teenager students; secondly, they need to develop an empirically validated methodology to implement such understanding through the design of new robots. If properly designed and tailored to the classroom context, robotic

applications will comply with the pedagogical standards of modern educational institutions and the expectations of their stakeholders.

One issued to be addressed, which is often discussed by pedagogists and educators, is the source and the legitimisation of the teacher's ability to maintain attention, preserve discipline, and demand honest and respectful behaviour from her students (Johnston 1991). Could technology erode the teacher's perceived authority, if her role relied too much on artificial aids and automated methods that surrogate the presence of the instructor? We don't think this is a realistic possibility, especially if technology is designed and implemented to represent an efficacious illustration and an extension of the teacher's agency, and hence a manifestation of her presence (even when the teacher is physically absent). Rather than being conceived as surrogates of the instructor's authority, technology in the classroom should be seen as a medium through which the instructor can express his/her role as an educator and moral guide. In this sense, robots are not different from other media like speech, written text, or multimedia presentations, although they offer unprecedented opportunities thanks to their distinctively embodied, interactive and autonomous features. In accord with these considerations, we believe that using robots as invigilators during exams does not undermine or replace the authority of the teacher but offers new scenarios to exert it in a considerate and professional way, reducing the operational and administrative burden for the teachers.

5 Conclusion and future work

In this empirical study, we show that a humanoid robot such as Nao has the potential to be incorporated in a classroom to deter students from dishonest or cheating behaviour. However, possibly due to Nao's unimposing appearance and responses, the students were significantly more talkative (and hence undisciplined) when the robot was their invigilator. Our study offers an original contribution to understand how Educational Robotics intersects Virtue ethics, as it provides researchers in both areas with a stimulus to discuss the implications of utilizing anthropomorphic robots in complex educational settings where students' moral values like honesty are an integral component.

Prior to articulating future directions emanating from our results, we would also like to acknowledge certain limitations of our study. Although we have attained some significant trends, generalisation of our results must be considered in light of our sample size. Furthermore, students in our sample had received no prior exposure to the Nao robot, which may have contributed towards their "overall" excitement when completing the test with Nao as an invigilator. Due to time limitations and other commitments of the

students, we did not conduct qualitative interviews or discussions with the children at the end of the sessions which could have revealed the possible causes behind the children being over-excited and their sentiments regarding the Nao robot.

In the future, new studies will be necessary to test our speculations and perfect the design of robot invigilators. Despite our efforts, we acknowledge that our experimental setting does not fully represent an obvious classroom or a typical invigilation setting. In addition, the experimental setting should be designed to consider the contextual motivation and the psychology of the students and what they know about the environment in which their behaviour is tested. Future studies could also involve pre-selected groups of students with a demonstrated tendency to cheat when undergoing group examination, more challenging tests, higher stakes and more desirable rewards, explicit indications about the functions and the powers of human and robotic invigilators, and familiar scenarios that resemble more closely the typical test environment. Arguably, a key factor is represented by the robot's look and behaviour: a taller, more imposing and active robot, programmed to track gaze behaviour and speech and respond appropriately to them, could contribute to attain significant results.

Acknowledgements The first author would like to acknowledge the academic development program (ADP) at Western Sydney University which provided the time and resources to carry out this research project. The second and third author would like to acknowledge UPAR Grant (31H125) from UAE University. All authors would also like to acknowledge the logistical support of Waleed Riaz towards the running of the empirical study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ahmad MI, Mubin O, Orlando J (2016) Understanding behaviours and roles for social and adaptive robots in education: teacher's perspective. In: Proceedings of the fourth international conference on human agent interaction. pp 297–304
- Ahmad M, Mubin O, Orlando J (2017) A systematic review of adaptivity in human-robot interaction. *Multimodal Technol Interact* 1:14
- Aroyo AM et al. (2018) Will people morally crack under the authority of a famous wicked robot? In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp 35–42
- Bae I-H, Han J-H (2017) Analysis on teacher's height and authority in robot-assisted learning. *J Digit Content Soc* 18:1501–1507
- Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. *Sci Robotics* 3:5954
- Bushweller K (1999) Student cheating: a morality moratorium? *The education digest* 65:4
- Check Inter-Rater Reliability (2019). <http://datavyu.org/user-guide/best-practices/checks/reliability.html>

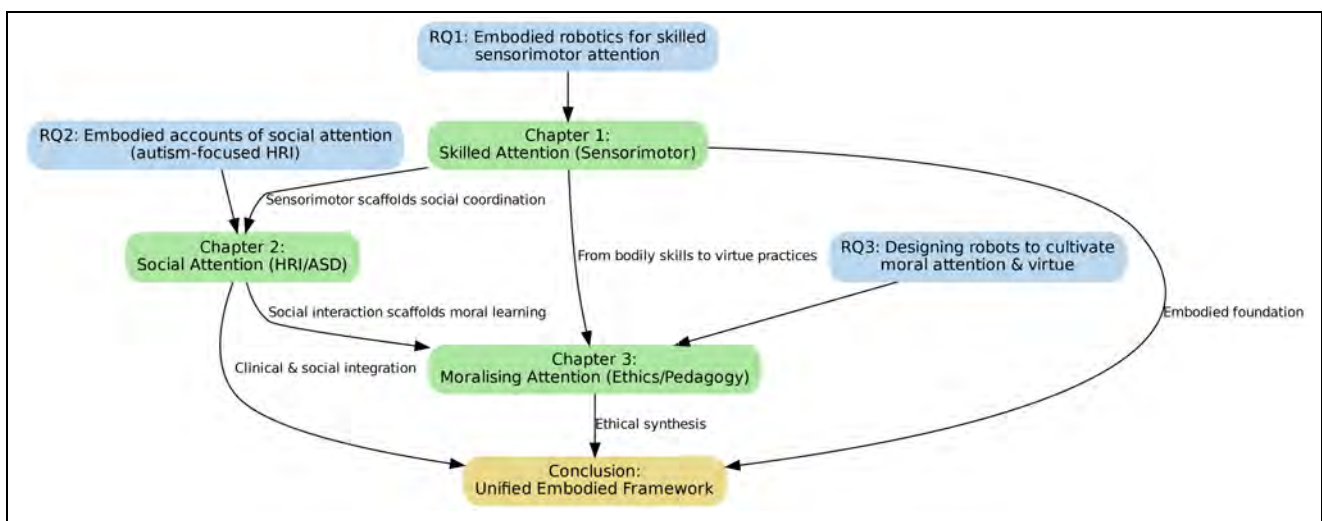
- Conti D, Carla C, Di Nuovo S et al (2019) “Robot, tell me a tale!”: a social robot as tool for teachers in kindergarten. *Interact Stud* 20:1–16
- Covey MK, Saladin S, Killen PJ (1989) Self-monitoring, surveillance, and incentive effects on cheating. *J Soc Psychol* 129:673–679
- D’Mello S, Olney A, Williams C, Hays P (2012) Gaze tutor: a gaze-reactive intelligent tutoring system. *Int J Hum Comput Stud* 70:377–398
- Drake CA (1941) Why students cheat. *J High Educ* 12:418–420
- Edwards BI, Cheok AD (2018) Why not robot teachers: artificial intelligence for addressing teacher shortage. *Appl Artif Intell* 32:345–360
- Edwards A, Edwards C, Spence PR, Harris C, Gambino A (2016) Robots in the classroom: differences in students’ perceptions of credibility and learning between “teacher as robot” and “robot as teacher”. *Comput Hum Behav* 65:627–634
- Forlizzi J, Saensuksopa T, Salaets N, Shomin M, Mericli T, Hoffman G (2016) Let’s be honest: a controlled field study of ethical behavior in the presence of a robot. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp 769–774
- Geiskkovitch DY, Cormier D, Seo SH, Young JE (2016) Please continue, we need more data: an exploration of obedience to robots. *J Hum Robot Interact* 5:82–99
- Gonzalez Billandon J et al (2019) Can a robot catch you lying? A machine learning system to detect lies during interactions. *Front Robotics AI* 6:64
- Gouaillier D et al. (2009) Mechatronic design of NAO humanoid. In: Robotics and Automation, 2009. ICRA’09. IEEE International Conference on, 2009. pp 769–774
- Hoffman G et al. Robot presence and human honesty: Experimental evidence. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, 2015. pp 181–188
- IELTS Academic Reading test – paper (2019). <https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-practice-tests/reading/academic>
- Itatu J (2015) Effects of indiscipline on academic performance of secondary schools in Mutito Sub-county, Kitui county. Technical report <http://197.243.10.178/handle/123456789/3032>
- Johnston DK (1991) Cheating: reflections on a moral dilemma. *J Moral Educ* 20:283–291
- Kelly-Hedrick M, Grunberg PH, Brochu F, Zelkowitz P (2018) Its totally okay to be sad, but never lose hope: content analysis of infertility-related videos on YouTube in relation to viewer preferences. *J Med Internet Res* 20:e10199
- Kennedy J, Lemaignan S, Belpaeme T (2016) The cautious attitude of teachers towards social robots in schools. In: Robots 4 Learning Workshop at IEEE RO-MAN.
- Klincewicz M (2019) Robotic nudges for moral improvement through stoic practice. *Techné* 23:425–455
- Kory Westlund J et al. (2016) Lessons from teachers on performing HRI studies with young children in schools. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp 383–390
- Lee M-H, Tsai C-C (2010) Exploring teachers’ perceived self efficacy and technological pedagogical content knowledge with respect to educational use of the World Wide Web. *Instr Sci* 38:1–21
- Lee E, Lee Y, Kye B, Ko B (2008) Elementary and middle school teachers’, students’ and parents’ perception of robot-aided education in Korea. *EdMedia + Innovate Learning*. AACE, Morgantown, pp 175–183
- Lewis R, Romi S, Qui X, Katz YJ (2005) Teachers’ classroom discipline and student misbehavior in Australia China and Israel. *Teach Teach Educ* 21:729–741
- Li J, Kizilcec R, Bailenson J, Ju W (2016) Social robots and virtual agents as lecturers for video instruction. *Comput Hum Behav* 55:1222–1230
- McCabe D (2005) Cheating: Why students do it and how we can help them stop. In: Guiding students from cheating and plagiarism to honesty and integrity: Strategies for change, pp 237–242
- Miller AD, Ramirez EM, Murdock TB (2017) The influence of teachers’ self-efficacy on perceptions: perceived teacher competence and respect and student effort and achievement. *Teach Teach Educ* 64:260–269
- Mubin O, Stevens CJ, Shahid S, Al Mahmud A, Dong J-J (2013) A review of the applicability of robots in education. *J Technol Educ Learn* 1:209–0015
- Muchai J (2018) An investigation into factors that contribute to cheating in examinations in technical institutions in central province. Machakos University, Kenya
- Rahim H, Rahiem MDH (2012) The use of stories as moral education for young children. *Int J Soc Sci Humanit* 2:454
- Reich-Stiebert N, Eyssel F (2015) Learning with educational companion robots? Toward attitudes on education robots, predictors of attitudes, and application potentials for education robots. *Int J Soc Robotics* 7:875–888
- Roizman M, Hoffman G, Ayal S, Hochman G, Tagar MR, Maaravi Y (2016) Studying the opposing effects of robot presence on human corruption. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp 501–502
- Sandoval EB, Brandstetter J, Bartneck C (2016) Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp 117–124
- Serholt S et al (2014) Teachers’ views on the use of empathic robotic tutors in the classroom. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp 955–960
- Serholt S, Barendregt W, Vasalou A, Alves-Oliveira P, Jones A, Petisca S, Paiva A (2017) The case of classroom robots: teachers deliberations on the ethical tensions. *Ai Soc* 32:613–631
- Shu LL, Gino F, Bazerman MH (2011) Dishonest deed, clear conscience: when cheating leads to moral disengagement and motivated forgetting. *Personal Social Psychol Bull* 37:330–349
- Softbank Robotics (2017). <https://www.ald.softbankrobotics.com/en/robots/nao>
- Williams R, Machado CV, Druga S, Breazeal C, Maes P (2018) My doll says it’s ok: a study of children’s conformity to a talking doll. In: Proceedings of the 17th ACM Conference on Interaction Design and Children. pp 625–631

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4. Conclusion

4.1 Answering the research questions

As illustrated by the diagram below, the previous three chapters have examined the role played by attentional behaviour in HRI, highlighting different ways for attention to shape the cognitive function of humans and robots, considered independently (acting autonomously), or during their reciprocal engagement (acting together and in team): Chapter 1 has considered the role of expert attention (the attention of an experienced performer), in skilled sensorimotor activities, such as golf putting or other sports tasks; Chapter 2 has considered the role of social attention (the attention of affective interactants), in joint and coordinated tasks; Chapter 3 has considered the role of moralizing attention (the attention of a superintendent or coach) in edifying and disciplining activities.



In the first case, attentional resources were spent by the agent to collect task-relevant information and for self-regulation to control goal-oriented tasks successfully (as theorized by Baumeister & Heatherton 1996): my Active Inference model has evidenced that the critical aspect was the focus of attention, and whether it was self-directed and reflective (self-monitoring via internal focus) or outcome-driven and pre-reflective (effect-anticipating via external focus). I concluded that the pre-reflective form of attention, the one that controls action by predicting its intended outcome, is the form of attention deployed by expert agents to perform skilfully. In the second case, attentional resources were gathered from the interactants to establish playful engagement and to maintain a curious and enthusiastic attitude in them: my empirical study with humanoid robots and autistic children has

evidenced that a critical aspect was the affective quality and the duration of attention, which in turn depended on whether attention was solicited through repetitive routines or rather by adaptive tasks. I concluded that the attention stimulated by varying activities, responding to the changing emotional state of the user, is the form of attention that a companion or mate needs to solicit to establish an enduring and truly engaging social relationship. In the third case, attentional resources were exchanged between robots and the human bystanders to solicit self-restraint and moral awareness in the latter: my model of human attitudes and then my empirical study with robot invigilators has evidenced that the critical aspects were the agent's presence and their credibility, and whether the agent looked sufficiently authoritative and physically imposing to command discipline and motivate compliance. I concluded that gaze can be leveraged to produce morally significant effect, exerting an edifying function, only if it operates in the right physical circumstances. The following table summarises the key findings of each chapter and how they answer the four research questions.

Fig. 40 – Addressing the Research Questions

Research Question	Addressed in Chapter	Key Findings	Contributions to HRI & Social Robotics
RQ1: <i>How can embodied approaches to sensorimotor attention enhance the design of robotic systems capable of skilled interaction in dynamic environments?</i>	Chapter 1 – Skilled Attention (Sensorimotor Domain)	Non-representational, adaptive coupling approaches enable robots to manage complex tasks without heavy reliance on internal world models. A comparative analysis of representational vs. non-representational robotics highlighted efficiency and adaptability.	Offers a framework for developing robots that can better adjust to unpredictable environments; advances the debate on embodied cognition vs. symbolic AI in robotics.
RQ2: <i>In what ways does social attention, particularly in the context of autism, illuminate the challenges and opportunities of embodied accounts of cognition?</i>	Chapter 2 – Social Attention (Clinical & Social Domain)	Reviewed theoretical models of autism (ToM, Weak Central Coherence, Executive Dysfunction, Broken Mirror, Sensory Overload) and reframed them under an embodied cognition lens. Showed how social robots can model or remediate atypical attention patterns.	Provides an integrated embodied framework (EAA) for interpreting ASD; bridges cognitive science with therapeutic HRI applications.
RQ3: <i>How can social robots be used to cultivate virtuous behaviour and moral attention in educational and developmental contexts?</i>	Chapter 3 – Moralizing Attention (Ethical & Pedagogical Domain)	Explored “virtuous robotics” and the potential for anthropomorphic design to scaffold character development. Identified both benefits (engagement, habit formation) and risks (dependency, deception, overgeneralisation).	Introduces the notion of robotic systems as pedagogical and moral partners; expands the ethical scope of HRI from “do no harm” to “promote the good.”

Research Question	Addressed in Chapter	Key Findings	Contributions to HRI & Social Robotics
RQ4: <i>What unifying principles can connect the domains of skilled, social, and moral attention within an embodied framework for HRI?</i>	Conclusion – Integrative Synthesis	Showed that sensorimotor, social, and moral dimensions of attention are not isolated but mutually reinforcing. Embodied coupling provides the common mechanism: motor learning supports social coordination, which in turn scaffolds moral development.	Proposes a holistic model of attention in HRI; establishes embodied robotics as a unifying paradigm across technical, therapeutic, and ethical domains.

4.2 Original contributions to the field

Despite the great diversity of these applications of attention, the communal element is the quality of the agents' embodiment and the physical details of their interactive behaviour. This conclusion confirms the hypothesis, motivated by ECT, that attention is not just an abstract cognitive function, a detached and vaporous mental state that operates in a non-physical world and that has only a representational, spectatorial function. Attention is embodied because it is a key component of the interaction between agents, and such interaction is embodied from the beginning to the end. Therefore, all considerations concerning the social, moral or epistemic value of attention, as well as its duration, quality, and span, must carefully take into account how attention is concretely realized, spatio-temporally situated, dynamically articulated, and physically instantiated in the context of rich reciprocal interactions among embodied agents.

This picture confirms that the two radical tenets of ECT are accurate and indicates how these principles can be serviceable to design new robots (more attuned to the circumstances of real world), and better interaction protocols (leveraging the physical appearance, the presence, and the behaviours of the interactants). That cognition is embodied means that these characteristics are constitutive and active elements of an embodied, situated, and distributed cognitive system, as opposed to merely exterior, contingent, and instrumental aspects of cognition. The first radical tenet, i.e. the idea that cognition is non-representational, is confirmed by the fact that, in every scenario we have considered, attention was not meant to reflect a 1:1 correspondence between a hypothetical external world and the internal replica constructed by the agent's mind; on the contrary, attention was physically enacted as an action, as an intention to affect and manipulate. It is something the agent does, more than something happening to the agent. The second radical tenet, i.e. the idea that cognition is distributed and encompasses material processes and peripheric systems beyond the core computational vehicles (brain, CPU), possibly even beyond the

canonical boundaries of the agent's body, is confirmed when we think that, in every scenario considered so far, attention was a channel to connect different bodies and to bridge the material gap between different perceptual and motoric and inferential systems to allow their real-time, fine-grained, goal-directed intelligent coordination.

Now, after having reviewed my previous research projects, I am in a better position to both defend and clarify the assertion that *attention is embodied*. I have gathered significant data and ideas concerning how the details of agent's physical presence and behaviour are not just instrumental to realize attention, but they also constitute, actively shape, and elevate the function of their attention. The analyses of my previous work allow me to distinguish at least four possible ramifications of this claim.

First, in the past, attention was often construed as an intellectual, disembodied faculty – more of a detached form of spiritual self-control than an intelligent regulation system realized by the agent's body. Today, however, cognitive scientists are much keener on accepting that attention has an important, constitutive, embodied dynamic. So much so that, without a body, attention is not a property of a real agent. I have described how a robot can invite, maintain, or re-direct the human user's attention using its own gaze, gestures, posture, or other features of its bodily endowment and behaviour. Secondly, attention is crucial in the management of cognitive faculties, e.g. in the dynamic allotment of the resources necessary to control multiple concurrent tasks or in the capacity to concentrate such resources towards a specific focus. Importantly, this doesn't mean only avoiding distraction by means of sustained attention, as distraction is not the only peril that risks jeopardizing expert activities: it means also avoiding reflective self-focus and explicit, analytic control, which compulsively hinders the natural control of fluid action. Thirdly, attention plays a crucial role in social interaction and can become the basis for pro-social and collaborative behaviours, including complex forms of communication: attention can be given to others, gently requested from them, or even forcefully taken; it can be self-centred or re-oriented towards an external focus; it can be overt, to be exerted explicitly and assertively, or concealed, to be exerted covertly. Fourth, attention is an instrument of objectification and as such it has the power to make us feel judged and scrutinized: this has an important function because it reminds us of our moral responsibilities, making us aware of our commitment to self-improvement, maturation and growth. In summary, a robot's embodiment and its effects on attentional resources are intertwined and play an important role in HRI insofar as physical presence and body design enhance the robot's ability to manage attention effectively, in a way that deeply influences the quality of interaction and human-robot teaming (joint tasks).

4.3 Policy Implications for HRI in Education and Therapy

The findings and arguments developed in this dissertation also carry important policy implications, concerning the use of social robots in education and therapy. Robots are increasingly introduced into classrooms and clinics, asking policymakers and practitioners to recognize the impact of technology beyond instrumental assistance, shape moral, aesthetic, and epistemic values. Regulatory frameworks and professional guidelines should therefore account not only for issues of safety, privacy, and accessibility, but also for the ethical and pedagogical dimensions of HRI. In education, this entails policies that promote responsible integration of robots as complementary tools for cultivating discipline, collaboration, and independent thinking, without displacing human teachers or undermining social bonds among peers. In therapy, it requires careful oversight to ensure that robotic interventions support, rather than substitute, the irreplaceable role of human caregivers and therapists, while also safeguarding against risks of dependency, deception, or inequitable access. At a broader level, policies should encourage interdisciplinary research and cross-sector dialogue, so that innovations in embodied HRI are evaluated not only as technological solutions but also as transformative forces in the socio-political and cultural sphere. To translate these reflections into practical guidelines, a set of four actionable recommendations can be outlined.

(i) *Establish Ethical Guidelines for Educational and Therapeutic Robots* – Regulatory bodies should define standards to ensure that robots are deployed to empower teachers and therapists, not to replace them or constrain their work. Ethical guidelines should promote transparency (users must know they are interacting with a machine), fairness (equal access across socioeconomic contexts), and accountability (clear lines of responsibility for desirable outcomes). Importantly, these guidelines should promote human-Centered design in policy frameworks: funding and approval mechanisms should privilege robot designs that enhance human agency, social interaction, and developmental growth. Anthropomorphic features or persuasive techniques should only be encouraged when they demonstrably promote learning or therapeutic outcomes without fostering dependency or deception. (ii) *Mandate Interdisciplinary Evaluation Before Deployment* – Policies should require pilot studies with multidisciplinary oversight—including educators, therapists, ethicists, and engineers—before robots are widely deployed in schools or clinical settings. Evaluation should include both technical performance and psychosocial impact (e.g., attention regulation, empathy, character development). One aspect of technology adoption to be carefully assessed is the potential for over-reliance and dependency: Educational and therapeutic institutions should be required to evaluate regularly whether robotic interventions are fostering autonomous

skill development in children or patients, rather than creating reliance on machine-mediated interaction. (iii) *Encourage Inclusive and Equitable Access* – Governments and funding agencies should support initiatives that ensure robotic interventions are accessible to underserved schools and clinics, preventing technology-driven inequalities in education and healthcare. This means also offering pathways and incentives for continuous training and support for practitioners: teachers and therapists using robots should be able to benefit from training in both technical operation and ethical considerations, supported by ongoing professional development programs. (iv) *Invest in Longitudinal Research on Developmental Impact* – Policymakers should allocate funds for long-term studies that assess how repeated exposure to social robots influences attention, learning outcomes, and moral development, thus ensuring policies are evidence-based.

4.4 The way forward: limitations and future directions

While this dissertation has outlined a comprehensive framework for understanding and designing embodied approaches to attention in HRI, it obviously suffers from inevitable limitations too. The empirical and theoretical work presented here is necessarily limited because selected to reflect and explain the path followed by its author during his career. Also, the case studies of skilled, social, and moralizing attention were chosen to emphasize domains of application that have been particularly successful during the last 20 years, but they cannot exhaust the diversity of possible scenarios in which embodied robotics may play a transformative role in the future, especially considering the unpredictable and groundbreaking impact of emerging technologies like multi-modal large language models, autonomous navigation, and general-purpose humanoid robots. Methodologically, the studies relied on experimental designs and interpretive frameworks that, while rigorous, may not fully capture the complexity of long-term human–robot encounters in naturalistic environments. In addition, the philosophical defense of non-representational and embodied accounts of cognition, while consistent with current findings in cognitive science, must be further articulated to take into account the novel challenges posed by hybrid architectures: for examples, the rapid developments in machine learning and generative AI, which combine representational and sensorimotor strategies in ways that seem destined to blur traditional distinctions between contentful and contentless models. Future work should therefore expand empirical testing to more diverse populations and ecological contexts, explore hybrid approaches that integrate embodied interaction with symbolic and principled adaptation, with a view to refine the conceptual apparatus for articulating the ethical and pedagogical dimensions of attention in HRI.

Despite these limitations, the research documented in this thesis demonstrates that the embodied approach to attention is promising and viable as both a scientific paradigm and a practical orientation for robotics. By grounding interaction in the bodily and social dynamics of attention, robots can serve not only as tools or assistants, but also as catalysts for human development. As companions, social robots display a significant potential to enhance skilled performance, scaffold social understanding, and nurture moral character. Such a perspective reframes HRI as a shared domain of growth, in which the value of robotic companions is measured less by their capacity to mimic human intelligence than by their ability to participate in and sustain the socio-cultural processes through which human beings become more attentive, virtuous, and socially connected. The embodied approach defended in this thesis thus offers a forward-looking, ecologically aware, and socio-culturally sustainable pathway for humanoid robotics. One that does not merely seek efficiency or automation but also aspires to enrich the quality of human life.

APPENDIX

Support Publication #1 **[SP1]**

Cappuccio M.L. (2009),
“Constructing the Space of Action. From Bio-Robotics to Mirror Neurons”
World Futures. The Journal of General Evolution, 65/2 (February 2009),
Taylor & Francis, London, pp. 126-132.

Reproduced with permission from Taylor & Francis. All rights reserved.

CONSTRUCTING THE SPACE OF ACTION: FROM BIO-ROBOTICS TO MIRROR NEURONS

MASSIMILIANO CAPPUCCIO
CREA-Ecole Polytechnique, Paris, France

This article distinguishes three archetypal ways of articulating spatial cognition: (1) via metric representation of objective geometry, (2) via somatosensory constitution of the peripersonal environment, and (3) via pragmatic comprehension of the finalistic sense of action. The last one is documented by neuroscientific studies concerning mirror neurons. Bio-robotic experiments implementing mirror functions confirm the constitutive role of goal-oriented actions in spatial processes.

KEYWORDS: *Action, AI, bio-robotics, canonical neurons, cog, embodiment, enaction, kinesthesia, mirror neurons, premotor cortex, space.*

The limitations of today's bio-robotics suggest ways of addressing the philosophical problem of the living body at least from the point of view of objective space constitution. This article distinguishes three archetypal ways of articulating our spatial experience.¹ The distinction aims to clarify that the complexity of our cognition of space is not reducible, *more geometrico*, to the mere quantification of its metric relationships.

WAY A: METRIC REPRESENTATION OF THE NATURAL-OBJECTIVE SPACE

Space is codified by the simplest artificial intelligence (AI) through syntactically defined logical-symbolic operations. This spatiality type consists of neutral and homogeneous parameters, as the quantitative representation merely grasps its exact metrics relationships, like in Euclidean geometry. During the execution of an action, the perceptive data (input) and the motor instructions (output) represent two opposite ends of a hierarchized procedure of algorithmic elaboration.² Action and perception are therefore correlated, as long as they mutually update the base of their own data. However, they remain two separate realities, substantially not unifiable, for their function and position inside the computational task is clearly distinguished. In the formalistic framework of classical computationalism, the

Address correspondence to Massimiliano Cappuccio, via Tolmezzo 12/7, 20132 Milano, Italy. E-mail: massimilianocappuccio@hotmail.com

space objectively elaborated as a metric representation is a virtual place, a space-without-body unfamiliar to the lived space of the embodied existence, but only concerned with its formal extrinsic description.

WAY B: ENACTIVE CONSTITUTION OF THE PERIPERSONAL BODILY SPACE

Referencing Varela, Thompson, and Rosch (1991), space—at this level—is “enacted” through *kinesthetic operations*, that is, perceptive acts endowed with an active and constitutive value. Kinesthetic space could be accessed by robots endowed with an enactive sensorimotor control. Perceptive information and motor programs are polyvalent complexes of competences embodied in neuronal circuitries. Perception and action do not exist as two opposite heads of a rationalistic formalism or as a one-way computational function; on the contrary, they emerge at each instant as moments of the same distributed process. Cognitive neurosciences suggest that perception and action constitute complex polyvalent functions emerging from the activity of a single integrated system or an encapsulated cognitive module. The studies by Alain Berthoz and his school have shown, for instance, the *intelligently predictive* active role of ocular saccades in vision stabilization processes. Some neuronal mechanisms of the vestibular system, elaborating efferent copies from visual information, allow the automatic reconfiguration of body scheme without the intervention of any objective space-centralized representation mechanism. In this case “a motor signal commanding the direction of the eye modifies the sensorial neuron. The action influences perception at its source!” Berthoz (1997, 78).³ Ascertaining the relationship of close co-implication subsisting between movement and vision, Berthoz states that “it is necessary to suppress the dissociation between perception and action” (15), considering them as two inextricably joined events, sprung by their common multimodal and embodied deep nature.⁴

Space unravels itself in front of the body as an enactive scenery for the execution of sensorimotor operations: bodily position in B-Way-Space is not exclusively characterized in terms of objective coordinates or metric relationships among objects. To the contrary it is also defined as a greater or smaller perspectual prominence assumed by objects in relationship to the situated body of an enactive subject. It is a space endowed with an egocentrated grading of the sense of distance characterising what is “here” and what is “there” (Husserl 1973; Petit 1997, 1999). This can be simultaneously reconstructed through multiple allocentrated references, reciprocally reversible and seeking a concordant synthesis, as shown by Alain Berthoz’s studies on the sense of movement (Berthoz 1997), and by their following phenomenological setup (Berthoz and Petit 2006). No longer, then, an homogeneous and neutral space, but a qualitatively characterized one, intrinsically related to the presence of his protagonist-subject and of his manifold potential alter-egos (Berthoz 2004, 268–272; Berthoz and Petit 2006, 131–139, 275–281).

What circumscribes this second description of spatial experience is that, from its perspective, the solely possible performative task is still only a kinaesthetic

action, that is, a sensorimotor variation redefining the correlation among geometric variables (see Clark 1997, 722–723).

WAY C: OPERATIONAL COMPREHENSION OF THE PRAGMATIC SPACE OF ACTION

A further layer of the enactive constitution of the *living-body*, not contrasting but complementing the function of kinaesthesia, needs to be considered. At this third level, the space is frequented as operational dimension of the actions executable by a living body, as a function of their pragmatic and teleological meaning.

If we assume the perspective of Heidegger's phenomenology, we can state that this spatial experience is unfolded as horizon of the "care," the condition characterizing the situated being of the living things: it connotes the spatial objects as virtual poles of possible actions, or rather as objects *ready-to-hand* available for an undetermined number of practical uses.⁵ The purpose of the action is disclosed as a primitive relationship of sensible familiarity toward the world, and not as a planned strategy of intervention (Gallese and Metzinger 2003).

The experiments conducted by Giacomo Rizzolatti's parmesan group show the role of premotor cortex in the elaboration of this class of pragmatical competences: for instance, the *mirror neurons* are selectively activated by specific typologies of finalized actions (Fogassi et al. 1998). Mirror neurons' activation does not depend on the identity of the subject performing the action, nor on the type of effector involved, nor on the wish or the desire of accomplishing the action, nor does it depend on the configuration of the assumed motor scheme (the topology of the movements of arm, hand, fingers, etc.). Therefore, it would be inaccurate to state that mirror neurons codify the spatial syntax of the action.⁶ More precisely, their activation specifically corresponds to the access to the "intentional meaning" of the gesture, or rather its pragmatic value, portraying it since its commencement as a finalized action endowed with a practical and transformative meaning (Rizzolatti and Sinigaglia 2007).

Besides motor properties, the so-called *canonical neurons* also possess visual ones. Like mirror neurons, they are involved in modeling specific varieties of actions. They are also activated by the sight of an object provided with salient characteristics evoking the execution of that same action. Unlike mirror neurons, canonical neurons discharge in front of the objectual correlate of a possible action, and not in front of the sight of the performed action itself. Canonical neurons attest, therefore, the presence of pre-conscious processes incorporating a pre-understanding of spatial objects in terms of practical usability.

The pragmatic role of these neuronal populations is experimentally confirmed by means of subtractive methods. Gallese et al. (1994), and then Fogassi et al. (2001), have shown that if muscimol microinjection temporarily inhibits some groups of canonical neurons involved in the modeling of precision prehension, in monkey's F5 zone, the execution of the prehension appears seriously uncoordinated, as if the monkey had been deprived of the general understanding of the operational sense of what it intended to do. The monkey does not succeed in pre-shaping, through simple observation of the object, the necessary movements to

perform the finalized action, which must now be performed in a hard-working fashion; the monkey needs repeated gestures of tactile exploration to reconstruct the correct motor scheme and hence conclude the precision prehension. The chemical intervention did not jeopardize the ability to reconstruct the gesture morphology; it has only impeded its holistic prefiguration, which would consent a harmonious performance endowed with practical sense. Muscimol did not inhibit motor coordinates calculus related to postural planning and to trajectory in the objective space, as these abilities efficiently continue in other cortical areas. The geometric representation of space and the horizon of endless virtual profiles of the objects are still available to the monkey. What fails to function is the understanding of the space in terms of possible *affordances* (Gibson 1979), which would allow the modeling and leading of the embodied actions on the basis of a purpose.⁷

Some characteristics of canonical neurons and of mirror neurons can be roughly reproduced in a machine. Recent bio-robotics experiments have pursued an approximation to their functions: for instance, the study by Giorgio Metta and others, documenting the functioning of an anthropomorphic robot and the pattern of its progressive learning from the visuo-motor and tactile experience.⁸

This model implemented a system of Bayesian nets set up for action *recognition* simulation within an artificial context. This system has been used as the computational heart of an anthropomorphic cybernetic device, called Cog, implementing a system of limbs so sophisticated that it can potentially execute a few manipulative gestures among those available to man.⁹

This experiment aims to provide the robot with essential cognitive equipment in order for it to succeed in autonomously modelling its own visuo-motor experience, progressively learning the causal structure that defines the rules of acting in space. The robot is therefore driven to acquire consecutive levels of motor competences, by using a statistical error proof method. This learning faculty must progress through different degrees of constitution of the practical experience.

First level. The robot learns fundamental kinematics laws allowing it to control its own limbs, to move them in the direction of the objects and to reach them. The robot can deal with elementary regularities in the process of operational constitution of the spatiality.

Second level. The robot learns some *affordances* among the ones encoded by canonical neurons. Not being endowed with a true hand of its own, the robot can only learn a few possible elementary operations on the objects: to touch, to strike, to push away, and to sideways sweep. Cog picks up the topological data inherent to the movement of its body while performing the explorative actions: the robot is left for a few days to feel the most disparate objects using the available simple motor approaches. The harvest of tactile data provides information inherent to the type of spatial answer passively offered by the object's morphology, by establishing a net of correlations between geometric forms of solids and action typologies corresponding to their haptic exploration in the space. The visual data picked up and elaborated by the optic apparatus of the robot provides the feedback for correcting the exploratory operations and for recording the geometrical changes occurred in the shape of the object during the variations of the point of observation. The cognitive procedure followed by the robot still shapes the space

in kinesthetic terms, because it proceeds by putting together somatosensorial data and dynamically adjoined proprioceptual information (Metta et al. 2006, 212). On this second level, therefore, the characteristics of the objects must be inferred from the type of actions executable on the object itself (Metta et al. 2006, 220).

Third level. The third level of motor competences learned by Cog attempts to emulate the visuo-motor function of the mirror neurons by following an inverse run: the category of a certain performed action must be recognized by deducting it from the topological characteristics of the manipulated objects, learning them through the manipulative familiarity acquired during the preceding phase.

This third level of competence seems to approximate the cognitive functions of *action recognition* and of imitation endorsed by mirror neurons (Rizzolatti, Fogassi, and Gallese 2001). The robot is predisposed for analyzing the visual data coming from the observation of a human agent accomplishing a determined type of action on an object. The action accomplished by the human being belongs to one of the classes of actions fitting in the pragmatic repertoire previously acquired by Cog. The robot is programmed to discriminate between different action types coupled with the previous experiences of motor exploration. This experiment was intended to empower the robot to recognize the action visuo-motor pattern type by starting from its practical “purpose.” *Practical purpose*, in this artificial context, refers only to the complex of kinematic consequences achieved by the action on the object and on the effectors involved.

Once the robot has acquired this ability, it can easily be programmed to imitate—by referring to its “goal”—the last action that a human being has completed in front of it. What Cog reproduces, therefore, are not the single components of movements, but the type of kinematic effects of the action.

Even if it cannot capture in its entirety the complexity of the finalism implicit in the simplest animal action, the study of the imitative abilities of Cog is extraordinarily clarifying, as it shows that: (1) the ability of imitating an action is strongly dependent on the understanding of its holistic value and of its finalistic sense; (2) the understanding of the finalistic sense of action is strongly facilitated by the performative experience previously acquired; (3) this performative experience is managed by a unique structure controlling both perceptive and motor functions; and (4) the enactive cognitive strategy based on the embodied simulation¹⁰ of the practical action is more economic and efficient in comparison to a centralized representation of the geometric components of gesture, as shown in studies on living systems (Woodward 1998).

NOTES

1. This idea originates from a personal interpretation of Corrado Sinigaglia’s teaching. I have to thank Vittorio Gallese for indispensable suggestions. Responsibility for eventual mistakes is just mine.
2. This concept is effectively developed by Gallese (2006).
3. Clark (1997, 29–31) discusses the *embodied* function of saccadis in terms of efficiency of the visual trials and computational economy, through models that are set aside from the use of detailed representations of the external world.

4. As will be better explained later, the study of premotor cortex visual properties, by Rizzolatti, Fogassi, and Gallese (2001) offers a further occasion to reject the cognitivist scheme opposing perception and action.
5. Paragraph 22 of *Being and Time* is devoted to “The Spatiality of the Ready-To-Hand Within-The-World.” Cf. Heidegger (1962, 134–137): here the crucial concept of “readiness-to-hand” (*Zuhandenheit*) is identified with the degree of “closeness” that qualifies the situation of “everyday dealings” entertained with the Within-The-World beings. Some studies recognize the constitutive role, for the manipulation of space, of this characteristic proximity to the living body. For instance, by approaching philosophically the experimental data acquired in Berti and Frassinetti (2000), a neurophenomenological theory of *Zuhandenheit* could be perhaps developed.
6. Merleau-Ponty (2003, 199–202) did clearly illustrate that syntax is not a primary feature in the motor construction of the space.
7. Binkofski et al. (1999) offer confirmations within a frame of clinical neuropsychology.
8. Metta et al. (2006). Similar researches appeared already in Tani, Ito, and Sagita (2004).
9. Some of the first descriptions of Cog can be found in Brooks and Stein (1993).
10. Cf. Gallese (2006), about embodied simulation and structuring of the practical space experience.

REFERENCES

- Berti, A. and F. Frassinetti. 2000. When far becomes near: Re-mapping of space by tool use. *Journal of Cognitive Neuroscience* 12: 415–420.
- Brooks, R. A. and L. A. Stein. 1993. Building brains for bodies. Memo 1439. Cambridge, MA: Massachusetts Institute of Technology Artificial Intelligence Lab.
- Berthoz, A. 1997. *Le sens du mouvement*. Paris: Odile Jacob.
- . 2004. Physiologie du changement de point de vue. In *L'Empathie*, ed. A. Berthoz and G. Jorland, 251–276. Paris: Odile Jacob.
- Berthoz, A. and J.-L. Petit. 2006. *Physiologie de l'action et phénoménologie*. Paris: Odile Jacob.
- Binkofski, F., G. Buccino, S. Posse, R. J. Seitz, G. Rizzolatti, and H. J. Freund. 1999. A fronto-parietal circuit for object manipulation in man: Evidence from an fMRI study. *European Journal of Neuroscience* 11: 3276–3286.
- Clark, A. 1997. *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.
- Fogassi, L., V. Gallese, G. Buccino, L. Craighero, L. Fadiga, and G. Rizzolatti. 2001. Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain* 124(3): 571–586.
- Fogassi, L., V. Gallese, L. Fadiga, and G. Rizzolatti. 1998. Neurons responding to the sight of goal directed hand/arm actions in the parietal area PF (7b) of the macaque monkey. *Society of Neuroscience Abstracts* 24: 257–265.
- Gallese, V. 2006. Corpo vivo, simulazione incarnata e intersoggettività. Una prospettiva neuro-fenomenologica. In *Neurofenomenologia. Le scienze della mente e la sfida dell'esperienza cosciente*, 241–270, ed. M. Cappuccio. Milano: Bruno Mondadori.
- Gallese, V., A. Murata, M. Kaseda, N. Niki, and H. Sakata. 1994. Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* 5(12): 1525–1529.
- Gallese, V., and T. Metzinger. 2003. Motor ontology: The representational reality of goals, actions and selves. *Philosophical Psychology* 16(5): 365–388.
- Gibson, J. J. 1979. *The ecological approach to visual perception*. Boston: Houghton Mifflin.
- Heidegger, M. 1962. *Being and time*. Trans. by John Macquarrie and Edward Robinson. London: SCM Press.
- Husserl, E. 1973. *Ding und Raum. Vorlesungen 1907*, ed. U. Claesges, in *Husserliana XVI*. Den Haag, The Netherlands: Nijhoff.

- Metta, G., G. Sandini, L. Natale, L. Craighero, and L. Fadiga. 2006. Understanding mirror neurons. A bio-robotic approach. *Interaction Studies* 7(2): 197–231.
- Merleau-Ponty, M. 1962. *Phenomenology of perception*. Trans. C. Smith. London: Routledge & Kegan Paul.
- Petit, J.-L. 1997. *Les neurosciences et la philosophie de l'action*. Paris: Vrin.
- . 1999. Constitution by movement: Husserl in light of recent neurobiological findings. In *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*, ed. J. Petitot, F. J. Varela, B. Pachoud and J.-M. Roy, 220–244. Stanford, CA: Stanford University Press.
- Rizzolatti, G., L. Fogassi, and V. Gallese. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* 2: 661–670.
- Rizzolatti, G. and C. Sinigaglia. 2007. *Mirrors in the brain: How our minds share actions, emotions, and experience*. Oxford, UK: Oxford University Press.
- Tani, J., M. Ito, and Y. Sagita. 2004. Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks* 17: 1273–1289.
- Varela, F., E. Thompson, and E. Rosch. 1991. *The embodied mind*. Cambridge, MA: MIT Press.
- Woodward, A. L. 1998. Infants selectively encode the goal object of an actor's reach. *Cognition* 69: 1–34.

Support Publication #2 **[SP2]**

Cappuccio M.L., Ilundain-Agurruza J. (2019),
“Embodied Cognition”

in Dieter Hackfort and Robert Schinke (eds.), *The Routledge International Encyclopedia of Sport and Exercise Psychology*, Routledge, 2019.

Reproduced with permission from Routledge (Taylor & Francis Group).

All rights reserved.

9

EMBODIED COGNITION

Massimiliano L. Cappuccio and Jesús Ilundáin-Agurruza

Introduction

Embodied or 4E cognition (EC) approaches emphasize the defining and constitutive role that the body plays in mental activities. Sport psychologists are inclined to embrace EC because it appreciates how a full range of human experience that includes the affective, perceptual, and motoric, not just the intellectual, constitutes legitimate cognitive activity. EC stresses that sensorimotor skills and bodily affects (denoting moods, emotions, and feelings) enable and scaffold cognitive faculties such as perception, attention, control, communication, and memory, shaping and orienting the development of more intellectual capacities that involve language processing and rational judgement. Given the pervasive intertwining among psychophysical phenomena in sport performances (where does the “mental” game end and the physical one begin, when a bad case of the yips shakes a golfer’s hands?), it is imperative to find philosophical accounts of the mind–body relation that are both non-dualistic and anti-reductionist. EC is strongly positioned to provide this. One reason is that it can meaningfully maintain a fundamental distinction between two forms of intelligence, representational and non-representational, while shedding light on how the former emerges from the latter. The representational forms of intelligence are intellectual, reflective, often linguistically mediated, and abstract – for example, a sportsperson or coach explicitly focuses on and analyses specific movements of the performance; the more basic, non-representational ones pertain to bodily engagements that are pre-reflective – that is, tacit, non-objectifying, and non-observational such that they perform unawares as it were. This chapter aims at illustrating how such understanding is needed to give a firm theoretical foundation to the efforts sport psychologists make to model sport performance and skilful expertise in athletes.

The Theory of Embodiment and the Embodied Approach to Cognition

Cognitive science is the interdisciplinary study of mental functions and intelligent systems at the level of their underlying processes and mechanisms. It aspires to answer fundamental questions about the functioning of human, animal, and artificial minds using a combination of empirical, synthetic, and phenomenological methods that involve both quantitative and qualitative analyses. Cognitive science can provide sport psychology with a theoretical and

methodological foundation based on a concrete evidential and normative background. In order to achieve scientific validity, the protocols of practical intervention adopted by practitioner psychologists need to match the theoretical models, the empirical data, and the integrative methods introduced by cognitive science research.

EC is not the only theoretical approach available in this domain. It is necessary to distinguish between a cognitivist and an embodied approach to cognitive science. The traditional approach to cognitive science is often referred as “cognitivist” because it typically describes mental functions in computational and representational terms, as high-profile, logico-causal mechanisms designed to store and manipulate information encoded in a representational format. This approach tends to assume that the structure of cognitive processes is language-like in nature and endowed with semantic content. It presupposes that cognitive processes are essentially algorithmic and formal in nature, and that the details of their physical realization and contextual implementation cannot be a part of cognitive activity.

The premises of cognitivism are internalist, instructionist, symbolicist, and largely intellectualist, as they reduce even the most basic forms of practical know-how to temporally abstract, purely syntactic, amodal information processing. Cognitivism supposes that these systems operate by means of acontextual models of the world stored in the brain and decoupled from online interaction with real-life experiential contingencies. Burdened by such intellectualist assumptions, cognitivism lacks the conceptual resources required to appreciate that athletic skill (for example, wrestling) is a legitimate form of intelligence, because it involves cognitive faculties that are essentially different from linguistic and representational ones and that, despite this, are not less sophisticated or complex than – say – mathematical reasoning (Matthen, 2014).

EC stances represent the main alternative to the cognitivist approach. Embodied cognitive science studies how the mechanisms and processes underpinning cognition are not only causally integrated with, but also actively constituted by, bodily processes. Within EC, the approaches range from moderate versions that propose a minimal embodiment to more radical ones that reject cognitivist computational and representational views altogether. Far from being incidental contingencies, the details of bodily implementation deeply shape cognitive functions and define their reach. EC promotes a deeper understanding of the body that takes seriously the interdependency of body and mind. It maintains that the brain is not the only part of the body involved in cognitive activity, as mental activity is not reducible to the neurons that implement information-processing in the brain. In turn, the extra-neural body is not just a biological vessel necessary to keep the brain alive. The body is more than a system of sensors and actuators instrumental to cognition, or the location in space where mental processes materially occur.

Indeed, for EC, the body works as an open spatio-temporal horizon, a place of opportunities for action and discovery, a dynamical boundary that defines the self and world in processual, relational, and reciprocal terms – that is, as elements of a precarious stability that emerges from a context of continuous unfinished negotiation between organism and subpersonal environment. The body is the dimension in which the original link between personal and subpersonal components of the mind – that is, one person’s consciousness and its underlying mechanisms and processes – can be clearly observed. EC studies this link using a “neurophenomenological” methodology that combines third-person empirical accounts and first-person phenomenological reports.

The phenomenological doctrines to which EC appeals conceive of the body as the place in which the self and world meet and constitute each other. They further see the body as the precondition under which their originating coupling is established. Thus, the body is a prerequisite to discover and make sense of the world as a meaningful experiential reality. The

construction of the world through bodily perception and interaction is what allows embodied agents – athletes, for instance – to make sense of their environment as an open horizon of action possibilities. Revealing this horizon, the bodily experience actively scaffolds the capability to learn and articulate meaningful actions in the environment.

The EC programme – especially the radical trends that defend anti-computationalist, anti-representationalist, and dynamicist views (e.g., Chemero, 2009; Dreyfus, 2002; Hutto & Myin, 2017) – reflects this non-instrumental conception of the body as a precondition for experience and cognition. In this respect, EC opposes cognitivism and offers an alternative to its internalist and instructionist assumptions about the nature of cognition that relegate the body to a simple instrumental function.

The Key Tenets of EC

The fundamental principle of EC is that the reach and the very nature of mental functions constitutively depend on the material and temporal details of their implementation, including the history of interactions between the embodied agent and its world (Gallagher, 2005, 2011). A simple example is how an athlete's movements and experiences, say those of a mixed martial arts (MMA) fighter, are pervasively informed by the history of her engagements with specific opponents in particular locations and times. This principle has two important implications.

First, many "low-level" regulatory and adaptive processes (i.e., either physiological processes in the non-neural body or physical processes in the extra-bodily environment) actively participate in the most basic forms of cognition (sensorimotor, affective, adaptive), scaffolding their development and defining their scope. Not only are perception and action intrinsically intelligent, but also the hormonal and immune systems (Marin & Kipnis, 2013), autonomic and peripheral responses (Aranyosi, 2013), and highly selective saccadic reflexes (Mann, 2019) can play an intelligent role in certain tasks. For example, an expert MMA fighter's response to her opponent is modulated and highly grained: She readily and rapidly picks up the slight raise and twitch on her opponent's right shoulder as it signals a punch. In the context of how the fight is going, this may be read as potentially damaging (she may be reeling from a previous hit) and lead to a feeling of being intimidated so that a cascade of hormonal and emotional processes condition her (re)action – for example, protecting her ribs and face. Models in embodied robotics validate the hypothesis that the anatomical and postural specificity of the body, with its inherent material properties (hardness, elasticity, etc.), and the intrinsic biomechanical and morphological constraints of the actuators (degrees of freedom, weight distribution, etc.) can play a cognitive role. For example, in the design of passive artificial walkers (Pfeifer, Iida, & Lungarella, 2014), the pendulum-like configuration of one's body parts can contribute to reducing the computational complexity of a task of bipedal motion. To return to the MMA fighter, her training has shaped her body in specific ways that mean she can absorb and adapt to hits that would instantly disable a non-fighter: She readily positions her arms in specific configurations in response to her opponent's moves such that, much as a physical chess match, these may cancel possible lines of attack.

Second, competent embodied dispositions and habitual patterns of perception and action, emerging from dynamic forms of adaptive coupling between embodied agents and their environment, contribute significantly to scaffolding and shaping forms of cognition (inference, language, symbolic representation) that tend to appear "disembodied" – that is, less immediately dependent on the details of their bodily implementation. In truth, the body

selectively modulates the development of “higher” cognitive faculties, such as abstract thought and rational decision, and filters or organizes their propositional contents. That is, although these faculties are not primarily aimed at sensorimotor engagement and physical interaction, they are nonetheless always informed by a primordial background of bodily familiarity and perceptual expertise. Strategic control and decision, perceptual discrimination, linguistic representation, memory, judgement, and creativity are not to be understood just as informational functions centrally controlled by the brain or abstract programmes detached from their material implementers: On the contrary, these cognitive functions can be correctly understood only when the fine-grained specificity of the physical and biological systems that instantiate them, and the real-life circumstances in which they have been developed and trained, are considered.

Bodily affects can dramatically curb perception and objective judgement for different reasons. In some cases, it may be that there are costs to our perceptual judgements, as when fatigue influences climbers’ perception of how far they can reach (Pijpers, Oudejans, & Bakker, 2007). In other cases, we find how hunger or other possible extraneous conditions (mental fatigue, self-regulation resources) may influence a judge’s rulings (Danziger, Levav, & Avnaim-Pesso, 2011). Accordingly, in his phenomenological account, Gallagher (2005) emphasizes that bodily dispositions and affordances in the environment can determine the outcome of evaluative and decisional processes more than any representation of the world internally stored by or manipulated within the central nervous system. Intelligent know-how, athletic skill, and embodied expertise emerge from the consolidation and diversification of an increasingly large repertoire of habitual interactions and affective patterns. By tapping into this repertoire, the agent finds the cognitive resources necessary to establish a structural coupling with real-world scenarios (Chiel & Beer, 1997).

Origins and Varieties of EC Approaches

The Embodied Mind (Varela, Thompson, Rosch, 1991), a seminal book that provided EC’s original formulation, has exerted a remarkable influence over various disciplinary areas and is linked with a number of doctrines and theories. In philosophy of mind, EC proposes an emergentist variety of non-reductive physicalism; in epistemology, it defends radical constructivism and pragmatism; methodologically, it promotes pluralism through an integration of phenomenological (first-person) and empirical (third-person) investigations; in theoretical biology, it embraces the autopoietic model of the living; in psychology, it advances the developmental and ecological approaches based on the agent–environment coupling through the development of enactive dispositions and practical habits; and, in neuroscience, it endorses dynamical systems theory and the explanatory models based on free-energy principle and protean (anti-localizationist) brain hypothesis.

After generating a prominent debate in cognitive science and philosophy of mind, the EC approach to cognition is today fragmented into a constellation of varying – often conflicting – positions and views (reviewed by Gallagher, 2011; Menary, 2010) concerning the role played by the body in cognitive processes. The label of 4E cognition – embodied, embedded, enacted, extended – is frequently used but, in light of recent developments, it is now more a matter of terminological convenience. Indeed, EC has further splintered into a number of camps that include, at the extremes:

- (a) Theories of minimal embodiment, which assert that the sensorimotor areas of the brain represent information in a “bodily format”, scaffolding higher cognitive functions.

- (b) Enactive (radical) theories of cognition, which claim that cognition builds on action–perception feedback loops emerging from the interaction between one’s body and environment.

And, as intermediate positions:

- (a) Embodied functionalism, claiming that some bodily systems, similar to other physical implementers, can at times play a non-trivial causal role in producing cognitive states and processes.
- (b) Biological embodiment, which maintains that the details of biological embodiment define the unique cognitive possibilities of each organism.
- (c) Semantic embodiment, which holds that the very nature of bodily experience shapes the possibilities of conceptualization and, in the case of humans, of linguistic categorization.

In particular, the interpretation of EC in cognitive philosophy is polarized around two leading approaches that contend the theoretical legacy of *The Embodied Mind*: the enactivist (or sensorimotor) and the extended (or distributed) approaches to cognition. Both accounts build on the premise that minds are causally co-determined and essentially co-constituted by the physical contingencies in which the body is contextually situated.

However, these two approaches hold different views about the major characteristics of embodiment: Enactivists, on the one hand, are generally oriented towards biological and dynamical accounts of the mind and stress that mental faculties are neither exactly localizable (because cognition is what the cognitive agent does, not an agent’s component or property) nor language-like in nature (because basic forms of cognition are not syntactically organized and do not involve propositional contents detachable from their linguistic vessels). On the contrary, according to enaction theory, cognition holistically emerges in the continuous adaptive interplay between the organism and its ecological surroundings. The genetic/constitutive role played by the sensorimotor coupling between the agent and his/her world environment is particularly emphasized over offline forms of intelligence.

Extended mind theorists, on the other hand, are typically sympathetic to functionalist and mechanistic accounts of the mind, as they stress that intracranial (neuronal, usually organic) and extra-cranial (non-neuronal, possibly artificial) components and processes have in principle the same capability to realize cognitive processes and produce cognitive states. By virtue of their equal status, internal and external vehicles of cognition are in principle interchangeable or inherently complementary, depending on whether they replace or complete one another. The same cognitive function can be implemented either by the central nervous system, by extra-neuronal bodily systems, by extra-bodily systems, or by a combination of them, if the systems in question constitute suitable functional vehicles of cognitive processing and content representation. This implies that internal mental processes such as memory can and often are offloaded on to artificial devices or socio-culturally instantiated by multiple interactive agents, as in the case of language or other intersubjective practices.

Remarkably, if most psychologists and cognitive scientists who work on sport and performance tend to endorse one variant or the other, all of them seem naturally inclined to embrace at least the core general tenets of EC (see Beilock, 2008; Gray, 2014). Such proximity is not merely coincidental, as sport psychology and embodied cognitive science share theoretical and historical roots.

This communal root is shown first in the characterization of skill acquisition as progressive automatization of complex action patterns. Such characterization was offered by some of the pioneers of the EC programme, which can be traced back as early as William James (1890)

and finds echo in the views of Dreyfus and Dreyfus (1980). These were in substantial accord with Fitts and Posner's (1967) model – a major contribution to establishing sport psychology as a scientific discipline – which posited that less cognitively demanding motor control is a distinctive mark of expertise. This is why experts can deliver their skilful performances in the form of unreflective action routines. More recently, the concept of reinvestment in motor skill acquisition research (Masters & Maxwell, 2008) congruently argues that, in expert action, subroutines are activated, and attention is reinvested into other attentional foci.

Precursor Theories – The Communal Roots of EC and Sport Psychology

EC resonates deeply with influential doctrines in developmental psychology and ecological psychology that paved the way for the foundation of sport psychology. These doctrines incarnate EC and sport psychology's common roots. Sport psychologists inherited important elements from both the Piagetian tradition (which stresses the pre-eminence attributed to "enactive learning" during early cognitive development) and the Gibsonian tradition (which has introduced the ecological concepts of "affordances" and "direct perception").

The legacy of ecological psychology is particularly remarkable for EC. The approach called "direct perception" or "direct realism" proposed by James J. Gibson (1966, 1979) stemmed, in many senses, from a reaction against the "computational" or "functionalist/information processing" approach that had been growing in the 1960s, which was largely inspired by Putnam's theory of machine-state functionalism in philosophy of mind (Putnam, 1960). Putnam's theory presupposed both the computational theory of mind (the thesis that cognitive processes are fundamentally algorithmic in nature) and the principle of multiple-realizability (which states that a cognitive function, like any other algorithmic function, can be implemented by any material support capable of physically instantiating the relevant formal operations of symbolic manipulation; see Putnam, 1967). This computational and functionalist/information-processing approach is often associated with the internalist view (typically illustrated by means of Putnam's famous "brain-in-a-vat" thought experiment; Putnam, 1981) that human cognition is entirely realized and carried out by the central nervous system alone, and the extra-neural body and the extra-bodily environment do not play any properly constitutive role in mental activity because they allegedly do not participate in information-processing.

The ecological approach proposed by Gibson offers an alternative to this functionalist, computationalist, and internalist view. It affirms the principle that animal cognition evolved to serve movement and action in the environment. In the case of human cognition, it emerges from the integration of sensory and motoric capabilities. These coordinate to generate intelligent actions that need neither representational content nor abstract inferences in order to work. In its most radical version, the Gibsonian argument rejects internal representation, characterizing cognition as direct connection from perception to action through recurrent feedback loops (this idea is also one of the precursors of the enactive/sensorimotor account of the mental; see Noë, 2005; O'Regan & Noë, 2001).

According to many researchers inspired by Gibson, human performance is primarily determined by the physics and biology of human anatomy (the muscular and skeletal physiognomy, the geometry and degrees of freedom of the effectors), and only secondarily by planning and control algorithms mediated by internal representations of the external world. The anatomical and physiological properties of the body are in many cases sufficient to explain skilful action and expertise, minimizing the role of information-processing and de-emphasizing

the importance of brain processes altogether (Carello & Turvey, 2016; Turvey, 1973; Kugler & Turvey, 1987).

The remarkable pre-eminence of Gibson's notion of affordance in the literature attests to the strong theoretical link between sport psychology and EC. The relationship between Gibsonian theory and EC is, however, multifaceted (see Chemero, 2009; Wilson & Golonka, 2013). The EC programme has tried to expand or revise the naïve realist epistemology embedded in direct perception theory, replacing it with the constructivist principle that cognition is a tentative process of unfinished negotiation and precarious adaptation to a world environment that is intelligible only to a living organism structurally coupled with it. EC has also updated some of the key ecological notions in light of recent empirical discoveries in the emerging neurosciences. For example, the notion of motor affordance was updated and expanded to accommodate the new data and models of action execution and motor control coming from neuropsychological research such as the discovery of canonical and mirror neurons in the premotor cortex.

In this updated ecological framework, EC does not emphasize just the action possibilities enabled by the physical structure of the body and the objects in the environment, but also the emotional, intentional, and motivational forces that fuel these actions, which are seen as fundamental components of skilful performance: built-in saliency detection and discrimination filters; powerful anticipatory mechanisms; adaptive systems for priority selection and allocation of interest, attention, and motivation; and flexible response habits.

The embodied nature of skill expertise does not only show the merely reactive and adaptive capability of the body. It also projects on to the capability to creatively improvise new actions: The notion of "enactive creativity" clarifies the experts' capability to strategically interpret their circumstances and adapt to unprecedented sensorimotor contingencies (Rucińska & Aggerholm, 2019; see also Hutto & Myin, 2017, for an enactivist account of imagination; llundáin-Agurruza, 2017, for a phenomenological and enactive analysis of imagination and improvisation). The embodied/enactive approach to cognition sees creativity as a capacity for active exploration of the limits of one's know-how. This capacity can be strengthened through training appropriately designed to diversify personal competences and adapted to unexpected opportunities.

In the last two decades, EC has proven advantageous for an expanding number of research areas pertaining to sport psychology. Below, to better connect theory with specific sport psychological applications, we discuss seven areas: skill acquisition and expertise; *risk sports*; attentional focus, flow, and choking in sports; tool use; social cognition; language processing in sports; and team dynamics.

Skill Acquisition and Expertise

The search for a global motor skill that would be the mark of skilled performers, posited as a "generalized motor ability" (GMA; McCloy, 1934; Schmidt, 1991) that underpins all other skills, never found sufficient supporting evidence. This led researchers to focus on specific skills with variations across groups and skill levels (Baker & Farrow, 2015). Today there are signs of a resurgence of interest in a GMA, endorsing an emergent and fluid model influenced by biology and environment over a lifespan (Hands, McIntyre, & Parker., 2018). EC can compellingly account for both GMA and specific abilities, as it relies on dynamic forms of adaptive coupling between embodied agents and their environment capable of describing the continuity between varying levels of generality in cognitive control. That is why EC can also explain the full range of expertise, from novice to advanced, while being particularly revealing with regard to experts' skills and their development. EC proves advantageous in both describing and explaining the complex

dynamics at play in top performance, which draw upon automatized processes as well as fully immersive attunement with tasks.

Ericsson, Krampe, and Tesch-Römer's (1993) *deliberate practice* concept, which proposed that expertise develops from sustained and purposeful involvement and specific training, was pivotal to the development of sport psychology. A number of foundational and influential – but non-sport-specific – works (Bloom, 1985; Starks & Allard, 1983) acted as trailblazers for sport-centred general research (Ericsson, 2003) and specific investigations thereof (Farrow, Baker, & MacMahon, 2013; Hodges & Williams, 2012; Vickers, 2016). An EC framework, given its emphasis on sportspersons' history of engagements within dynamic systems, may serve as a better theoretical foundation for such explanations.

Contemporary research in expert action in a variety of contexts is challenging long-held views in the field of sport psychology. Interestingly and controversially, opposing views in the current debate on skilful expertise correspond to different articulations of the EC approach. Fitts and Posner's principle, Dreyfus's contention, and reinvestment views posit that highly skilled performance is essentially unreflective, or that it operates solely subpersonally and by way of subroutines. Contrary to this view, more recent stances argue, from an ecological perspective, that expert athletes evince some measure of highly tuned self-awareness (Christensen, Sutton, & McIlwain, 2013, 2016; Sutton, McIlwain, Christensen, & Geeves, 2011). Experts rely not only on attuned motor control processes, but also on a tactical intelligence that dynamically adapts to contingent in-game conditions. To capture their connections, Sutton and colleagues propose a unitary model for the architecture of skilful action that integrates motor control with perception and cognition.

Risk Sports

Along these lines, *risk* or *extreme sports*, such as big wave surfing, free solo climbing (without using ropes or gear except for shoes and chalk), or wingsuit flying, push the practical and theoretical envelope. Ecological views (Immonen, Brymer, Davids, Liukkonen, & Jaakkola, 2018) highlight how traditional theory-laden psychological analyses of *extreme sports*, such as edgeworks, sensation-seeking, and psychoanalysis, have proven inadequate to capture the depth and nuances of experiences of risk sports participants – that is, such notions as adrenaline and thrill-seeking motivations prove too simple to capture the practitioners' deeply meaningful engagements. As an alternative, they advocate for studies that capture the phenomenology and ecological dynamics of these sports. Empirical evidence de-emphasizes the role of deliberate behaviour and top-down approaches that might counterbalance subpersonal routines by means of complex mental representations (Ericsson, 2006) when performers risk serious injury or even death, as Eriksen (2010) avows. That is, an EC approach is better positioned than alternative models to explain expert performances in high-stakes sports. Research on circus trapeze acrobats shows that, even if actions are highly automatized as subroutines, when much is on the line, performers rely on extensive conscious control (Eysenck, 2004). This reliance on explicit conscious control is corroborated by phenomenological analyses of experts across a range of sports (Breivik, 2013). Cognitivism is at odds with these findings: It posits subpersonal, amodal, syntactic processes that decouple online interaction from real-life experiential contingencies, whereas such studies show that the personal level and the body as attentional focus are constitutive elements of expert performance. Through neurophenomenology, EC precisely factors in both subpersonal and personal levels, advancing a holistic model of body-mind as the foundational locus for our dynamic worldly engagements – and not just a body or subcomponents thereof. That is why a highly discriminating attunement and fluid relation between online and offline processes are emblematic of EC.

Congruent with EC's framework, once a naturalistic perspective is adopted, even within cognitive systems engineering (CSE), the evidence corroborates that performers rely on a mix of formal, informal, and affective cognitive processes (Klein, 1998), and that experts depend less on formal aspects than novices (Beach & Lipshitz, 1993). Moreover, CSE forsakes informational models à la cognitivism in favour of modal description of cognitive workflows and supporting structures based on highly skilled performers' accounts (Lintern, Moon, Klein, & Hoffmann, 2018), something that neurophenomenological methodology is poised to elucidate. Moreover, the musculoskeletal, kinematic, neural, organic, and affective adaptations of expert athletes both constrain and enable the cognitive processes and reduce the cognitive load that underwrite superior performance.

The extent to which mental representations play a role in the type of intelligent and highly skilled action characteristic of experts is contested. EC approaches, by virtue of their dynamic, situated, online, and holistic (affective, intellectual, motoric) transactions with their environment, can accommodate hybrid online/offline situational requirements more parsimoniously than standard cognitivist and computational approaches. This is true even of more radical versions of enactivism that de-emphasize the role of representation in basic forms of cognition (Hutto & Myin, 2017). As previously mentioned, some EC variants – that is, enactive and biological ones – push against mental representation, opting for scaffolded cognitive models with both non-representational and representational features at different levels. Other versions – that is, minimal, functionalist, or semantic – accommodate representations all the way up and down the cognitive process. At stake is whether sport psychologists' and neuroscientists' notion of mental representation conforms to the canonical philosophical conception of representation, or whether representations are conceived as encoded neural and biological instantiations. The canonical conception understands representations in terms of veridicality, accuracy, or truth-functional conditions (do these accurately reflect states of affairs?), which is quite different from actual physiological embedment.

Enactive and biological variants support the latter view, and, indeed, the evidence seems to favour this. In the empirical literature, there are different notions of how sporting action and knowledge are represented, but none fit the canonical notion of mental representation. Even in cases where mental representations stand at the very centre of action control and organization, as in Schack and Land (2016), these correspond to dendrograms (diagrammatic taxonomic tree representations) derived from subjective labelling. Others, such as the abovementioned Lintern et al.'s (2018), explicitly discard information-processing and favour knowledge-elicitation to guide sportspersons' subjective reports, thereby operationalizing representations not as mental phenomena, but rather as a mapping tool for researchers. Accordingly, sport psychologists' notion of mental representations fit pictorial stand-ins for neurally encoded processes at best, instead of philosophical epistemic conditions.

Attention and Effort

Attentional focus, particularly in relation to effort and effortlessness, elicits less controversy, but the cognitive underpinning of effortless attention remains somewhat enigmatic. Judgement and decision-making are related to effort but distinct from it, as these involve testing of actual choices and judgements and rely on a different skillset and domain of applicability (Raab & MacMahon, Volume 1, Chapter 16, in this encyclopedia). Normally, the more effort an action requires, the harder it is to keep attention focused, with the effort eventually reaching a maximum point that it is not possible to surpass. Effort can be measured both objectively, in terms of metabolism (as in calories consumed, METs, or watt output), and subjectively, as how effortful a particular action is

perceived to be. Puzzlingly, in some circumstances, the sportsperson, fully engaged in the activity, can reach a state in which action and attention seem effortless. The objective requirements do not change putatively, but the subjective exertion perplexingly remains the same and may even decrease. Current analyses fail to accommodate this, having neglected and often relegated this issue to the realm of esoteric explanations (Bruya, 2010). Csikzentmihalyi (1991) theorized this phenomenon as an autotelic experience – vernacularly known as “flow states” – which extends across many domains, including sport and performance (Jackson & Csikzentmihalyi, 1999). The phenomenon of *quiet eye* – a technique where gaze behaviour modification improves task outcome – has also been related to flow states and effortless attention (Vine, Lee, Moore, & Wilson, 2013). EC, broadly interdisciplinary and methodologically diverse, can integrate the objective accounts based on empirical findings with the subjective phenomenological accounts.

Novice and competent sportspeople generally experience flow states accidentally. Usually, these are seen as an epiphenomenon not under the control of the athlete. Hence, typically flow states are not expressly sought in Western sports. This contrasts with the deliberate cultivation of such states in its East Asian sport and martial practices (among others). Flow states are central to martial arts with a Buddhist undertone, such as Kyudo or Kendo, where they are referred to as *mushin* (lit. without mind) states (Ilundáin-Agurruza, 2016). In these practices, *mushin* is touted as a mark of highly skilled performers and their signature fluid action. Flow and *mushin* states are characterized by consistency, economy of movement, efficacy, accuracy, and improvisational effectiveness. Whereas these typify many sports, others that unfold in more chaotic circumstances (downhill skiing, big wave surfing) belie some of these elements. Exploring the similarities and differences between flow and *mushin* transactionally – that is, considering the physiological, phenomenological, and cultural levels – is beyond the cognitivist scope but well within EC’s purview given its situated and extended view of mind–body–environment relations. Likewise, novice/expert and inter-sport variances in such states are best examined when understood as embodied dynamics that arise from adaptive agent–environment couplings.

Whether such effortless action and attention also lead to objectively improved performance is less clear, although there seems to be evidence to this effect (Swann, Keegan, Crust, & Piggott, 2016; Wulf, 2007; Wulf & Lewthwaite, 2010). This research, consistent with an EC model that endorses both subpersonal and reflective engagement in sports action, posits that it is external focus on movement, action targets, or stimuli, and not paying attention to internal sensations (whether these be proprioceptive, kinaesthetic, nociceptive, etc.), that leads to better results across the whole performance range. For example, when pedalling a bicycle, paying attention to cadence (revolutions per minute) results in better performance than focusing on muscular sensations in the quadriceps. In this case, explicit attention to external stimuli results in performance gains. Putatively, external foci allow for automatism of basic biomechanical actions and underlying mechanisms and processes. Such foci also centre attentional resources on objective target acquisition that correlates to performance objective maximization of performance goals: pedalling cadence correlates more readily with higher speeds, whereas muscular pain does not. Again, this is fruitfully analysed from an embodied perspective where physiological and mechanical conditions scaffold and structure performance as being intelligent and superior performance.

Sports Choking

Another challenging phenomenon that EC is poised to help clarify is choking under pressure in sport – the flip side to flow states, in a sense. At its simplest, when athletes choke, they fail to complete tasks that they have already mastered (Beilock, 2010). Normally, deliberative

strategies lead to better performance, but real-world and do-or-die scenarios can increase the sense of effort and lead to worse results. This can become chronic or acute and develop into a pattern vernacularly referred to as “a slump”. There are three main kinds of empirical account: (1) self-focus or self-monitoring theories (Baumeister, 1984), which attribute *choking* to paying too much attention to oneself and/or the task; (2) distraction theories, which posit that not enough attention is paid when, for instance, working memory is disrupted; and (3) overload views, in which cognitive demands overwhelm the sportsperson (these are usually conflated with and – for simplicity here – subsumed under distraction theories).

Integrating self-focus and distraction/overload accounts is not impossible, as they rely on two sets of explanations that do not entirely exclude each other (Beilock & Carr, 2001). Nonetheless, it is necessary to emphasize the different explanatory advantages of these two theories. Self-focus views best account for novice versus expert differences, particularly in terms of fast action conditions, and seem to better explain why distractors succeed in addressing choking problems (Beilock, Bertenthal, McCoy, & Carr, 2004; Beilock & Gray, 2012); distraction theories appear to more readily incorporate non-neural aspects (including environmental factors) as pressure conditions (Eysenck & Calvo, 1992). Both accounts share the underlying assumption that *choking* results from a misallocation of the attentional resources that make peak performance possible. What they contend is the nature of these attentional resources and the causes and specific modes of their misallocation.

One charge that applies to both theories is the ecological frugality that characterizes them. In other words, many of the experiments lack validity outside the laboratory and fail to reflect real-world conditions (Christensen et al., 2013; Montero, 2016). For instance, soccer players are told to dribble a ball while listening to aural cues (Beilock, Carr, MacMahon, & Starkes, 2002) or while paying attention to their feet (Ford, Hodges, & Williams, 2005), but these are not the sorts of task for which players train, nor do they mimic the competitive situations they face. Similarly, the role that pragmatic interests play in motivating athletes is sometimes underdetermined, and the modest monetary rewards offered to experiment participants as motivational stimulus hardly compare to being in contention for an Olympic medal.

EC, in virtue of the very plethora of theoretical stances and methodological resources, is well poised to supplement empirical findings, mediate in internecine theoretical disputes, and redress the ecological paucity that besets many studies. At the heart of sport psychology and embodied cognition – as the opening overture to this entry suggests – lies a recognition that cognitive activity is not merely intellectual, but also affective, perceptual, and motoric. Moreover, the high specificity inherent in sport means that, although there may be a general phenomenology to *choking* (Ilundáin-Agurruza, 2015), there are a variety of choking instances. A fully grained account of *choking* needs to account for the holistic continuities involved in skilled action and the disruptive discontinuities that characterize *choking*. To wit, physical qualities and skill, rather than isolated from each other, are so closely intertwined that skill and performance are affected by variables such as fatigue or pressure or affective state (Gabbet, 2015).

A number of studies can meet such demands from a variety of angles; for instance, Christensen et al. (2016) developed a theory they call Mesh, which “proposes that cognitive control plays an important ongoing role in advanced skill, with cognitive and automatic processes being closely integrated” such that performative responsibilities are assigned hierarchically (p. 280). This has the advantage of being able to explain in an ecologically thorough fashion how experts mesh or integrate cognitive and automatic processes or exploring how and when these decouple. Cappuccio (2017) and Cappuccio, Gray, Hill, Mesagno, and Carr (2019) assess *choking* in relation to consciousness and self-consciousness,

respectively, from an EC perspective, maintaining that understanding how either factor is involved in choking episodes needs to discriminate and correlate with nine dimensions of consciousness and recognize that self-consciousness's effect on expert skills is articulated at multiple parallel levels and causal pathways (these include emotional, situational, and other factors). EC approaches combining radical enactivism and dynamic systems theory (Hutto & Sánchez-García, 2015) provide the requisite level of complexity for such a complex phenomenon. Some (Vine et al., 2013) have studied the relationship between *choking* and *quiet eye*. The studies have concluded that failure under pressure is due to disruptions in attentional control, demonstrating that "pressure has a greater impact on visual attention during the execution of the movement than it does during the preparation of the movement" (Vine et al., 2013). These findings further line up with EC, reflecting the constitutive role that it ascribes to sensorimotor processes and perception.

Tool Use

Tools add another level of complexity to theoretical and empirical sports research, as, in addition to the often theorized bodily and mental schemas, there is a need to integrate tool use as another layer of the performative workflow. Maravita and Iriki (2004) demonstrate how sportspersons incorporate tools and instruments into their body schema, much as Merleau-Ponty's (1962) blind man's cane becomes part of his body schematic processes. In athletics, this is often referred to as "transparent equipment": Athletes integrate a wide range of inputs, from varied tactile information (skin, temperature, pressure) to proprioceptive limb position or kinaesthetic dynamics (muscular, vestibular, nociceptive, etc.), and visual data. For example, Biggio, Bisio, Avanzino, Ruggeri, and Bove (2017) studied how, in measuring sportspeople's peripersonal space and its long-term modulation, expert tennis players accurately discriminated between their personal racquet and a common one such that "the tool daily used during sport activity is stably embodied in the peripersonal space of tennis players." (p. 54). Expert players had reached a much higher level of familiarity and integration. This can be extrapolated to many other sports where sportspeople rely on equipment – for example, cycling or surfing. The tight connection between our embodiment, our tools, and the sorts of sport we devise also leads to other interesting conundrums. Sporting equipment, for example, may afford different ways to modulate the transfer of skills such as gliding, which is usually operationalized in terms of the sense of balance. Research (Krenn, Werner, Lawrence, & Valero-Cuevas, 2015) shows that, in more static sports (rifle shooting, soccer, or golf), elite athletes show better balance, whereas, in the case of alpine skiing or surfing, it is not the case. Clearly, these issues fit the methodological and theoretical parameters and basic tenets that underlie the EC paradigm.

As remarked before, EC propounds that our cognition is extended: Out-of-skull extra-skin support systems such as notebooks or phones become constitutive of cognition and not merely causally relevant factors. Some favour a functionalist/mechanistic stance (Clark & Chalmers, 1998), whereas others endorse predictive and enactive positions (Kirchoff, 2015). Both argue that tools are fully integrated as *de facto* cognitive elements. The issue now is whether tool use is simply a causally relevant factor (Adams & Aizawa, 2008) or a constitutive element. Kirchoff (2015) argues that the pertinent sense of constitution is not synchronic; rather, body, brain, and environment (tool) dynamically couple diachronically. In this view, processes "distributed across different factors/levels (neural, behavioural, environmental) and across different timescales are constituted in a temporally integrated dynamic system" (Gallagher, 2016, p. 9). Thus, by EC's lights, sports equipment is not just an external object but becomes integrated as part of the performer's body schema. This theoretical framework, when combined with

dynamic systems analyses, for instance, helps explore how embodiment constrains and enables our cognitive and skilled engagements.

Social Cognition – Embodied Approaches to Language and Action

The EC programme provides sport psychologists with updated theoretical tools to understand social interaction, informed by the recent advances in cognitive neuroscience. An example that sport psychologists (Beilock, 2015) particularly cite as evidence that neurocognitive functions and their bodily realizers must be studied together is that reading comprehension and memory improve if the subject, while reading the descriptions of certain physical activities (e.g., baseball actions), performs physical manipulations that are consistent with them (Glenberg, 2011). The “action-based” theory of reading comprehension, confirmed by neuroimaging studies, asserts that the sensory and motor systems are involved during the process of understanding, imagining, and remembering an action described in a written story, as if the reader were actually perceiving or executing that action (Glenberg & Gallese, 2012).

This investigation of the embodied scaffoldings of language confirms the results of Beilock, Lyons, Mattarella-Micke, Nusbaum, & Small (2008), who demonstrated in neuroimaging examinations of brain activation that fans of a sporting activity with scarce experience “on the field” and athletes who actually play the sport show different patterns of brain activation when watching videos of game play. Compared with fans, players’ brains show much more activation in motor-planning and motor-control regions – the areas that would be active if the subjects were actually playing rather than merely watching a match.

This and similar results – for example, those concerning motor resonance during observation of basketball tasks (Abreu et al., 2012; Aglioti, Cesari, Romani, & Urgesi, 2008) – have been interpreted as evidence of a common embodied root of motor execution, imagination, and action understanding (Gallese, 2007; Beilock, 2015). According to this interpretation, TV sports fans reflectively evaluate those actions in a way detached from their motoric dimension, whereas expert players rely on their own motor repertoire to mentally re-enact the execution of the actions they are watching. As attested by brain imaging, performative familiarity with the execution of observed actions matters for the comprehension of those actions. According to Gallese (2007), what can be lived out in imagination and in remembering depends crucially on an internal “embodied” simulation of what has been previously experienced, which employs one’s established reservoir of motor competences (for a similar conclusion suggested by a different body of evidence, see Witt, South, & Sugovic, 2014; for a review of the literature on embodied simulation informed by mirror neuron theory, see Ferrari & Rizzolatti, 2014; for a review more specifically related to sport and athletic skills, see Abreu, Esteves, & Aglioti, 2019; Ikegami, Nakamoto, & Ganesh, 2019).

Team Dynamics

Expanding the inquiry into team dynamics, the challenge is, much as with other aspects of sport, to properly analyse the many and varied kinds of team dynamics. Compared with other contexts, such as interview settings or workplace observation, Loehr, Kourtis, Vesper, Sebanz, and Knoblich (2013) confirm the unique and compounded challenges that team sports raise. EC addresses sport psychology’s need for thick, flexible, and highly textured ecological accounts that can properly analyse the many and varied kinds of team dynamics. At one extreme we find synchronized, and at the other asynchronous, performances. In between, there is an array of possible team interactions – for example, military units, firefighters, surgery

teams, and various types of team sports – with unique intersubjective manifestations, some of which are more collaborative than others and under more or less pressurized environments (Gallagher & Ilundáin-Agurruza, *in press*).

As far as synchronized performance, we find pre-reflective synchrony with teammates or partners. Musical performance is arguably the best and most studied example outside sport (Glowinski et al., 2013; Gnecco et al., 2013; Repp & Su, 2013). For instance, Soliman and Glenberg (2014) show that, when coordinating tasks such as cutting through candles with a wire that needs to be under tension, the participants modulate their actions in ways that are measurable both neuronally and behaviourally. In a sporting context, players would develop “joint body schemas” that integrate not only equipment, but also other players, inclusive of role as either adversary or ally (for joint action as a theory of event coding, see Hommel et al. 2001; for task sharing as in a doubles tennis table game, see Wenke et al., 2011). That is, rather than discrete individual body schemas, the EC stance would argue that such synchronous action is evidence for a larger system, based on joint-body schemas, that amounts to more than the sum of its parts (Gallagher & Ilundáin-Agurruza, *in press*). This fits best for cases where there is synchronization between or among agents, rather than situations with different constraints, as when antagonistic interactions are involved. At the other extreme, we find asynchronous interactions where there are adversaries who are actively working against one’s own tactics and synchronized plays. If, in the former case, the level of attunement is high and consistent, in the latter, the adversarial factor means that such consistency is missing. Surprise is the norm. This results in phenomenological and performative differences between self- and other-awareness, as well as between effective attentional focus and letting go of such focus.

More specific to sports, as far as awareness of others is involved, whether of teammates or adversaries, we find on the one hand closed-skill team sports, where there are fixed sets of movements – for example, synchronized swimming. These are closer to the analysis where synchrony is central and finds parallels in non-improvised musical performance and dance. On the other hand, open-skill team sports such as rugby, which involve continuous and dynamic adjustment to competitors’ moves, are quite different. Lacking any sort of pre-accorded movements or choreography, players must adapt to the opponents’ attempts to thwart their tactics while trying to implement their own. That is, there is a need to synchronize with teammates while working with the asynchronous surprises that adversaries unfold. In this case, EC’s methodological pluralism affords a more refined description of phenomena and more highly textured analysis at various levels of the complex dynamics involved, as there is a need to account for perceptual factors, affective states, physiological processes, biomechanical, kinetic, and kinaesthetic dynamics, and intersubjective relations, among others. Along these lines, Hoffmann et al. (2018) provide a multidisciplinary model that, taking into account the dimensions of movement dynamics, multivariate measures, and dynamic statistical parameters, aims at integrating action and cognition so as to be able to incorporate the dynamics of actions into theoretical advances.

Conclusion

EC theory can delineate a distinctive normative ideal to guide the disciplines that study sensorimotor abilities in sports activity, physical exercise, and athletic performance. Most of these disciplines focus exclusively on the medical and physiological aspects of sports performance – for example, the biomechanical and metabolic factors. The received story is that sport psychology differs from these disciplines because, unlike them, it is specifically concerned with the “mental” factors that affect sports activity. Said mental factors include the cognitive-behavioural, attitudinal, emotional, and

motivational elements that underpin one's performance. It is true that sport psychology seeks to maximize performance by acting upon these elements, primarily to optimize training methods and skill development and, secondarily, to promote other distinctively "mental" dimensions of sport, such as talent identification, team consolidation, motivation building, and the pedagogy of sport ethics. But it is wrong to assume that a dichotomous view of the mind-body relation could accurately delineate the distinctive scope of sport psychology.

Insofar as it is expected to be compatible with the naturalistic paradigm of health and medicine research, sport psychology cannot afford to found its scientific status on such a dichotomous metaphysical understanding (whether based on a dualism of substances or functions). On the contrary, EC motivates sport psychology to interpret the mind-body as a reciprocal relationship of interconnected poles or, better yet, interrelated levels of the same complex systemic reality. EC and sport psychology naturally complement each other as they arise from the same endeavour to understand the human as a dynamic body-mind synthesis: On the one hand, embodied cognitive science plays a privileged role in defining the field of application of sport psychology by highlighting the reciprocal body-mind intertwinement in sporting activity; on the other hand, EC finds the most striking confirmations of its theoretical claims in the accounts of sport performance and athletic skill provided by scientific psychology. In conclusion, EC adequately responds to the foundational requirements of sport psychology and matches its practical expectations in various fields of applied research.

References

- Abreu, A. M., Esteves, P. T., & Aglioti, S. M. (2019). Action understanding, motor resonance, and embodied simulation in sports: Bridging ecological and neuroscientific approaches. In M. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology*. (pp. 359–380). Cambridge, MA: MIT Press.
- Abreu, A. M., Macaluso, E., Azevedo, P. T., Cesari, P., Urgesi, C., & Aglioti, S. M. (2012). Action anticipation beyond the action observation network: A functional magnetic resonance imaging study in expert basketball players. *European Journal of Neuroscience*, 35, 1646–1654.
- Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. Malden, MA: Blackwell.
- Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, 11, 1109–1116.
- Aranyosi, I. (2013). *The peripheral mind*. Oxford: Oxford University Press.
- Baker, J., & Farrow, D. (2015). A [very brief] review of the historical foundations of sport expertise. In J. Baker & D. Farrow (Eds.), *Routledge handbook of sport expertise* (pp. 1–7). London: Routledge.
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46, 610–620.
- Beach, L. R., & Lipshitz, R. (1993). Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In G. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.), *Decision making in action: Models and methods* (pp. 21–35). Norwood, NJ: Ablex.
- Beilock, S. (2008). Beyond the playing field: Sport psychology meets embodied cognition. *International Review of Sport and Exercise Psychology*, 1(1), 19–30.
- Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to*. New York: Simon & Schuster.
- Beilock, S. (2015). *How the body knows its mind: The surprising power of the physical environment to influence how you think and feel*. New York: Atria Books.
- Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, 130(4), 701–725.
- Beilock, S., Carr, T. H., MacMahon, C., & Starkes, H. L. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology Applied*, 8, 6–16.

- Beilock, S. L., Bertenthal, B. I., McCoy, A. M., & Carr, T. H. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review*, 11, 373–379.
- Beilock, S. L., & Gray, R. (2012). From attentional control to attentional spillover: A skill-level investigation of attention, movement, and performance outcomes. *Human Movement Science*, 31, 1473–1499.
- Beilock, S. L., Lyons, I. M., Mattarella-Micke, A., Nusbaum, H. C., & Small, S. L. (2008). Sports experience changes the neural processing of action language. *Proceedings of the National Academy of Sciences*, 105(36), 13269–13273.
- Biggio, M., Bisio, A., Avanzino, L., Ruggeri, P., & Bove, M. (2017). This racket is not mine: The influence of the tool-use on peripersonal space. *Neuropsychologia*, 103, 54–58.
- Bloom, B. S. (1985). *Developing talent in young people*. New York: Ballantine.
- Breivik, G. (2013). Zombie-like or superconscious? A phenomenological and conceptual analysis of consciousness in elite sport. *Journal of the Philosophy of Sport*, 40, 85–106.
- Bruya, B. (2010). *Effortless attention: A new perspective in the cognitive science of attention and action*. Cambridge, MA and London: MIT Press.
- Cappuccio, M. L. (2017). Flow, choke, skill: The role of the non-conscious in sport performance. In Z. Radman (Ed.), *Before consciousness: In search of the fundamentals of mind* (pp. 323–371). Exeter: Imprint Academic.
- Cappuccio, M. L., Gray, R., Hill, D. M., Mesagno, C., & Carr, T. H. (2019). The many threats of self-consciousness: Embodied approaches to choking under pressure in sensorimotor skills. In M. L. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology* (pp. 137–155). Cambridge, MA: MIT Press.
- Carello, C., & Turvey, M. (2016) Dynamic (effortful) touch. In T. Prescott, E. Ahissar, & E. Izhikevich (Eds.), *Scholarpedia of touch* (pp. 227–240). Paris: Atlantis Press.
- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Chiel, H. J., & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, 20, 553–557.
- Christensen, W., Sutton, J., & Mclwain, D. (2013). Cognitive control in skilled action. Online draft. <https://waynechristensen.wordpress.com/2013/08/15/cognitive-control-in-skilled-action/>
- Christensen, W., Sutton, J., & Mclwain, D. J. (2016). Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind & Language*, 31, 37–66.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 7–19.
- Csikzentmihalyi, M. (1991). *Flow: The psychology of optimal experience*. New York: Harper & Row.
- Danziger, S., Levav, J., Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings National Academy of Sciences USA*, 108, 6889–6892.
- Dreyfus, H. L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, 1, 367–383.
- Dreyfus, H. L. (2005). Overcoming the myth of the mental: How philosophers can profit from the phenomenology of everyday expertise. In *Proceedings and Addresses of the American Philosophical Association* (pp. 47–65). Newark, DE: American Philosophical Association.
- Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. (No. ORC-80-2) Berkeley, CA: Berkeley Operations Research Center, California University.
- Ericsson, A. (2003). How the expert performance approach differs from traditional approaches to expertise in sport: In search of a shared theoretical framework for studying expert performance. In J. L. Starks & K. A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sports expertise* (pp. 371–402). Champaign, IL: Human Kinetics.
- Ericsson, K. A. (2006). The influence of experience and deliberate practice on development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge: Cambridge University Press.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363–406.
- Eriksen, J. (2010). Mindless coping in competitive sport: Some implications and consequences. *Sport, Ethics and Philosophy*, 4, 66–86.
- Eysenck, M. W. (2004). *Psychology: An international perspective*. Hove: Psychology Press.
- Eysenck, M. W., & M. G. Calvo. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion*, 6, 409–434.

- Farrow, D., Baker, J., & MacMahon, C. (2013). *Developing sports expertise: Researchers and coaches put theory to practice*. Oxford: Routledge.
- Ferrari, P. F., & Rizzolatti, G. (2014). Introduction: Mirror neuron research: the past and the future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1644), 1–4.
- Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Monterey, CA: Brooks/Cole.
- Ford, P., Hodges, N. J., & Williams, A. M. (2005). Online attentional-focus manipulations in a soccer dribbling task: Implications for the proceduralization of motor skills. *Journal of Motor Behavior*, 37, 386–394.
- Gabbet, T. (2015). Physical qualities of experts. In J. Baker & D. Farrow (Eds.), *The Routledge handbook of sport expertise* (pp. 121–129). London: Routledge.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.
- Gallagher, S. (2011). Interpretations of embodied cognition. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 59–70). Exeter: Imprint Academic.
- Gallagher, S. (2016). The practice of thinking: Between Dreyfus and McDowell. In T. Breyer (Ed.), *The phenomenology of thinking* (pp. 134–146). London: Routledge.
- Gallagher, S., & Ilundáin-Agurruza, J. (in press). Self- and other-awareness in joint expert performance. In E. Fridland & C. Pavese (Eds.), *Routledge handbook on skill and expertise*. London: Routledge.
- Gallese, V. (2007). The shared manifold hypothesis: Embodied simulation and its role in empathy and social cognition. In T. F. D. Farrow & P. W. R. Woodruff (Eds.), *Empathy in mental illness and health* (pp. 448–472). Cambridge: Cambridge University Press.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
- Glenberg, A. M. (2011). How reading comprehension is embodied and why that matters. *International Electronic Journal of Elementary Education*, 4, 5–18.
- Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, 48, 905–922.
- Glowinski, D., Mancini, M., Cowie, R., Camurri, C., Chiorri, A., & Doherty, C. (2013). The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves. *Frontiers in Psychology*, 4, 841.
- Gnecco, G., Badino, L., Camurri, A., D'Ausilio, A., Fadiga, L., Glowinski, D., & Volpe, G. (2013). Towards automated analysis of joint music performance in the orchestra. In *International Conference on Arts and Technology* (pp. 120–127). Heidelberg: Springer.
- Gray, R. (2014). Embodied perception in sport. *International Review of Sport and Exercise Psychology*, 7, 72–86.
- Hands, B., McIntyre, F., & Parker, H. (2018). The general motor ability hypothesis: An old idea revisited. *Perceptual and Motor Skills*, 125, 213–233.
- Hodges, N. J., & Williams, A. M. (2012). *Skill acquisition in sport*. London: Routledge.
- Hoffmann, S., Borges, U., Bröcker, L., Laborde, S., Liepelt, R., Lobinger, B. H., & Raab, J. M. (2018). The psychophysiology of action: A multidisciplinary endeavor for integrating action and cognition. *Frontiers of Psychology*, 9, 1423.
- Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA, and London, England: MIT Press.
- Hutto, D. D., & Sánchez-García, R. (2015). Choking rectified: Embodied expertise beyond Dreyfus. *Phenomenology and the Cognitive Sciences*, 14, 309–331.
- Ikegami, T., Nakamoto, H., & Ganesh, G. (2019). Action-driven and prediction-driven contagions in human actions. In M. L. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology* (pp. 381–411). Cambridge, MA: MIT Press.
- Ilundáin-Agurruza, J. (2015). From clumsy failure to skillful fluency: A phenomenological analysis of and Eastern solution to sport's choking effect. *Phenomenology and the Cognitive Sciences*, 14, 397–421.
- Ilundáin-Agurruza, J. (2016). *Holism and the cultivation of excellence in sports and performative endeavors: Skillful striving*. London: Routledge.
- Ilundáin-Agurruza, J. (2017). Muscular imaginings—A phenomenological and enactive model for imagination. *Sport, Ethics and Philosophy*, 11, 92–108.
- Immonen, T., Brymer, E., Davids, K., Liukkonen, J., & Jaakkola, T. (2018). An ecological conceptualization of extreme sports. *Frontiers in Psychology*, 9(1274), 1–9.
- Jackson, S. A., & Csikzentmihalyi, M. (1999). *Flow in sports*. Champaign, IL: Human Kinetics.

- James, W. (1890). *Principles of psychology*. New York: Dover.
- Kirchoff, M. (2015). Extended cognition and the causal-constitutive fallacy: In search of a diachronic and dynamical conception of constitution. *Philosophy and Phenomenological Research*, 90, 320–360.
- Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.
- Krenn, O., Werner, I., Lawrence, E., & Valero-Cuevas, F. J. (2015). The lower extremity dexterity text quantifies sensorimotor control for cross-country skiing. In E. Müller, J. Kröll, S. Lindiger, J. Pfusterschmied, & T. Stögl (Eds.), *Science and skiing V* (pp. 439–445). Aachen: Meyer & Meyer.
- Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. London: Routledge.
- Lintern, G., Moon, B., Klein, G., & Hoffmann, R. R. (2018). Eliciting and representing the knowledge of experts. In A. K. Ericsson, R. R. Hoffman, A. Kozbelt, & M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 165–191). Cambridge: Cambridge University Press.
- Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., & Knoblich, G. (2013). Monitoring individual and joint action outcomes in duet music performance. *Journal of Cognitive Neuroscience*, 25, 1049–1061.
- Mann, D. (2019). Predictive processing in the control of interceptive motor actions. In M. L. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology* (pp. 651–668). Cambridge MA: MIT Press.
- Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8, 79–86.
- Marin, I., & Kipnis, J. (2013). Learning and memory ... and the immune system. *Learning and Memory*, 20, 601–606.
- Masters, R., & Maxwell, J. (2008). The theory of reinvestment. *International Review of Sport and Exercise Psychology*, 1(2), 160–183.
- Matthen, M. (2014). Debunking enactivism. *Canadian Journal of Philosophy*, 44, 118–128.
- McCloy, C. H. (1934). The measurement of general motor capacity and general motor ability. *Research Quarterly*, 5(Supp. 1), 46–61.
- Menary, R. A. (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences*, 9, 561–578.
- Merleau-Ponty, M. (1962). *Phenomenology of perception* (trans. C. Smith). London: Kegan Paul.
- Montero, B. G. (2016). *Thought in action: Expertise and the conscious mind*. Oxford: Oxford University Press.
- Noë, A. (2005). *Action in perception*. Cambridge, MA: MIT Press.
- O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24, 939–973.
- Pfeifer, R., Fumiya, L., & Lungarella, M. (2014). Cognition from the bottom up: On biological inspiration, body morphology, and soft materials. *Trends in Cognitive Sciences*, 18, 404–413.
- Pijpers, J. R., Oudejans, R. D., & Bakker, F. C. (2007). Changes in the perception of action possibilities while climbing to fatigue on a climbing wall. *Journal of Sports Science*, 25, 97–110.
- Putnam, H. (1960). Minds and machines. In S. Hook. (Ed.), *Dimensions of mind* (pp. 138–164). London: Collier-Macmillan.
- Putnam, H. (1967). Psychophysical predicates. In W. Capitan & D. Merrill. (Eds.), *Art, mind, and religion*. Pittsburgh, PA: University of Pittsburgh Press. Reprinted as "The nature of mental states," in H. Putnam (1975), *Mind, Language, and Reality* (pp. 429–440). Cambridge: Cambridge University Press.
- Putnam, H. (1981). Brains in a vat. In H. Putnam (Ed.), *Reason, truth and history* (pp. 1–21). Cambridge: Cambridge University Press.
- Repp, B. H., & Su, Y. H. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review*, 20, 403–452.
- Rucińska, Z., & Aggerholm, K. (2019). Embodied and Enacted Creativity in Sports. In M. Cappuccio (Ed.), *Handbook on embodied cognition and sport psychology* (pp. 669–694). Cambridge, MA: MIT Press.
- Schack, T., & Land, W. M. (2016). Mental representation and learning. In R. J. Schinke, K. R. McGannon & B. Smith (Eds.), *Routledge international handbook of sport psychology* (pp. 723–737). Abingdon, UK: Routledge.
- Schmidt, R. A. (1991). *Motor learning and performance: From principles to practice*. Champaign, IL: Human Kinetics.
- Soliman, T. M., & Glenberg, A. M. (2014). The embodiment of culture. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 207–220). London: Routledge.
- Sutton, J., McIlwain, D., Christensen, W., & Geeves, A. (2011). Applying intelligence to the reflexes: Embodied skills and habits between Dreyfus and Descartes. *Journal of the British Society for Phenomenology*, 42, 78–103.

- Swann, C., Keegan, R., Crust, L., & Piggott, D. (2016). Psychological states underlying excellent performance in professional golfers: "Letting it happen" vs. "making it happen". *Psychology of Sport and Exercise*, 23, 101–113.
- Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review*, 80, 1–52.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Vickers, J. (2016). The quiet eye: Reply to sixteen commentaries. *Current Issues in Sport Science*, 118, 1–18.
- Vine, S., Lee, D., Moore, L., & Wilson, M. (2013). Quiet eye and choking: Online control breaks down at the point of performance failure medicine and science. *Sports and Exercise*, 45, 1988–1994.
- Wenke, D., Atmaca, S., Holländer, A., Liepelt, R., Baess, P., & Prinz, W. (2011). What is shared in joint action? Issues of co-representation, response conflict, and agent identification. *Review of Philosophy and Psychology*, 2, 147–172.
- Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology*, 4, 58.
- Witt, J. K., South, C. S., & Sugovic, M. (2014). A perceiver's own abilities influence perception, even when observing others. *Psychonomic Bulletin and Review*, 2, 384–389.
- Wulf, G. (2007). *Attention and motor skill learning*. Champaign, IL: Human Kinetics.
- Wulf, G., & Lewthwaite, L. (2010). Motor skill learning and performance: A review of influential factors. *Medical Education*, 44, 75–84.

Support Publication #3 [SP3]

Cappuccio M.L., Wheeler M. (2012),
“Ground-level Intelligence:
Action-Oriented Representation and Dynamics of the Background”
in Zdravko Radman (ed.), *Knowing without Thinking. Mind, Action, Cognition and the
Phenomenon of the Background*,
Palgrave Macmillan, London.

Reproduced with permission from Palgrave (Springer Nature). All rights reserved.

1

Ground-Level Intelligence: Action-Oriented Representation and the Dynamics of the Background

Massimiliano Cappuccio and Michael Wheeler

Studies of embodied intelligence have often tended to focus on the essentially responsive aspects of bodily expertise (for example, catching a ball once it has been hit into the air). But skilled sportsmen and sportswomen, actors and actresses, dancers, orators, and other performers often execute ritual-like gestures or other fixed action routines as performance-optimizing elements in their *pre*-performance preparations, especially when daunting or unfamiliar conditions are anticipated. For example, a recent movie (*The King's Speech*) and a book of memories (Logue and Conradi, 2010) have revealed that, just before broadcasting his historic announcement that the United Kingdom was entering the Second World War, King George VI furiously repeated certain tongue twisters in a resolute effort to overcome his relentless stutter. Such ritualized actions don't merely change the causal relations between performers and their physical environments (although this may well be part of their function), but they provide performers with the practical scaffolds that summon more favourable contexts for their accomplishments, by uncovering viable landscapes for effective action rather than unassailable barricades of frightening obstacles. In other words, while the kinds of embodied skills that have occupied many recent theorists serve to attune behaviour to an actual context of activity, whether that context is favourable or not, preparatory embodied routines actively refer to certain potential (and thus non-actual) contexts of a favourable nature that those routines themselves help to bring about, indicating the possibilities of actions disclosed by the desired context. As we shall see, this sort of transformative event, which is exemplified by, but not confined to, the ritualized gestures and routines of skilled performers, is a regular occurrence in everyday skilled activity, not the crowning achievement of a few talented individuals; so the capacity in question belongs centrally to our ordinary suite of bodily skills. The theoretical ramifications of that embodied capacity are the topic of this chapter.

Our ability to skilfully indicate and reconfigure contexts is intimately intertwined with the widely cited, but (we think) incompletely understood, phenomenon of the *background*. Following others, we take the background to be the implicit and plausibly endless chains of preconditions (bodily, attitudinal, social, cultural) that provide the context-dependent meaning and normative relevance conditions for any specific intelligent action. Consider, for example, the innumerable preconditions of this kind which are in play when an intelligent being walks into a BBC studio and recognizes it as the uncanny setting of an historically momentous, personally challenging, public speech. Our remarkable capacity to navigate the open-ended and shifting structure of a human life by way of the background is what Dreyfus (2008) calls ‘background coping’ or ‘ground-level intelligence’. In what follows, we shall argue that the defiantly nonrepresentational conception of ground-level intelligence developed and defended by Dreyfus himself, and by others who share his general approach, is ultimately unable to do justice to the distinctive dynamics of background, precisely because that conception, at least partly as a consequence of its representation-shunning character, fails to encompass the particular, transformative, background-involving embodied capacity so strikingly illustrated by the King’s routine.

1. Absorbed coping, the background, and the problem of relevance

According to one generic, orthodox view, traditionally dominant in areas such as cognitive science and artificial intelligence (AI), intelligent behaviour is internally mediated, in multiple and varied ways, by models of the world, specifications of pre-given goal-states, and/or problem-solving procedures based on stored rules and heuristics. This theoretical package – call it *cognitivism* – is no longer mandatory, even in cognitive-scientific circles. One source of disquiet comes from a philosophical approach to intelligence that, in the hands of some thinkers, has recently and controversially been converted from a hostile brake on the ambitions of cognitive science into an emerging alternative conceptual framework within which cognitive science might be developed. (For some of the details of this transformation, see for instance Kiverstein, forthcoming; Wheeler and di Paolo, forthcoming.) First advanced by Dreyfus (for example, 1992, 2002a, 2002b), and then in various forms by, for example, Kelly (2000, 2002), Rietveld (2008, forthcoming) and Wheeler (2005, 2008, 2010; Cappuccio and Wheeler, 2010), the approach in question draws its inspiration from phenomenological thinkers such as Heidegger and Merleau-Ponty. In its most prominent form (to be placed under scrutiny here), the view takes everyday intelligent activity to be most revealingly, characterized by a mode of engagement with environmental entities that Dreyfus (2002a) has dubbed ‘absorbed coping’, understood as the skilful and fluid adjustment of behaviour to context-dependent

contingencies by way of a richly adaptive, direct (that is, unmediated by representations or any subject–object interface), situated sensitiveness to what is relevant.

The notion of absorbed coping will be important in what follows, so it is worth pausing here to unpack it a little. The behaviour of an absorbed, coping agent, motivated to achieve his practical goals, is not regulated by any set of truth conditions attached to a pre-specified goal-state (that is, a description of the state of affairs in which a belief that some goal-state obtains would be true; cf. means-end analysis AI algorithms that produce plans for reducing the difference between the current state of the world and some explicitly represented goal-state). Rather, the behaviour of an absorbed, coping agent is regulated by that agent's capacity to sense intensities of deviations from a contextually determined optimal balance with her environment, coupled with her instinctive practical knowledge of how to adjust her behaviour to improve her performance and thus reduce her sense of being out of balance. With absorbed coping identified as a kind of zone of expert performance, skill acquisition and direct sensitiveness to contextual relevance may be seen as developing together and as implicating each other, since coming to have a mature performative competence for a certain action means becoming able to recognize what situations afford it appropriately, and distinguishing the fine-grained articulation of a situation means knowing how to navigate it competently (Dreyfus, 2002a).

One might articulate this picture further by identifying the absorbed copers' direct sensitivity to relevance – his unmediated sense of being in or out of balance with her environment – as indicating the property of *thrownness* (Heidegger, 1996). As thrown, the absorbed copers always find themselves in a world that *matters* to him in some way or another. This always-already-meaningful world into which he is thrown is encountered fundamentally in terms of practical dispositions towards context-dependent affordances (possibilities for action presented by the environment). But, crucially, the contentful structure (or significance) of this world is neither fully articulated within the practical knowledge that enables the copers to negotiate situations, nor is it fully articulable in some ideal theoretical register. This is because that structure encompasses a vast and indeterminate web of implicit preconditions for sense making, “an unexplicated horizon” or *background*, providing “the vantage point from out of which” every experience matters to one in certain way (Taylor 1993, p. 325). As we shall understand this horizontal background, it contains every interwoven element of our adaptive know-how (whether bodily, attitudinal, social, or cultural) that is presupposed by our concrete practical engagements. The background is thus itself a body of adaptive know-how, although one which is implicitly presupposed by, rather than on open display in, everyday patterns of skilled activity. This suggests that our epistemic relationship with the background (our familiarity with, and capacity to smoothly navigate, its unarticulated patterns of

significance) is itself a kind of coping, hence Dreyfus's (2008) term 'background coping'. Moreover, since the background is, in effect, a transcendental condition for absorbed coping, it constitutes a deep-structural feature of skilled expert performance; so background coping constitutes our *ground-level intelligence* (see again Dreyfus, 2008).

If we view intelligent behaviour in the light of its thrownness, we can offer a compelling diagnosis of why cognitivist AI, as identified above, has struggled to provide any general solution to the so-called *frame problem* (for this claim, see Dreyfus 1992, 2008; for discussion, see Wheeler 2005). Here, the frame problem is to be interpreted in the widest possible sense, as the problem of building a naturalistically discharged system (for example, a computational machine) that can process information and produce behaviour in a manner that is fluidly and flexibly sensitive to context-dependent relevance. In this general form, the frame problem might simply be re-named the *problem of relevance*. Here is how the problem manifests itself. Faced with the challenge of determining which of its behaviour-generating rules and representations are relevant in the present context, the cognitivist agent might naturally deploy second-order rules and representations that determine first-order contextual relevance. But this strategy can succeed only in pushing the issue of relevance one stage back, for the system then needs to decide which of its stored heuristics or potentially context-specifying representations are currently relevant, a challenge which requires a further, higher-order set of heuristics or representations, and so on. Put another way, to the extent that cognitivist AI persists in attempting to capture the background (the preconditions of our context-situated intelligence) in terms of explicit representations and rules, it runs headlong into an infinite regress of context-specifying structures. Once we adopt the perspective of thrownness, however, the conditions that generate this sort of infinite regress are never established, because appropriate behaviour selection and modification are rooted in a non-representational direct coupling between agent and environment which already embodies sufficient sensitivity to relevance. In particular, because the normative preconditions of an action are implicitly embedded in the context that the agent inhabits as a practical scenario – as features of that agent's background – they don't need to be scrutinized as aspects of a problem to be decoded and solved.

At this point, things get more complicated, for it seems to us that there are two different dimensions to the problem of relevance. First there is an *intra-context* problem, which challenges us to say how a naturalistically discharged system is able to achieve appropriate, flexible, and fluid action within a context. Then there is an *inter-context* problem, which challenges us to say how a naturalistically discharged system is able to flexibly and fluidly switch between an open-ended sequence of contexts in a relevance-sensitive manner (Wheeler 2008, 2010). If this distinction between an

intra-context and an inter-context problem of relevance is indeed genuine (criticisms of the distinction will be considered later), an intriguing question suggests itself: are the nonrepresentational processes that we have met so far under the banner of Dreyfusian ground-level intelligence sufficient to account not only for our within-context sensitivity to relevance, but also for our capacity for relevance-sensitive, open-ended context-switching? It is to this question that we shall now turn.

2. Solicitation and summoning

If we add a little more detail to the picture of ground-level intelligence sketched so far, an affirmative answer to our question comes into view. The key here is an application of Merleau-Ponty's (1962) notion of the *intentional arc*, according to which skills are not internally represented, but are realized as contextually situated *solicitations* by one's environment that tend to become more fine-grained with experience (see Dreyfus 2008, 340). Thus, as Gallagher (2008) explains, when poised to engage in the action of climbing a mountain, the skilled climber does not build an inner representation of the mountain and infer from that plus additionally represented knowledge of her own abilities that it is climbable by her. Rather, from a certain distance, in particular visual conditions, the mountain 'simply' looks climbable to her. Her climbing know-how is 'sedimented' in how the mountain looks to her and thus may *solicit* the action of climbing from her. Clarifying this idea further, Rietveld (forthcoming) usefully draws a distinction between different kinds of affordance. Given a specific situation, some affordances are mere possibilities for action, where 'mere' signals the fact that although the agent *could* respond to them in some way, such a response would be contextually inappropriate. In the same situation, however, some affordances, precisely because they are either directly contextually relevant to the task at hand, or have proved to be relevant in similar situations in the past, prime us for action and thus, as Rietveld (forthcoming) puts it, render us ready to act in appropriate ways by being bodily potentiating. It is affordances of the latter kind that are identified by Rietveld as Merleau-Pontian solicitations, divided into figure solicitations and ground solicitations. Figure solicitations are those with which we are actively concerned. Ground solicitations, by contrast, are those with which we are not currently concerned, but for which we are currently potentiated, and which are thus poised to *summon* us to act (see Rietveld forthcoming). According to this phenomenological analysis, the background structures of ground solicitations, together with the process of summoning that those structures support, are the non-representational conditions that explain our capacity for adaptive context-switching.

It is important to note that the distinction between figure and ground solicitations should be conceived not in terms of two fundamentally

separate kinds of element governed by distinct proprietary mechanisms, but in terms of different perspectival depths within the same overall structure of significance. These perspectival depths correspond respectively to the proximal concerns of an occurrent practical context and the potential retrieval of relevant sense-making elements from an entire life and history (an immanent and finite condition of being-in-the-world; Heidegger 1996). In harmony with the idea that intelligent behaviour involves a complex but ultimately continuous landscape of varying perspectival depths, Dreyfus (2008) argues that it is at root the same practical, competent disposition to maintain a dynamical balance with contextual circumstances that underlies within-context absorbed coping and cross-contextual summoning, and that this competence is materially implemented at the subpersonal level by the same set of complex, dynamical, non-representational structures and processes. To provide an example of such subpersonal structures and processes, Dreyfus calls on the neurodynamical framework developed by Freeman (2000), in which the brain is conceptualized as a non-representational dynamical system primed by past experience to actively pick up and enrich significance. The constantly shifting attractor landscape of such a system physically grounds Merleau-Ponty's intentional arc, by causally explaining how newly encountered significances may interact with existing patterns of inner organization to create new global structures for interpreting and responding to stimuli.

Taken to the limit, the line of thought that we have been laying out generates scepticism about the very distinction between intra-context and inter-context versions of the problem of relevance. Thus Rietveld (forthcoming) argues that sensitivity to within-context relevance essentially coincides with the broader sensitivity to global relevance that guides non-representational background coping. If context-specific activity were genuinely encapsulated (that is, if it relied only on segregated portions of the background), then our responses to local contingencies, even if successful, would be at constant risk of neglecting the background on which their broader appropriateness depends. Rietveld illustrates this risk by drawing an analogy between contextually encapsulated relevance-sensitivity and a neuropsychological pathology called 'utilization behaviour'. Patients with this disorder are excessively sensitive to the local affordances provided by the immediate environment, and suffer from an inability to discriminate between appropriate and inappropriate behavioural responses (Lhermitte, 1986, p. 342). As a result, contextually inappropriate actions are not inhibited, even though broader background knowledge should discourage them. For example, a patient exhibiting utilization behaviour may start compulsively making the bed that she sees before her, even if this bed is in someone else's house. Rietveld imputes such behaviour to a disruption of the normal tendency to situate the local context of action within its holistic background, a tendency that he calls sensitivity to 'real relevance'. In his view, this impairment in

sensitivity to real relevance reflects what would happen in normal subjects if context-specific activity were encapsulated.

For Rietveld, then, ground-level intelligence underwrites a behavioural profile in which solicitation makes sense *only* where there is an *intrinsic* possibility of summoning. This, he concludes, destabilizes the distinction between intra-context and inter-context sensitivity to relevance. The point might be put like this: if the preconditions of all skills, including the most specialized ones, are part of an holistic, unrepresentable “background sense of reality [...], something we possess in – that is inseparable from – our actual dealings with things” (Taylor, 1993, p. 327), then arguably there is no precise boundary separating the context-specific forms of coping from background coping, but only local ways for the latter to be revealed within the former. To return to our opening example, the context-specific task of speaking on national radio wouldn’t have been such a tremendous ordeal for George VI, if that task wasn’t interpenetrated by a range of background conditions, including some excruciating personal circumstances, institutional and familial expectations, the fact that he had recently become king in a very dramatic political contingency, and so on.

Rietveld proceeds to deny the target distinction on additional, neurofunctional grounds. Here he targets previous work by Wheeler (2008, 2010), in which it is argued that the intra-context problem of relevance may be neutralized by a sub-class of encapsulated mechanisms that he (Wheeler) calls special-purpose adaptive couplings (SPACs). The precise details of Wheeler’s analysis and argument need not concern us here (although a little more detail is given below, during the discussion of Mataric’s robot, Toto), but the general thought in play is that because SPACs become activated correctly only in the presence of the right, contextually relevant input, the intra-context problem of relevance never arises for them. Rietveld suggests that the best candidates for neural SPACs are certain context-specific installed routines present in the lateral premotor areas of the brain (routines that involve mirror neurons, canonical neurons, and other action-specific families of motor neurons). But, he observes, ‘preaffference shows that the triggering of SPACs [...] presupposes the proper functioning of the medial frontal system that determines the sensitivity of the sensory system’ and ‘tunes it to respond to what is currently significant to the individual’ (Rietveld, forthcoming, p. 23; cf. Dreyfus, 2008, p. 350). Importantly, this modulation of the system’s sensitivity is something that normally happens before the stimulus is picked up. So the proper functioning of the premotor system that instantiates SPACs is dependent on the prior activity of the medial premotor system, the system that, according to Rietveld, orients the sensory expectations of the agent, tuning her general sensitivity to the overall context in which she is situated. Once again, then, the distinction between intra-context and inter-context sensitivity to relevance is placed under threat.

We introduced this section by wondering whether the kind of non-representational account of ground-level intelligence advocated by Dreyfus has the resources to account for all forms of sensitivity to relevance. By now, it may seem that a compelling affirmative answer has been given to that question, and that the boot has been transferred to the other foot. In other words, rather than worrying about the scope of the non-representational account of background coping, it seems we should be asking whether it is possible for background coping ever to engage representations. So let's ask that question.

3. Ground-level representations

At this point, we need to back up slightly. We wish to propose two principles pertaining to the nature of the background, principles that, we think, any adequate account of the background would need to respect. The first, strongly suggested by our foregoing discussion of Dreyfus and Rietveld, is what we shall call the 'principle of unity'. This states that the background cannot be decomposed into self-sufficient, encapsulated parts without sacrificing the holism that characterizes its transcendence. The second is what we shall call the 'principle of mutuality'. The justification for this principle comes from Merleau-Ponty's (1962, p. 159) pregnant observation that 'movement and background are, in fact, only artificially separated stages of a unique totality'. We take this observation to imply that background coping and context-specific coping enjoy a background-foreground relationship (recall the discussion of different perspectival depths in a single landscape of significance; see also Cappuccio and Wheeler, 2010), but also, and crucially, that the relationship in question is the one in which the two modes of coping, conceptualized (somewhat artificially, as Merleau-Ponty points out) as separate stages in the process of sense-making, are able to *dynamically shape each other*. The principles of unity and mutuality enable us to re-ask the question of whether or not there exist ground-level representations. The principle of unity suggests that background coping may discriminate between contexts by somehow representing the criteria for such discrimination, if and only if some contingent configurations of the background could act as representations of the background itself, without this implying a decomposition of the background into self-sufficient, encapsulated parts. The principle of mutuality suggests that the non-representationalist account of ground-level intelligence advocated by Dreyfus and Rietveld is unable to do justice to the complex and subtle dynamics of the background. It might seem as if this claim must be wrong, since it might seem that the Dreyfus-Rietveld view *embraces* the principle of mutuality. In what follows, however, we shall argue that such a view encompasses an overly restrictive, essentially static account of the relationship between context-specific coping and background coping, an account that supports a relation of one-way

dependence, but not one of dynamic co-dependence. But if that is right, then the possibility that there exist ground-level representations is back on the agenda. The challenge would then be to show that there are situations which are consistent with – or perhaps, to raise the bar further, suggested by – the principle of mutuality, in which the presence of ground-level representations is not merely possible, but to be expected. We think that this challenge too can be met, and in its more demanding form.

To put us on the right road, here is an opening thought. From a Heideggerian phenomenological perspective, there is every reason to think that there exists a mode of context-specific coping that is representational in form. The key here is a way of encountering entities that Heidegger (1996) calls *un-readiness-to-hand*. This is in contrast with *readiness-to-hand* (the transparency of equipmental entities in hitch-free use, which is correlated with non-representational absorbed coping) and *presence-at-hand* (the explicit consciousness of entities as full-blown objects, which is correlated with the kind of fully representational, decontextualized reasoning prioritized by the cognitivist picture). Un-readiness-to-hand emerges paradigmatically when skilled practical activity is disturbed by broken or malfunctioning equipment, discovered-to-be-missing equipment, or in-the-way equipment, although, as Cappuccio and Wheeler (2010) have stressed, un-readiness-to-hand does not need to be the result of some actual disturbance to a *pre-existing* flow of absorbed coping, but may be established by an attitude on the part of the agent. When encountered as un-ready-to-hand, entities typically solicit the agent to temporarily monitor and plan his actions, in order to solve context-specific practical problems posed by the environment. For instance, if I am on the way to work, a broken watch is encountered by me not as a lump of metal of measurable mass (that is, as a present-at-hand object removed from the range of my practical concerns), but as a faulty timekeeper, a damaged tool that constitutes a hindrance to my train-catching activity. It thus solicits certain remedial, context-dependent, problem-solving actions (for example, the use of alternative time-keeping resources, planning a visit to a watchmaker).

Heidegger's analysis suggests further that the kind of practical problem-solving distinctive of un-readiness-to-hand involves representational states (Wheeler, 2005, 2008, 2010). Crucially, however, these are not the full-blooded cognitivist representations that plausibly mediate epistemic access to the present-at-hand. When revealed as present-at-hand, an entity will be experienced in terms of properties that are action-neutral, specifiable without essential reference to the representing agent, and context-independent. By contrast, when revealed as un-ready-to-hand, an entity will be represented in terms of properties that are action-specific, egocentric, and dependent on a particular context of activity. So, for example, the online, task-engaged navigator may represent the external environment by way of an egocentrically defined space in which obstacles appear only as regions to

be avoided, positioned in terms of roughly specified bearings relative to her own body. The character of the kinds of representations distinctive of un-readiness-to-hand is an issue to which we shall return. For the moment, the key point is that, although practical problem-solving in the domain of the un-ready-to-hand is not absorbed coping (by hypothesis, absorbed coping has been disturbed), it nevertheless remains a form of context-specific coping (since it is paradigmatically a skilled adaptive process oriented towards re-establishing a dynamical balance with a specific extant situation). From now on, then, we shall use the term *skilled coping* as an overarching term covering *both* forms of context-specific coping. Using this terminology, the point on the table right now is that skilled coping may come in non-representational and representational forms.

Perhaps surprisingly, Dreyfus agrees that, in cases of un-readiness-to-hand, skilled coping within specific contexts occurs in a representational modality. He argues, however, that 'all coping, including unready-to-hand coping, takes place on the background of [a] basic nonrepresentational, holistic, absorbed, kind of intentionality, which Heidegger calls being-in-the-world' (Dreyfus 2008, 345–6). In other words, skilled coping, whether representational or not, is derivative on background coping, which, by Dreyfusian lights, is necessarily non-representational. Thus although representations may figure in certain forms of intelligent behaviour within the more fundamental condition of thrownness (that is, provided that background coping is already in place), they cannot figure in the ground-level process of background coping itself. On this model, then, the dependency relation between background coping and skilled coping is *unidirectional* and *bottom-up*. And that's why the Dreyfusian model, as built on by Rietveld, ultimately falls short of being a satisfying account of the background. Its commitment to a one-way dependency relation between skilled coping and background coping collides head-on with the Merleau-Pontian prerequisite of a strict dynamic co-constitution between an action and its background; in other words, it flouts the principle of mutuality.

So now what happens if we wholeheartedly pursue a model in which background coping and skilled coping are genuinely and dynamically co-dependent? To set the scene, let's return once more to our opening example. It is true both that George VI was coping with his background when he forced himself to give his speech on the national radio, and that his oration constituted an instance of skilled coping directed at producing a certain performance. Now, on the non-representational account of ground-level intelligence, some aspects of this story remain decidedly unexplained. What really did push King George VI to give his speech, in spite of his debilitating stutter, taking on an institutional duty that intimidated him and that his brother had declined not long before? It seems difficult to claim that his decision to confront the local context of that public announcement was fluidly coupled with, and unproblematically derived from his personal

and political history (and thus from his background coping), as the emotional tone of the situation seemed strongly to discourage that decision; so much so, in fact, that even his personal identity and institutional function, and not just the contingent circumstances of the speech, could be seen as an anguishing state of affairs to be avoided (more on *Angst* below). After all, George had always thought of himself as a marine officer, and was totally unprepared to be crowned until a few weeks earlier. We can infer from this that neither the king's existential background nor the local context of his action established preconditions or affordances that positively summoned or solicited him to accomplish his duty. In truth, both the global background and the local context hindered his action so strongly that he had to suspend his direct coupling with them, and actively create a more favourable context into which to transfer his performance. This shift could be achieved only through a resolute effort on the part of the king to become aware of, and to change, the conditions of his contextual situation, and not through a process in which he unreflectively accommodated himself to those conditions. Hence the disclosive and transformative function of the king's embodied preparatory routines, his furiously repeated tongue twisters. Some skilled embodied practices (like preparatory gestures), then, don't merely achieve an appropriate response to the actual contextual opportunities of action, but rather address the context itself as a problem, in order to reconfigure it and to disclose new opportunities for action.

This capacity to reconfigure the background, in order to achieve local context-creation, is an idea that bubbles away tantalizingly, alongside Dreyfusian summoning, in Sutton's (2007) compelling recent analysis of expert batting in cricket. Sutton observes that certain 'individualized "pre-ball routines" as the batsman prepares and takes guard act [...] as a transportable sequence of consistent and comfortable signs which prepare the mindful body for action' (p. 774; more later on this crucial idea of embodied routines as signs). These preparatory embodied routines are then linked, in a way that is suggestive of context-creation, to the ability of expert batsmen to reset their response profiles at key moments in the game.

When the match situation is changing rapidly and continually – over the crucial dying overs of a decisive one-day game, for example – good players will be constantly resetting their response repertoire in ways which may have been discussed or partly planned out in advance, either deliberately or simply as the result of the sedimented history of relevant experience. This doesn't mean deciding in advance that only one stroke is allowable 'no matter what,' but rather altering the probabilities of attempting certain shots to certain ranges of possible deliveries [...] One successful case was when, during the one-day internationals before the 2005 Ashes series, Andrew Strauss set himself more than once to get way across to the offside, outside the line of good-length balls from Jason Gillespie and use

the pace to lift them over fine leg, a shot unthinkable in less audacious circumstances. (Sutton, 2007, p. 775)

At this point, a clarification (or perhaps a complication) is in order. Although further analysis is required, it is arguable that our examples of preparatory embodied routines do not reward the kind of ‘voluntaristic’ interpretation that one might naturally find most tempting. Indeed, in our view, there is reason to be suspicious of any claim that the reconfiguration of King George VI’s background was the outcome of an absolutely free choice on his part, a choice in which he employed some sort of latent ‘inner power’ against the external circumstances. Heidegger (1996, p. 176) warns us that to interpret the phenomena of resoluteness and deliberation in these sorts of voluntaristic terms would be to miss the fundamental fact that there cannot be any decision that is not always already decided by its background preconditions. This Heideggerian principle, it seems, should apply to those decisions that concern the background itself. After all, one might see King George VI’s transformation of his predicament as itself situated within an acceptance of certain broader circumstances, meaning that the decision to deploy specific preparatory embodied routines was the only possibility disclosed to the king at that time, the only thing he could do to respond to the dramatic call of the situation. So decisions cannot be abstracted from their background altogether. Nevertheless, that doesn’t alter the fact that some decisions must be taken, those that serve to abstract the agent’s behaviour away from *certain extant configurations* of the background while producing *new specific configurations*. As we have seen, the king’s speech simply couldn’t have been delivered, if he hadn’t used his preparatory embodied routines to gather his resolute decision, in part as a way of ignoring those context-specific solicitations that would have positively hindered his skilled performance.

Preparatory embodied routines enable a special kind of context-switching in which an appropriate set of action-soliciting conditions is called forth actively by those very routines. Context-switching is here not a response to shifting environmental circumstances, but an agent-driven intelligent strategy for adaptively structuring behaviour. Our proposal is that the functioning of such routines needs to be understood as one which is representational in form. An immediate reason for thinking that representational language is appropriate here draws on our earlier Heideggerian claim that un-readiness-to-hand (the disruption of absorbed coping) ushers in a representational mode of agent–environment engagement, because it opens up a problem-solving cognitive distance between agent and environment. It seems that preparatory embodied routines assume the same kind of problem-solving cognitive distance between agent and environment as is operative in the domain of un-readiness-to-hand, suggesting an extension of the representation-involving explanatory template from the latter to the

former. Turning this suggestive analogy into a compelling explanatory perspective requires some additional conceptual machinery.

4. Signs as action-oriented representations

There are situations in which intelligent behaviour requires us to deal with the whole context of our current coping activity. This happens when we need to recall the whole network of practical preconditions that define our present possibilities of action, possibly to contrast them with other, non-actual possibilities that are connected to them. In Heidegger's phenomenological analysis, this experience is initiated by *Signs*.

In the theoretical sense that matters here, the class of Signs is not limited to graphic signs or conventional means of communication, but potentially includes any situational element (event, state, process, or thing). When revealed as Signs, situational elements make us aware of our context and delineate its normative boundaries. A 'sign is something ontically ready-to-hand which is not just this particular equipment, but functions as something that indicates the ontological structure of readiness-to-hand, referral-totality and worldhood' (Heidegger, 1996, p. 82). Any situational element can become a Sign, just as soon as it emerges out of the background of our skilled coping activity and starts to orient our attention towards those background conditions themselves. For example, malfunctioning or unfamiliar pieces of equipment are likely to become Signs of the problematic situation in which they show up, introducing un-ready-to-hand interactions. Thus a road sign informing the skilled driver of a detour announces that her absorbed coping with the road is interrupted, and indicates that she must withdraw from her current activity of automatic driving, possibly switching to an alternative, reflective modality of driving, in order to follow or find an alternative route home. For a stuttering orator who is on the threshold of giving an important speech, the smallest hesitation in pronouncing a single syllable may become a Sign of his challenging situation. As a Sign, his hesitation reminds him that his oratorical task is risky, that he must focus carefully on the training received, that he cannot fail for the sake of his personal, familial, and institutional duties. This attention combines with a sudden awareness of many preconditions (about his present situation, past training, future historical consequences) that were buried in the background of his oratorical activity until the instant that he hesitated on that syllable. Even when coping proceeds smoothly, ready-to-hand pieces of equipment may morph into Signs, disclosing to our attention that the whole context of action, rather than just this or that piece of equipment, matters to us and constrains our conduct. Thus a delicious smell coming from the oven is a Sign reminding us of our current activity of cooking and making us aware that it is time either to check the progress of our almost baked cake or to switch to the context of preparing the table for dinner. Signs

thus operate in ready-to-hand and in un-ready-to-hand scenarios. The main difference is that while, in the former kind of scenario, Signs create the possibility of reflecting on one's absorbed coping (including the possibility of disrupting that coping) as a successful outcome of the coping process itself, in the latter kind of scenario, in which an interruption to absorbed coping has already happened, Signs more directly establish the shape of the appropriate reflective response.

So, by standing out from their inconspicuous surroundings, Signs call the skilled agent to pay attention to the circumstances, the preconditions, and the goals of her activity. So conceived, Signs often allow us to adjust our behaviour to the present context of action. Equally, however, because of those same *signs'* standing out, background-disclosing function, Signs may provide the instigating means by which adaptive context-switching – whether reactive *or proactive* in nature – may occur. It is at this point that it becomes illuminating to conceptualize preparatory embodied routines – routines whose function is precisely to reconfigure the background, in order to transform the context for action – as a special sub-class of Signs. This Heideggerian gloss, we think, allows us to appreciate the full scope and importance of Sutton's (2007, p. 774) cricket-related remark that 'individualized "pre-ball routines" as the batsman prepares and takes guard act [...] as a transportable sequence of consistent and comfortable *signs* which prepare the mindful body for action' (emphasis added).

The next piece in the theoretical jigsaw we are building is to conceptualize Signs, and thus preparatory embodied routines, as representations. What this would establish is that, contra Dreyfus, context-switching may be enabled and driven by representational structures. Of course, Dreyfus holds that representations cannot solve the problem of relevance. Indeed, he goes further, by suggesting that, from a Heideggerian phenomenological perspective, the problem of relevance is revealed, at least partly, to be an artefact of representationalism. As he puts it, 'for Heidegger, all representational accounts are part of the problem' (Dreyfus 2008, p. 358). If Dreyfus is right, our attempt here to interpret the relevance-sensitive structures that are Signs, and thus to interpret preparatory embodied routines, in representational terms is doomed to failure. But is it right? To resist Dreyfus's anti-representational scepticism, we need to remind ourselves that representations come in more than one conceptual flavour. As we have seen, in the fully decontextualized mode of presence-at-hand, representations in experience will encode properties that are essentially action-neutral, specifiable without any necessary reference to the representing agent, and context-independent. A map in a Cartesian co-ordinate system would be an example of a present-at-hand representation. In the mode of un-readiness-to-hand, however, entities will be represented in experience in terms of properties that are action-specific, egocentric, and intrinsically embedded in a particular context of activity. Representations with this profile have been dubbed

action-oriented representations (henceforth AORs; for further discussion, see for instance Wheeler, 2005).

It will be useful to deepen our appreciation of the nature of AORs, by investigating the sub-personal, mechanistic counterparts of the phenomenologically identified cases on which we have concentrated so far. A seminal example of sub-personal AORs is provided by Mataric's sonar-driven mobile robot Toto (Mataric, 1991). Toto wanders around its office environment following walls and avoiding obstacles. As it proceeds, it constructs an internal map encoded in terms of patterns of sensorimotor activity. For example, if Toto keeps detecting proximally located objects on its right-hand side, while its compass bearing remains unchanged, then a 'right-wall' is encoded in the sub-agential map, not as some objective entity, but in terms of the robot's sensorimotor 'experience' at the time. These structured sensorimotor 'experiences' (Toto's landmarks) are stored as connected nodes in a distributed graph, and this record of the robot's own sensorimotor history constitutes its sub-personal map of the spatial environment. Toto is then able to navigate its way around using paths encoded as sequences of past, current, and expected sensorimotor 'experiences'.

Toto's maps thus exhibit the range of inter-related properties that were previously disclosed by a phenomenological analysis of un-readiness-to-hand, and which are distinctive of AORs. They are action-specific, in that they are tailored to the job of producing the specific behaviour required. They are egocentric, in that they encode the environment in terms of the robot's own history of sensorimotor 'experiences'. And they are intrinsically context-dependent, in that because those maps are embedded in the kind of domain-dependent, task-specific mechanisms that we earlier identified as SPACs (meaning that they wouldn't be any good for working out things like the objective distance to the snack bar), Toto never confronts the problem of selecting, out of a vast sea of potentially available spatial information that could have been retrieved and internally stored, the subset of such data relevant to the navigational context in which it operates. Another way of explicating this final property is as the reason why Toto never confronts the frame problem, a reason which turns directly on the conceptual profile of AORs.

As a result of the operative profile just described, AORs are what might be called *minimally representational* in character. Rather than determinately specifying some detailed objective content (knowledge-that), AORs indicate an open-ended set of possible actions (a kind of under-specified knowledge-how), just like the detour sign that vaguely suggests the many directions that, given one's practical knowledge, one might take to drive home. Mention of the detour sign suggests a crucial observation. If we permit ourselves to interpret egocentricity more generally as a kind of perspectival deictic indication, Signs may be counted as AORs. As AORs, Signs indicate the different possibilities of coping that essentially define the current context or, via

the background, a range of potentially associated contexts. Signs are not present-at-hand representations, because their referential function is not action-neutral or context-independent. Rather, their function is essentially action-orienting and context-specific, warning us that something must be done in a certain situation (for example, ‘take the detour’). Given their minimally representational status, however, the precise form of this ‘something that must be done’ will be determined in detail by the trajectory of our competent engagement in the world.

A Sign, understood as an AOR, will determine a standpoint that is disclosed from within the background, but from which we relate to that background, allowing us to redirect our attention to access different depths or levels of coping, and thereby alter the relevance of various action-shaping preconditions. At any moment, some background conditions will matter more than others, and Signs move us from one perspectival angle to another, actively transforming the potential availability of different contexts of actions, and not merely the potential availability of alternative actions within the same context. Tongue twisters and other preparatory rituals are thus best interpreted as Signs that a performer, under the pressure of unfamiliar or problematic contingencies, deploys as a useful scaffold to change his dispositions towards the immediate contingencies, not only to better recognize the available affordances (he is already trained and ready to deal with the physical environment), but to establish a more appropriate perspective of self-interpretation in the light of a different configuration of his background. To see the environment as a familiar scenario for our actions requires recognizing it as compatible with certain background motivations, expectations and intentions that Signs prioritize in place of others that are debilitating or disruptive. We have seen that preparatory embodied routines, like Signs in general, enable us to manipulate the background preconditions and criteria that determine relevance, from within the current context. And since Signs are AORs, we can conclude that our capacity to navigate the background – our ground-level intelligence – is sometimes contingent upon the availability of minimal representations and is actively shaped by those structures.

The role that we have found for ground-level minimal representations might seem to contravene the principle of unity. The principle of unity would indeed be violated if the position we have sketched required the background to be divisible into encapsulated components, but the fact that the background can be articulated in different ways (that is, the fact that we can distinguish different modes or configurations of the global preconditions of one’s situated sense of reality) doesn’t imply encapsulation in any strict sense. If some forms of representation can in truth produce articulation without encapsulation, they will remain consistent with the holistic unity of the background. Our account keeps intact the unity of the background, by combining it with the idea that the background is accessed and manipulated perspectivally, from distinct angles. Each of these angles

indicates a certain global reconfiguration of the totality of the background, which is however never exhaustively represented by any of these indications. Indeed, if a part or portion of the background, configured in representational form, attempted to encode the totality of the background, this would lead to an infinite regress (like an encyclopaedia containing itself and its full description); but this regress is never established, if we assume that a local representational configuration of the background provides only an indication of some possible way of accessing the totality of the background (like an encyclopaedia containing the index of its own chapters and contents). AORs (including Signs and thus preparatory embodied routines) indicate how to inhabit the background strategically; they do not attempt to build a complete model of the background. If this is right, and if such representation-driven strategic inhabitings provide the basis for events such as fluid context-switching, then minimally representational processes are indeed at the heart of the ground-level intelligence.

5. *Angst* and the problem of relevance

As we have seen, Signs ordinarily prompt an agent to stand before this or that context of action. However, a particular experience that Heidegger calls *Angst* allows an agent to stand apart from the totality of all the possible contexts in which she can act, intuiting the extreme edge that globally delimits her possibilities as a thrown individual. “*Angst* provides the phenomenal basis for explicitly grasping the primordial totality of being of Da-sein” (Heidegger, 1996, p. 176). Despite this explicit grasping of the totality of what is intelligible, *Angst* remains a phenomenon that occurs to situated agents within the world, not as a de-worlded ‘view-from-nowhere’, but as an immanent mode of attunement within the world that discloses the world as world. Intriguingly, a link may be forged between *Angst* and the problem of relevance.

To bring this link into view, we can begin by noting that although *Angst* can be announced by Signs, Signs cannot actually indicate the object or objects of *Angst*. That is because *Angst* has no object, or at least no definite one. Imagine that, while the preoccupied George VI is preparing himself to give a speech, he glimpses the austere gaze of an ancestor in an old painting, or sees a blinking light on the microphone watching him like a malevolent red eye, or hears his own voice hesitating on a certain syllable. These Signs, harmless in themselves, may disrupt his attention and fill his heart with vague senses of discomfort, apprehension, and finally anxiety. The possibility of a complete failure looms, a possibility that shakes his confidence in his social skills, his political determination, and his historical mission. What this example indicates is that, in certain situations, even the most innocuous Signs can unearth such profound layers of the background, that the agent is shaken by deeper and deeper degrees of uncertainty. This uncertainty

impairs, rather than facilitates, his capability to recommence a fluid process of coping. In this spiral of growing anxiety, not only the agent's acquaintance with this or that situation, but his whole sense of reality, his sense of being-in-the-world, is eventually challenged by questions such as: 'why am I here?', 'what is the real meaning of my actions?', 'what am I supposed to do now?', and so on. Manifested by a feeling of uncanniness, *Angst* doesn't have a precise object but expresses a wrenching of the agent from her meaning-giving background. Heidegger (1996, p.175) describes this phenomenon as a 'not-being-at-home', the disorienting experience of being confronted by a meaningless world that appears alien and unintelligible.

Through *Angst*, we encounter the fact that all the possible ways things usually matter to us depend on our situated perspectives in thrownness. But since it is thrownness itself that is exposed by *Angst*, all our absorbed coping activities are revealed as devoid of intrinsic meaning. Those coping activities are now being observed from the perspective of our fundamentally precarious, ultimately ungrounded existential situation. As a consequence, even if all our competences are still in place, our ability to use them stops mattering to us. Our skilled coping becomes the 'merely occurrent' (Heidegger, 1996, p. 103) object of an anxious contemplation that disconnects affordances from the responses that they habitually summon in us. Thus 'the totality of relevance discovered within the world of things at hand and objectively present is completely without importance. It collapses. The world has the character of complete insignificance' (Heidegger, 1996, p. 174).

Crucially, the 'knowledge' gained through *Angst* never amounts in itself to a complete, fully representational depiction of the background. Rather, the alienation that *Angst* introduces engages a restless interrogation on the part of the agent regarding the meaning of things. The questions that he poses to himself about this meaning cannot find a theoretical answer, because any such answer would presuppose the world that *Angst* had made meaningless. So, even if that question originates fundamentally from the background of that meaning, the very fact that the question is formulated in *Angst* makes the meaning of the background indiscernible. Although it is clear that the knowledge of the background revealed by *Angst* is not fully representational, it seems plausible that it must be at least minimally representational, as it clearly produces a modulation of the background by adopting an explicit perspective within the background, assuming the dynamic co-moulding of cognitive distance and skilled action that is typical of AORs and the kind of troubled coping regime that may be introduced by Signs, although this attunement with the background is of a very strange, uncanny nature. Crucially, for Heidegger, the minimally representational phenomenon of *Angst* precedes – ontologically, if not chronologically – even the possibilities of our skilled coping. As he puts it: 'Tranquillized, familiar being-in-the-world is a mode of the uncanniness of Da-sein, not the other way around. Not-being-at-home

must be conceived existentially and ontologically as the more primordial phenomenon. And only because *Angst* always already latently determines being-in-the-world, can being-in-the-world as being together with the “world” take care of things and attuned, be afraid, (p. 177).

Conceived as a fundamental structure of human life, *Angst* provides a new perspective on the frame problem. Structurally destined to care, and constitutively exposed to the possibility of the total loss of meaning through *Angst*, human beings seem realistically open to something tantalizingly close in form to the inter-context frame problem that affects cognitivist AI. Seduced by the dreadful Signs that lead her to withdraw from her context, the agent experiences a more and more dramatic loss of the sense of mattering, a loss that paralyzes her responsiveness to the circumstances and makes her aware of the necessity to deliberate, even if no stable foundation for deliberation is available. The endless series of questions that the agent is driven to ask about her own being, as she endeavours to delineate contextual meaning (what matters) is analogous to the infinite regress of meaning-targeting structures that are indicative of the frame problem. The difference is that, in the case of humans, the *Angst*-driven ‘frame problem’ derives from a contingent suspension of the holistic coupling with the background that usually gives meaning to our actions, while for cognitivist AI systems this holistic direct coupling is structurally missing due to the system’s lack of thrownness (see above). And that is why, when the human agent forgets about the network of endless open questions in which it is trapped in *Angst*, it buries the fundamental sense of uncanniness under its habitual commerce with worldly things, and becomes ‘tranquillized’: it can then go back to its skilled coping with the world, re-activating its background coping, a possibility that is fundamentally unavailable to machines that lack thrownness.

6. Utilization behaviour revisited

As a bonus to the main business of this chapter, our analysis of ground-level intelligence has left us in a position to counter Rietveld’s arguments against the distinction between intra-context and inter-context sensitivity to relevance. In particular, as far as we can tell, the phenomenon of utilization behaviour doesn’t undermine this distinction, but rather demonstrates that the human capacity to produce appropriate actions is severely disrupted when the boundary in question collapses, allowing any local solicitation to acquire the role of a global precondition. The patient exhibiting utilization behaviour has not lost his sensitivity to the holistic meaning of his actions (he has a background, and he is still a thrown agent, in Heidegger’s sense); the problem is that this meaning is now totally absorbed by the contingent situation. He can still perfectly cope with beds to be made, and this means that he knows the normative preconditions of this action, but he can’t separate the contingent contextual solicitations from the holistic consideration

of the whole situation, as if the bed was the totality of the world with which he must deal.

Normally, the background of our actions is not given to us as an all-encompassing and all-determining universal structure, but is articulated perspectively within the local contexts in which it may be involved. Just as the veins in a not-yet-sculpted piece of marble suggested to Michelangelo the forms to be carved out from it, Signs delineate the different contextual articulations that a situated agent will produce by addressing pragmatically (that is, with a specific purpose in her mind) the unitary structure of her background. Indeed, Signs testify that the distinction between different contexts of action, as well as the very distinction between local contexts and the holistic background, is ontologically grounded, even though it is virtual (embedded in possibility). The appearance of Signs delineates the phenomenological boundaries separating different contexts of action, revealing whether or not it is possible to switch between them by adopting different perspectives for accessing the background. In cases where Signs announce a relevant transition between contexts, it means that genuine ontological boundaries exist within the background.

Things are very different for patients exhibiting utilization behaviour. For them, the background is entirely explicated by the actual focus of interest, such that no other contexts are possible. The dissolution of the separation between different contexts of action is demonstrated by the fact that the patient fails to experience the Signs that usually warn us to monitor our actions as soon as we cross the borders of appropriateness. The patient fails to interpret the surprised face of his host as a sign that his actions are not appropriate in her house, because the relevant contextual borders no longer exist for him. Such borders clearly exist in normal experience, which is why we immediately respond to the presence of Duchamp's famous fountain in a museum as out-of-context. From the twentieth century, the fine arts systematically employed the alienating experience of out-of-context objects and events in order to produce, in observers, an active reflection on the background preconditions of their own notion of a work of art. This reflection couldn't be achieved if the observers in question were not comparing the global background of the broader situation with the local factive circumstances in which it is perspectively disclosed. Signs help us to understand when we need to switch from one context to another to continue producing appropriate behaviour. But they also tell us when our actions are crossing contextual borders into a territory of the non-appropriateness.

What now of Rietveld's neurofunctional argument? Rietveld's overall position requires (on the philosophical level) that (1) sensitivity to background conditions is never informed by context-specific competences, and (on the neurofunctional level) that (2) sensitivity to background conditions is underpinned by brain mechanisms that globally tune the sensory systems, a process which is necessary for goal-specific motor structures (dedicated to

both action execution and the understanding of affordances) to underpin context-specific competences. This picture, according to Rietveld, is confirmed by the evidence that (3) the tuning of the sensory system in accordance with the agent's contextual expectations determines the conditions under which the contributions of context-specific motor competences are relevant. While we agree with (2), we have already argued that (1) represents both an inappropriately static model of the relation between background coping and skilled coping, and a transgression of the principle of mutuality. If background coping is a unitary process, then access to it is always perspectively oriented in accordance with the way the agent engages her local context, such that different context-specific competences can determine some of the background preconditions. For example, according to the premotor theory of attention (Rizzolatti and Craighero, 2010), the same frontal-parietal circuits that control ocular movements towards specific spatial locations actively reconfigure the attention towards the same locations in order to prime the detection of salient elements (Rizzolatti *et al.*, 1987). In an analogous way, the experimental data on canonical neurons show that the same pre-motor circuits that control the execution of certain goal-specific transitive actions are recruited by the process of detection and recognition of the objects that afford those actions (Rizzolatti and Luppino, 2001). These findings suggest that spatial attention and detection of the environmental features related to the relevant context of action do not result from a dedicated control mechanism, but from the same mechanisms (corresponding to SPACs, in our hypothesis) that control the execution of the appropriate actions in those contexts: 'as in the case of spatial attention eye movement preparation selects a given spatial location, the preparation of a grasping movement selects an object with specific intrinsic characteristics' (Rizzolatti and Craighero, 2010). But that means that (3) tells only one half of the neurofunctional story, because, in accordance with the principle of mutuality, the medial system that tunes our capability to pick up and interpret relevant sensory stimuli can be tuned by our motor engagement with the local context with which we are interacting.

7. Conclusions: the king's intelligence

The conclusions of our analysis may be expressed as a tri-dimensional pattern of agreement and disagreement with the account of ground-level intelligence that may be extracted from the work of Dreyfus and Rietveld.

First, we agree that the background is inhabited as a unitary set of holistic conditions, but this doesn't mean that the whole network of background preconditions is always equally involved in every kind of coping, because, for each form of skilled coping, some elements or modality of background coping may exclude others. The background is a vast web of significance that implicitly underpins the mattering of things during coping, but it is

a structure that we always inhabit from a situated perspective, and different profiles of the background matter in different degrees at different times. Were background knowledge always equally present in all its aspects, there would be no need for the disclosive function of Signs, structures that highlight the background preconditions that are most relevant in the current situation or that may potentially provide a bridge to other contexts of action.

Secondly, we agree that skilled coping, whether absorbed or minimally representational, is always underpinned by background coping, but, on the basis of the principle of mutuality, we recognize that the background can in turn be modified by ongoing, concrete acts of skilled coping. The background provides an ontological platform for our situated experience, but not as an immobile ground providing an ultimate, self-sufficient, autonomous foundation, like the set of basic axioms for a formal system of first-order logic. Rather, the background is a relatively stable scaffold that orients our everyday engagements within the world. Skilled coping continuously re-founds the background by dynamically modifying the normative preconditions it embodies, but that in turn provides the normative preconditions for further skilled coping. If it is correct to conceptualize the background as foundational at all, then, as Heidegger argues, it should be characterized as a bottomless abyss of preconditions, each of which requires others, according to the general schema of the hermeneutic circle.

Finally, we agree that background coping is not guided by full-blooded representations of a traditional kind, because the background can't be reduced to a body (however vast) of explicitly represented information, beliefs, or stored heuristics. That said, we have argued that our access to the background is often mediated and articulated by action-oriented representations, sometimes appearing in our experience as Signs. This means that, in the 'right' circumstances, the production of the wrong syllable by a stuttering speaker or an unexpected road sign appearing in front of the driver can prompt a global reconfiguration of the background. So, in the end, it is a minimally representationalist approach to intelligence that brings the background and its dynamics into proper view. The problem of relevance indicates that background coping cannot be understood on the cognitivist model, as a rational process of deliberation using full-blooded representations. But, as we have argued, neither can background coping be understood, or at least not exhaustively so, on a Dreyfusian model, as the unreflective and nonrepresentational selection of past contexts. In other words, at ground-level, the king's intelligence, like yours and like ours, is in part a matter of negotiating the dynamics of the background by way of action-oriented representations.

Acknowledgements

Some sections include short passages of text adapted from Wheeler (2010), and Wheeler and Di Paolo (forthcoming).

References

- Cappuccio, M. and M. Wheeler (2010) 'When the twain meet: Could the study of mind be a meeting of minds?' in J. Reynolds, J. Chase, J. Williams, and E. Mares (eds.), *Postanalytic and Metacontinental: Crossing Philosophical Divides* (London: Continuum).
- Dreyfus, H. L. (1992) *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press).
- Dreyfus, H. L. (2000) 'How Heidegger defends the possibility of a correspondence theory of truth with respect to the entities of natural science' in T. R. Schatzki, K. K. Cetina, and E. von Savigny (eds.) *The Practice Turn in Contemporary Theory* (New York: Routledge), 151–62.
- Dreyfus, H. L. (2002a) 'Intelligence without representation: Merleau-Ponty's critique of mental representation', *Phenomenology and the Cognitive Sciences*, 1, 367–83.
- Dreyfus, H. L. (2002b) 'Refocusing the question: Can there be skillful coping without propositional representations or brain representations?', *Phenomenology and the Cognitive Sciences*, 1(4): 413–25.
- Dreyfus, H. L. (2008) 'Why Heideggerian AI failed and how fixing it would require making it more Heideggerian', in P. Husbands, O. Holland, and M. Wheeler (eds) *The Mechanical Mind in History* (Cambridge, MA: MIT Press), 331–71.
- Freeman, W. (2000) *How Brains Make Up Their Minds* (New York: Columbia University Press).
- Gallagher, S. (2008) 'Are minimal representations still representations?', *International Journal of Philosophical Studies*, 16(3), 351–69.
- Heidegger, M. (1996) *Being and Time*, John Stambaugh (trans.) (Albany, NY: State University of New York Press). Originally published as (1927) *Sein und Zeit* (Tübingen: Max Niemeyer Verlag).
- Kelly, S. D. (2000) 'Grasping at straws: Motor intentionality and the cognitive science of skillful action' in M. Wrathall and J. Malpas (eds.), *Essays in Honor of Hubert Dreyfus*, Vol. II. (Cambridge, MA: MIT Press), 161–77.
- Kelly, S. D. (2002) 'Merleau-Ponty on the body: The logic of motor intentional activity', *Ratio-New Series* 15(4), 376–91.
- Kiverstein J. (forthcoming) 'Introduction', in J. Kiverstein and M. Wheeler (eds.) *Heidegger and Cognitive Science* (Basingstoke: Palgrave Macmillan).
- Lhermitte, F. (1986) 'Human autonomy and the frontal lobes. Part II: Patient behavior in complex and social situations: The «environmental dependency syndrome»', *Annals of Neurology*, 19, 335–43.
- Logue, M. and P. Conradi (2010) *The King's Speech* (London: Quercus).
- Mataric, M. (1991) 'Navigating with a rat brain: A neurobiologically inspired model for robot spatial representation' in J.-A. Meyer and S. Wilson (eds.) *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, (Cambridge, MA: MIT Press), 169–75.
- Merleau-Ponty M. (1962) *Phenomenology of Perception*, C. Smith (trans.) (London: Routledge).
- Rietveld, E. (2008) 'Situated normativity: The normative aspect of embodied cognition in unreflective action', *Mind*, 117/ 468, October.
- Rietveld E. (forthcoming) 'Context-switching and responsiveness to real relevance' in J. Kiverstein and M. Wheeler (eds.) *Heidegger and Cognitive Science* (Basingstoke: Palgrave Macmillan).
- Rizzolatti, G. and L. Craighero (2010) 'Premotor theory of attention', *Scholarpedia*, 5(1), 6311.

- Rizzolatti, G. and G. Luppino (2001) 'The cortical motor system', *Neuron*, 31, 889–901.
- Rizzolatti, G., L. L. Riggio, I. Dascola, and C. Umiltà (1987) 'Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention', *Neuropsychologia*, 25, 31–40.
- Sutton, J. (2007) 'Batting, habit, and memory: The embodied mind and the nature of skill', *Sport in Society*, 10(5), 763–86.
- Taylor, C. (1993) 'Engaged agency and background in Heidegger' in C. B. Guignon (ed.), *The Cambridge Companion to Heidegger* (New York: Cambridge University Press), 317–36.
- Wheeler, M. (2005) *Reconstructing the Cognitive World: The Next Step* (Cambridge, MA: MIT Press).
- Wheeler, M. (2008) 'Cognition in context: Phenomenology, situated robotics and the frame problem', *International Journal of Philosophical Studies*, 16(3), 323–49.
- Wheeler, M. (2010) 'The Problem of Representation' in S. Gallagher and D. Schmicking (eds.) *Handbook of Phenomenology and Cognitive Science* (Berlin: Springer), 319–36.
- Wheeler, M. and di Paolo E. (forthcoming) 'Existentialism and cognitive science' in J. Reynolds, A. Woodward, and F. Joseph (eds.) *Continuum Companion to Existentialism* (London: Continuum).

Support Publication #4 **[SP4]**

Cappuccio M.L., Froese T. (eds.) (2014),
Enactive Cognition at the Edge of Sense-Making. Making Sense of Non-Sense, Palgrave Macmillan, London.

Reproduced with permission from Palgrave (Springer Nature).
All rights reserved.

1

Introduction

Massimiliano Cappuccio and Tom Froese

Don't *for heaven's sake*, be afraid of talking nonsense!

Only don't fail to pay attention to your nonsense.

(Ludwig Wittgenstein, 1998, p. 64)

1.1 Wittgenstein's philosophical challenge: speaking about non-sense, without speaking non-sense

Ludwig Wittgenstein, in the attempt to take as far as possible a “naïve” correspondentialist and representationalist view on language (Hutto, 2004), had to face a resilient philosophical puzzle on the nature of nonsensical propositions (Wittgenstein, 1921/2001, 4.01): if our knowledge of the world is necessarily delimited by its representations, then what should we say about nonsensical expressions that represent nothing at all and that, nonetheless, allow us to relate to what is most valuable in our lives? If non-sense is situated outside the “limits of our language”, which bi-univocally correspond to the articulations of ontology (the “limits of our world”), then why is it the case that our words can capture the very notion of non-sense, allowing us to concur on its meaning? How is it possible that we know how to efficaciously respond to nonsensical expressions, and our beliefs seem affected by them in ways that are characteristic and significant?

Non-sense is imputable neither to the falsity or vagueness of representational content, as these flaws could well be compatible with the normativity of representationalism (inexact isomorphism does not undermine isomorphism as a criterion of truth), nor to mere absence of sense: according to Wittgenstein, our language is “senseless” (*sinnlos*) when it tries to represent the preconditions of its very semantic function,

and this can be disconcerting for a representationalist view of language, but there is nothing nonsensical in such a limitation. The propositions of logic – such as tautologies and contradictions – or mathematics, or the pictorial form of our pictures, have *no sense*, because they do not stand for any state of affairs in the world (Wittgenstein, 1921/2001, 4.0312). Senseless but significant, endowed with truth-values and an identifiable declarative form, these propositions “scaffold” the very possibility of contentful representation and, even if they do not represent anything within the boundaries of the world, they are all we have to gesture at these boundaries. That is why the very possibility of representing (“the logic of the facts”, *idem*) seems nonrepresentable: not only the illogical structure of ungrammatical expressions, but also the well-formed structure of meaningful expressions, are beyond possibility of representation.

Compared with the senseless, proper non-sense gestures at a more specific kind of non-representability, which is not just confined to the borders of our language/world, but ranges beyond them, displaying an ulterior domain (*ibid.*, 5.61). According to Wittgenstein, nonsensical propositions are a subcategory of the senseless ones; thus, his characterization of non-sense is narrower and even more puzzling. Truly nonsensical (*unsinnig*) propositions are not just devoid of clear denotational content, like the senseless ones: as they do not exhibit any recognizable logical form, they seem to transcend the very purpose of representation. Not only is their declarative content uncertain, but also the very fact that they have a declarative form, and that is why their truth conditions appear unknown and mysterious. “Pseudo-propositions” such as “Socrates is identical” (*ibid.*, 5.473), but also formal concepts like “There is only one zero”, “there are objects”, and “2+2 at 3 o’clock equals 4” (*ibid.*, 4.1272) are not just devoid of a clear denotational referent: their truth conditions – if they have any – are not even thinkable, as they gesture at “things that cannot be put into words”; but, adds Wittgenstein in a famous remark, “they make themselves manifest. They are what is mystical” (*ibid.*, 6.522), that is, they suggest a domain of significance that transcends positivistic representation of the world while having a pervasive presence in our life.

Surprisingly, whether such propositions are well-formed or not, their communicative intent can be intuitively efficacious in affecting our beliefs and motivating our actions: all the philosophical propositions of ethics and aesthetics (*ibid.*, 6.421) and traditional metaphysics belong to this group of utterances (*ibid.*, 6.53). Puzzlingly, most of our knowledge (and most of the *Tractatus* itself!) is made of such nonsensical

propositions (ibid., 4.0003). While the logic of their use is not rule-based, it compels us to regulate our life, and this idea will famously inspire the “language games” theme in Wittgenstein’s late production (Wittgenstein, 1953/2009, § 2, 23, 65; see Hutto, 2004, pp. 137–138). However, in the *Tractatus*, Wittgenstein is still attempting to account for the paradox of non-sense (i.e., that which is meaningless, but effective in our practices of knowledge) by contrasting two incomparable modes of signification, referred to as “what can be said” and “what can be shown” (Wittgenstein, 1953/2009, 4.1212, p. XI), which we could reinterpret as a *contentful* and a *directed* form of cognition, respectively (this distinction solicits various questions on the possibility of direct perceptual experience before and below conceptual categorization. Even if recontextualized in a different debate, these are some of the philosophical questions challenged by Beaton’s chapter in this book). In fact, if a merely representational approach to meaning might not account for all the possibilities of cognition, then another – non-representational – form must be possible and necessary: an immediate presentation of the existent, one that does not require internal mental states to be matched with external states of affairs, and for which “truth” does not indicate a norm of adequate correspondence.

1.2 Enactive theory and non-sense as challenges to the non-representationalist approaches to cognition

Wittgenstein treats non-sense as a paradox situated at the core of his philosophy of language, but his reflections are significant for a broader consideration of non-sense as a crucial experiential event that, once understood in its typical form and prelinguistic preconditions, can reveal an underlying embodied cognitive architecture. Interestingly, the same puzzles that affected Wittgenstein’s representational theory seem to be challenging today the non-representational theories of cognition and knowledge formation. During the last 20 years, the cognitivist approaches to cognitive science have been gradually supplanted by embodied-embedded and enactive approaches (e.g., Dreyfus, 2002a, 2002b; Gallagher, 2005; Varela, 1995; Thompson, 2007; Noë, 2009; Di Paolo et al., 2010). The latter, in opposition to the former, consider representation neither as a primitive explanatory element of intelligence nor as a constitutive ontological building block of the mind. Not to minimize this trend, but in recognition of its significance, this book wants to highlight that the problems concerning the nature of non-sense, unsolved by the representationalist cognitive science, need even more

urgently a solution by the non-representationalist one: the latter – as opposed to the former – cannot account for non-sense in terms of failed or impossible representation, as it does not recognize the same normative efficacy to the distinction between representational (well-formed) and non-representational (anomalous) forms of intelligence.

In fact, the enactive-embodied theory assumes that what was considered an anomaly by the tradition is actually the norm, as mental functions prevalently emerge from a background of practical engagement in which meaning is pure know-how learned through unprincipled interaction with the world. On this view, cognition is primarily a relational form of meaningful engagement, or *sense-making* (Weber and Varela, 2002). But what kind of process does regulate sense-making, then, if nonsensical, senseless, and fully meaningful experience all equally depend on an undifferentiated background of situated engagement? Or, to put it differently, if all cognition is situated sense-making, then how shall we account for cognition of the absurd? How should the enactive theories characterize non-sense, and the fact that it plays a major role in our cognitive life through abstract and symbolic concepts? The problem is not only that enactivism cannot rely on wrong, missing, or “blank” representations to differentiate between meaningful and nonsensical events, but also that it often seems to implicitly assume that every directed form of practical engagement with the world not only *can* but – to some extent – must be inherently productive of sense (Froese, 2012). But is it correct to assume that the cognitive horizon described by enactivism is entirely saturated with sense?

To illustrate this problem, we must introduce the non-representational embedded-situated theories of sense-making, which interpret adaptive intelligence as an immediate organismic responsiveness to the relevance of contextual contingencies; in particular, according to the enactive approach to cognition, all organisms are structurally predisposed to make sense of their world-environment in terms of opportunities of perception and movement, whose reciprocal co-implication is established during interaction with real-life circumstances. These opportunities are disclosed either to individual agents or, importantly, in a participatory way, when their social interaction brings about coordinated forms of perception and joint awareness (De Jaegher and Di Paolo, 2007; Fuchs and De Jaegher, 2009; Torrance and Froese, 2011). Interestingly, not only is the world of the agents defined by their coordinated social interaction, but the identity of the agents themselves, with their specific features, including, importantly, gender and gender perception (see Merritt’s chapter in this book for a circumstantiated analysis). Intelligence (including social

intelligence) does not find its foundation in explicit decisional processes mediated by internal models, supra-modal representations of an objective external world, stored heuristics, declarative contents, or formal inferences; on the contrary, it emerges from the unprincipled adaptation of a living system whose intrinsic pragmatic dispositions and narrative habits progressively adjust to reach a structural attunement with the environment, tending towards a dynamical balance with its contingent fluctuations (Gallagher, 2008a; Gallagher and Hutto, 2008). This notion of intelligence as direct know-how opposes the intellectualist biases of classical cognitivism (Hutto and Myin, 2013), modeled in the image and likeness of exquisitely human practices of knowledge, preferably based on literacy proficiency and culture-specific aptitude in formal thought. Enactivism challenges the anthropocentrism implied by this view (see in particular Leavens' chapter in this book for a systematic argument against such an anthropocentric approach to cognition); it remarks that these forms of cognition constitute the explanandum in the scientific discourse on the evolution of intelligence, not the explanans, and claims that a better awareness of the biologically and historically situated foundations of cognition is needed. Paradigmatically, enactivism revolves around animal processes of meaning formation that at root are organism-centered, environment-specific, and goal-directed, that is, incarnated into dynamic relationships rather than stored informational contents, shaped by material contingencies rather than computational algorithms (Thompson, 2007, 2011).

1.3 Sense-making as adaptive coupling between living body and world-environment

Inspired by a well-established tradition in bio-semiotics (von Uexküll, 1934/1957; Barbieri, 2006) and phenomenology (Merleau-Ponty, 1945/1962), the enactive approach to cognition (originally introduced to describe the role of embodied action in perception, see Varela, Thompson, and Rosch, 1991) assumes that the distinctive characteristic of a cognitive system is its continuous engagement in the active constitution of a meaningful "world-environment" (*Umwelt*). The world-environment is essentially irreducible to the natural world described by the empirical sciences (the scientist himself, while running his experiments, is a cognitive system that enacts a peculiar world-environment, with its peculiar experimental truth, as Bitbol's chapter in this book reminds us). The system intimately belongs to its world-environment and always has a situated, teleologically oriented perspective on it (Di Paolo, 2005);

perspectivism, in turn, is an ineliminable, constitutive feature of the world-environment itself. Autonomously orienting itself towards what is relevant to its subsistence, the organism actively adjusts the anticipatory trajectories of its behavior to dynamically regulate its conduct in the fuzzy, risky scenarios of real-life. This regulation can legitimately be called intelligent because it does not mechanically react to limited sets of occurring stimuli on the basis of the statistical repetition of previous experiences, but also flexibly prioritizes between novel contingencies based on their contextual relevance for the survival of the organism, anticipating the incoming changes (various philosophers characterized this pre-reflective adaptation in terms of “motor-intentionality” guided by a situated normativity, rather than representations of rules and stored heuristics; see, e.g., Merleau-Ponty, 1945/1962; Kelly, 2000, 2002; Dreyfus, 2002a; Rietveld, 2008, 2012). This rich coupling with the environment, at once compensatory and anticipatory, exploratory and balance-seeking, is structured in accord with a protentive/retentive temporal dynamic (Varela, 1999) and is essentially realized by the sedimentation of habitual responses that are progressively refined, either in ontogenesis or in phylogenesis, through a history of successful interactions.

Enactivism stresses also that perceptual engagement and motor expertise are pre-eminent defining factors for the consolidation of the cognitive self (Thompson, 2005), as the interactive exploration of the peripersonal space (through either manipulation or navigation) is the primal situated experience for learning how to map the transformative meaning of one’s own actions into the opportunities offered by the perceived environmental affordances; in turn, it also emphasizes that the perceptual sensitivity to relevance is embodied in the physical constitution of the organism, either as passive filters realized by its morphology, or as anticipatory mechanisms relying on the formation of sensorimotor feedback loops (O’Regan and Noë, 2001; Noë, 2004). Before and below explicit judgment, sense-making depends on the interactive habits matured by the system that establishes a dynamic coupling with its world: this reciprocal belonging of living body and world-environment is the defining, nonmetaphoric underpinning of cognition itself, so that living and cognizing are modes of the same sense-making capability and therefore are, in their essences, coextensive (Thompson, 2004). Materially extended over the dynamic interplay between nervous system, extra-cranial body, and extra-bodily environment (Virgo et al., 2011), the sense-making processes of an organism constitute its world as a significant context of action and experience, the bottomless backdrop

of implicit and nonrepresentable practical meanings that at once are informed by and inform the patterns of preferred intervention, the living system's intrinsic sensitivity to relevant stimuli, and the characteristic emotional tones that globally modulate the ongoing body-environment adjustments (Colombetti, 2014).

Unsurprisingly, the enactivist concept of sense-making builds on a general constructivist epistemology of the living organisms situated in their specific niche. Enaction theory represents the cognitive-psychological complement of the paradigm of autopoiesis in theoretical biology, a key descriptive and explanatory doctrine that understands the living being as a distinct type of homeostatic system whose peculiarity consists in counterbalancing environmental fluctuations to build and preserve its own internal functional organization (Varela et al., 1974; Maturana and Varela, 1980, 1987; Maturana, 2002). This compensatory process is an always-precarious negotiation that tends towards a relative stability through the constitution of self-organizing hierarchies of transient structures. The supervenience of self-organization over the component processes of the system defines the virtual boundaries between the internal and the external world of the organism, and endows the internal domain of the living being's systemic processes with a distinctive identity (but territorial spatial concepts cannot account for the dialectic process that defines this identity: see Stewart, this book, for a deeper contextualization of this notion of identity). Autopoietic systems are defined as autonomous because, in spite of their precarious negotiation with the external fluctuations and the transitory nature of their constitutive processes, they actively preserve a stable organization that distinguishes them from the world in which they are situated. Their phenomenology is governed by the principle of operational closure, which asserts that the transitions occurring within the organism can find a functional characterization only in reason of the specific horizon of organism-centered meanings that it enacts: this is the holistic backdrop against which single processes and events make sense to a particular living being, the immanent hermeneutic precondition of its sense-making capability (Stewart, 2000).

The precariousness of the living condition is not merely a contingent aspect of its realization, but essential for the enaction of meaning: without the ever-present possibility of the cessation of life there could be no sense of concern for life. The potential end of all sense-making is at the same time a necessary condition of possibility of all sense-making (Jonas, 1992). The complexity and the variety of the meanings of an organism's world-environment depend on the level of organization

of the autonomous systems considered (single-celled and multicellular organisms, and large societies of multicellular organisms, can all be described as autopoietic entities, and the nervous system itself is a system governed by operational closure; Varela, 1991). Even if some theorists suggest distinguishing enaction theory from the classical autopoietic paradigm (e.g., Froese and Stewart, 2010), the naturalistic approach based on their combination, committed to mechanistic explanatory models, phenomenology of first-person experience, and procedures of empirical validation, has produced groundbreaking theoretical models that deeply changed our way of describing the relationship between cognition and life (Froese and Stewart, 2012). If we take life to be autopoiesis plus adaptivity, then living is sense-making, and cognition is a form of sense-making (Thompson, 2011). On this view, there is no mind without life and no life without mind; there is no such thing as philosophical zombies or brains-in-the-vat (Hanna and Thompson, 2003; Cosmelli and Thompson, 2010). The main advantage of this approach, which interprets cognition as a sense-making capability of the living, is that it accounts for the radical context-sensitivity and rich adaptivity of intelligence, allowing a deep appreciation of its embodied and situated dimension against the intellectualist narratives tailored to higher forms of intelligence.

1.4 The puzzle of non-sense and the phenomenology of the uncanny

In spite of the originality of the enactive paradigm, its philosophical foundations are still haunted by the specter of Wittgenstein's puzzling question: if cognition is essentially a process of sense-making, then how does the enactive approach account for non-sense? Is non-sense characterized by any specific form or content, or is it, rather, the lack of these that makes the occurrence of the nonsensical recognizable? What does exactly happen in the mind of a cognitive agent when he becomes aware of facing a nonsensical experience? These are the key questions that this book wants to address and articulate analytically, as they pressure the unitary paradigm of cognition as adaptive coupling, reaching deeply into the theory of life as a sense-making system. In order to present them, this introduction will initially build on the phenomenology of the *uncanny* as the prototypical form of nonsensical experience; subsequently it will discuss the theoretical options available to model its underlying mechanisms, under the assumption that first-person descriptive analysis can clarify the general architecture of cognition. Phenomenology, in fact,

is the systematic and methodologically controlled examination of the conscious, qualitatively irreducible experience lived by a situated agent, aiming at highlighting the internal regularities and typical patterns of her intentional life (Gallagher and Zahavi, 2008). As phenomenology is the philosophical tradition that has exerted the deepest influence on the foundation of enactivism and embodied-embedded cognition (Varela, 1996; Rudrauf et al., 2003; Lutz and Thompson, 2003; Thompson, 2005), it is useful to refer to the existentialist phenomenology that first explored the intervention of the uncanny in everyday life.

In his most prominent philosophical treatise, *Being and Time*, dedicated to the situatedness of human life and the background of its meaning, Martin Heidegger famously captures the essence of non-sense with one of the categories of his existential analysis, the “uncanny” (*Das Unheimliche*; Heidegger, 1927/1966, § 40): a sense of disconnection experienced by a conscious intentional agent, indicated as *Dasein* (usually translated as “being-in-the-world”), when the familiar sense of its typical world-environment is lost. As the world of the cognitive agent corresponds, in its phenomenological constitution, to the ego-centered structure of its own perceptual field, falling outside the perimeter of its familiar environment inevitably coincides with losing the coordinates of its embodied self. Heidegger’s phenomenology backs the embodied-embedded and enactive models of cognition (Wheeler, 2005; Dreyfus, 2007; Kiverstein and Wheeler, 2012; Wheeler and Di Paolo, 2011), in that for both of them the concept of familiarity offers the implicit norm of a competent engagement with typical circumstances. This norm characterizes the subject’s readiness to practically deal with the nonrepresentable background of precategorical meanings (being-in-the-world) that scaffold, silently guide, and fill with experiential concreteness her activities (Taylor, 1993). This pre-reflective, fluid, expert mode of interaction is called “ready-to-hand” by Heidegger (*Zuhandenheit*; see Heidegger, 1927/1966, § 15) and “absorbed coping” by Dreyfus (2002a).

The most radical events of suspension of familiarity are accompanied by a disturbing atmosphere of alertness and detachment that Heidegger (1927/1966, § 40) dubs “anxiety” (*Angst*). It resembles ordinary psychological affects such as fear, estrangement, and surprise, but it is both deeper and vaguer: it is not directed towards any particular intentional object, but towards a ubiquitous menace, which solicits a hypersensitive attentiveness that apprehensively interrogates *Dasein*’s familiar world, bringing to light its implicit and nonthematic features, reorienting *Dasein*’s concern from the foreground to the background of the very structure of its consciousness. An anxious agent perplexedly

scrutinizes not the items of his ordinary experience, but the preconditions of the intentional relation he entertains with them, as he is sensitive to the possible instability and arbitrariness of this very relation. The by-product of this attitude is that it unintentionally reifies the totality of his perceived world, which now stands before him as a brute fact, a neutral collection of inert objects devoid of intrinsic purpose or significance. This anxious contemplation suspends the typical relationship of coupling (“attunement”) with the world in which the subject had been pre-reflectively embedded (Ratcliffe, 2002): when tinted by anxiety, the facts of the world, while structurally unchanged, suddenly lose their characteristic practical or affective value, as they stop evoking in the agent its typical responses. For Heidegger (1927/1966, p. 189), this neutralization of the expected affordances is alienating but revelatory, because it shows (by interrupting it) the subject’s uncritical and unquestioned (“tranquilized”) absorption in his routine relationship with the familiar (“being-at-home”), and solicits an estranged examination of the habitual subject–world coupling (“not-being-at-home”). This nomadic exploration is existentially more fundamental and authentic than being settled into a given frame of established customs and procedures, as the disenchanting wandering through the desert of bare existence – which entails the loss of the center and the examination of the boundaries of our world – discloses a harsher awareness of one’s own background, breaking any conformist compliance with everyday routine. This theme is later further emphasized by Sartre in both his essays (1946/2007) and novels (1938/2000). That is why anxiety, while annihilating our usual coupling with the things, may turn it into an epistemological opportunity of awareness and presence. Developing a Nietzschean motif, we could associate the uncanny with a convalescence of sense: the recovery from the despairing fevers of non-sense promises a genuine rediscovery of what had always been unperceived because implicitly taken-for-granted, submerged by the obviousness of the familiar, of the normal, of the healthy.

1.5 Pathological anxiety and the un-ready-to-hand

That is why deviant and even morbid instances of non-sense can shed an insightful light on the ordinary constitution of sense-making. Through the systematic study of depressive and schizophrenic patients, the phenomenologically informed approaches to psychopathology and psychotherapy have extensively examined the alienating and paralyzing effects of the uncanny in the clinical expressions of

anxiety (Jaspers, 1959/1997; Binswanger, 1959). These effects occur when the patient is compelled to look at his absorbed relationship with the world from outside himself, as if the centers of his agency and ownership were shifted beyond the stratified horizon of expert habits that he typically embodies (Fuchs, 2005). Interactive engagement with particular activities and contexts becomes unwieldy as the latter stand before the subject as an insignificant fact, a puzzling problem that concerns his own existential condition, prompting open-ended questions on the justification and the destination of his project of life (Froese, Stanghellini, and Bertelli, 2013). The imaginative elaboration on the nonsensicality of one's own existence can override the perception of one's reality, motivating delusional narratives that follow excruciatingly rigid and obsessively repetitive self-deconstructive patterns: the uncanny exposure of the background of sense, which had always been implicit and silently assumed before anxiety, turns into a painful admission of its lack of foundation, disclosing its ultimate meaninglessness.

The cognitive trajectory of this process is what is most relevant in our discussion. If non-sense emerges from anxiety, this is not because the fluid stream of habitual coping with the world had been overlooked, forgotten, or impaired, but because it was objectified under the focus of hyperreflective consideration, turning into a petrified body of factual information virtually separated from its cognizer (for a rich analysis of this alienation, see Dibitonto's chapter in this book). Non-sense originates from anxiety because the agent cannot re-enact the spontaneity of his habits and routines when he reflectively stands before their presence, while painfully trying to rationalize them as mere contemplative facts, without engagement or participation. Importantly, the nonsensical atmosphere of the uncanny is not simply produced by a malfunction of our adaptive skills, but by their efficacious exaptation under a new regime of decoupled, nonengaged, and hyperreflective assessment, in which the agent-world coupling has lost its spontaneity. The dramatic alienation that follows from pathological anxiety confirms by subtraction that it is by means of a continuous process of embodied and situated sense-making, as opposite to representation, that a typical cognitive system constitutes his familiar world-environment. But it also bears testimony that the breakdown associated with hyperreflection, through detached representation, can authentically be productive of sense, rather than merely destructive, because – if nothing else – it confronts the cognitive agent with the disturbing, but revealing, contemplation of the lack of objective foundation of cognition. Indeed, the groundlessness of mind

and cognition was one of the original motivations for the enactive approach (Varela et al., 1991).

Anxiety is not the only possible neutralization of the subject–world coupling; it is just the most dramatic and disruptive form, involving the totality of one’s existence. Existential analysis describes lesser, non-pathological forms of loss of familiarity in everyday life. Heidegger (1927/1966, § 16) describes the “un-ready-to-hand” (*Unzuhandenheit*), that is, the moment of unexpected breakdown of the expert ready-to-hand interactions. This event prompts a troubled, uncertain modality of practical engagement that (as confirmed experimentally by Dotov et al. 2010) occurs when the pre-reflective flux of interactions with familiar worldly circumstances is interrupted because of the malfunction, unreachability, or obstructiveness of tools, producing challenging practical circumstances that force us to critically stand back from the context and reflectively develop a novel problem-solving strategy (see Dotov and Chemero, this book). The un-ready-to-hand is the in-context, task-specific analogue of the world-involving, boundless experience of anxious uncanny.

This experience, according to Heidegger, also has revelatory implications, but, as opposed to the uncanny, these are local rather than global: the required practically oriented depictions of the ongoing situation can become manifest so as to plan appropriate decisions, indicating substitutive or emendatory directions of intervention. According to Wheeler (2005), “action-oriented representations” are the kind of cognitive devices that, extended over the brain and the environment, and encompassing both subpersonal mechanisms and personal-level intentional acts, realize this function. They enable a new critical scrutiny of the practical context of activity in which the subject operates, soliciting her to explicitly depict the content of her cognitive acts. While adaptive skills operate in a pre-representational non-reflective mode during routine engagements with familiar contexts, the tasks carried out in a mode of un-readiness-to-hand require aware self-monitoring and deliberation, mediated by explicit knowledge expressed in a declarative and representational format. Anxiety is similar to the un-ready-to-hand in that both of them disclose the brute factuality of the context in front of the subject, prompting a concerned interrogation on the preconditions of the expert subject–world coupling, and asking the subject to establish what norms govern that context of activity. Deliberation suspends the unreflective absorption in the automatic flux of expert activity in both anxiety and un-ready-to-hand. However, while anxiety reaches into a bottomless abyss of familiarity that is the existential situation of the

subject, unveiling its groundless, limitless backdrop, the un-ready-to-hand discloses through a clear-cut representation the isolated contexts of familiarity and the dedicated adaptive competences coupled with them.

According to Wheeler (2005, 2008), whose analyses are largely in line with Dreyfus (1991, pp. 69–85), the un-ready-to-hand is the moment in which enactive skills, facing their contingent failure as purely pre-reflective dispositions, evolve into a more reflective and detached form of intelligence, soliciting the development of proto-representational or minimally representational modes of cognition. Cappuccio and Wheeler (2012) further elaborate on this idea, suggesting that symbolic culture is based on the production of action-oriented representations during social practices (e.g., joint attention through index finger). Enactive and interactionist approaches to social cognition are inclined to reject this minimally representationist model, pointing out either that action-oriented representations reintroduce by the back door some outdated cognitivist assumptions (Hutto, 2013), or that, in spite of their name, they are actually not representations at all (Gallagher, 2008b). On the other hand, it has been argued that only such a minimally representationalist line of speculation allows us to appreciate how the most sophisticated forms of joint attention necessarily involve a public awareness of the discontinuity in the communal engagement of a plurality of subjects (Cappuccio and Shepherd, 2013; Cappuccio, 2013). Therefore, action-oriented representations constitute a theoretical option that enactivism may or may not want to follow.

If the line of argument based on action-oriented representations is both phenomenologically and empirically correct, then the enactivist picture would need to be complemented by contents that are only minimally representational: this means that, in contrast to the representations theorized by cognitivism, action-oriented representations are not necessary building constituents of our intelligent skills, but tentative scaffolds of precarious practical activities, which contingently emerge against a background of failed adaptive skills when these skills are reused to face unfamiliar cognitive tasks. This model would match the phenomenological evidence that, in an un-ready-to-hand activity, non-sense is produced by self-monitoring that does not paralyze action, but discloses entirely new opportunities of action-planning through minimal forms of representation that are context-specific, goal-oriented, transitory, and perspectival. This practical modality of representational intelligence compensates for the discontinuities in the usual sensorimotor engagement and administers the economy of the delays in the sense-making processes (Clark, 2006).

Importantly, it is also creative and productive of new layers of sense, not only because it facilitates mindful problem-solving and decisional processes, but especially because it indicates how to modify one's own subjective perspective on a certain context of action. This modification is instrumentally obtained through the manipulation of the representation of the context of action, which offers a distinct normative content to evoke the psychological scenario for one's performances (Cappuccio and Wheeler, 2012). In fact, the Heideggerian notion of "signs" (Heidegger, 1927/1996, p. 71) essentially captures the capability of action-oriented representations to identify and stand in for entire contexts of practical use, explicitly highlighting the rules of the context in which we are situated. They can become vicarious means to manipulate information on these contexts in their absence, or to switch between different contexts guided by their temporary and partial representations, allowing an agent to actively deliberate on how to intervene in them and on them. Action-oriented representations offer the most basic way to deal with everyday non-sense in a detached fashion, transforming unfamiliarity into a cognitive resource necessary to rationalize risky or uncertain decisions.

Insofar as they mediate between the sense and the non-sense of direct perceptions, manipulating the awareness of the entire context of actions, action-oriented representations might help explain the cognitive precursors of symbolic practices. In fact, as Heideggerian signs evoke entire contexts of activities, symbolic content is not conveyed to solicit action, but to reflectively appreciate the context in which this action is possible. The non-sense manipulated by action-oriented representations probably plays a role in the origin of symbolic intelligence, as it interferes with our immediate sense-making dispositions, and, if symbolic experience is really based on the suspension of these dispositions, then an advanced cognitive system dedicated to the manipulation of non-sense might actually play a major role in explaining the early emergence of symbolic practices. This system would be required to create a delay between the percepts and their associated responsive dispositions by neutralizing the direct responsiveness to perceived affordances: this delay could be the precondition to establishing a supplementary sense-making system that evokes inhibited actions to refer to distal, absent, or virtual contents associated with them, scaffolding the advent of the human faculty of imagination.

Or else, enaction theorists might well decide to stick to their dispositionalism and try to account for the specificity and the origin of symbolic practices without involving any kind of minimal representations. Froese (2013), for example, speculates that prehistoric rituals, such as initiation

rites and other rites of passage that involve a bracketing and breakdown of habitual behavior, may have originally served as social aids for the enculturation of more symbolic modes of cognition. However, following Gallagher (2008b), Hutto (2013), and others, it is interesting to see whether this idea can be developed without appealing to subpersonal representations. Indeed, the general challenge of Varela's enactive approach is to account for specifically human forms of detached cognition, while resisting the temptation to fall back on representationalism (the chapter by Short, Shearing, and Welchman in this book deepens this point, clarifying how the enactive genesis of symbolic practices is rooted in an original coupling, rather than in a representational function). This non-representationalist account may be achieved in various ways (see review by Froese, 2012). One interesting possibility is to consider the effects of altered states of consciousness (as systematically done by González, this book). For example, the self-sustaining neural dynamics that are unleashed during certain kinds of altered states may have been involved in the prehistoric origins of more abstract cognition and imagination, because they can lead to hallucinations of geometric forms that are imbued with significance, while at the same time partially decoupling the brain from outside influences (Froese et al. 2013).

1.6 Other practices at the edge of non-sense: humor and surrealist arts

Non-pathological experiences of non-sense are continuously encountered in many ordinary contexts. Because sense-making is not a deterministic process, but the dialectic outcome of uncertain negotiation between the world-environment and an autopoietic self, the uncanny manifestations of non-sense accompany life as an open possibility that is ready to be triggered at any time. Often confined to the safe terrain of pure imagination, simulation, and pretense, the revelatory advent of non-sense does not have to be psychologically disruptive. One hypothesis that this book wants to explore is that the intrusion of non-sense into the otherwise saturated horizon of our everyday sense-making practices allows the production and the fruition of particular practices such as sense of humor, surrealist art, and Zen meditation, informing their respective cognitive structures. A stronger hypothesis is that non-sense is involved in any kind of symbolic medium that creates critical distance from the direct affects engendered by our perceptual and motoric environment, to appreciate an imaginary world of absent, fictional, or virtual entities.

Support Publication #5 [SP5]

Cappuccio M.L. (2017),
“Flow, Choke, Skill: the role of the non-conscious in sport performance”
in Zdravko Radman (ed.), *Before Consciousness. In Search of the Fundamentals of
Mind*, Imprint.

Reproduced with permission from Imprint Academic Publisher. All rights reserved.

Massimiliano Cappuccio

Flow, Choke, Skill

The Role of the Non-Conscious in Sport Performance

1. Embodiment and Background: Philosophy and Psychology on the Sporting Field

Do expert athletes rely preferably on conscious or non-conscious¹ mental processes in order to peak their sport performances? Today, this question is passionately debated by sport psychologists and cognitive scientists. Given the elusive nature of the notion of consciousness, it could hardly be answered without simultaneously addressing some interrogatives that are exquisitely normative in nature or that importantly involve philosophical clarification: what truly is consciousness, and what does make an experience conscious? What is the relationship between conscious and non-conscious behaviours? Providing a definitive answer to these interrogatives is well beyond the scope of this chapter, which only aspires to concretely illustrate how the question of peak performances and the question of the nature of consciousness are intertwined and need to be addressed together. That is why this chapter will not simply assume that the philosophical clarification of the notion of consciousness is a prerequisite to understand peak

¹ In describing non-conscious psychological processes, I will assume a classical tripartition. All of the cognitive processes and information that are accessed by the system at the subpersonal level, without them explicitly informing phenomenal experience (as they are outside the field of one's conscious awareness/attention), are called *non-conscious*. The domain of the non-conscious comprises two subsets: *subconscious* dynamics (cognitive processes and bodies of information that currently are not, but that in principle can be and could have been, consciously accessed at the personal level, if the cognitive agent directed its attention to them); and *unconscious* dynamics (cognitive processes and bodies of information that are not, and that in principle cannot be, accessed by consciousness, as structural limitations or intrinsic incompatibilities in the cognitive architecture prevent the cognitive agent from consciously accessing them at the personal level).

performance; it will also show that, in turn, an empirical understanding of the psychological conditions that enable peak performance can significantly enrich our philosophical understanding of consciousness. In order to do this, it will examine the cognitive processes and psychological states that allow professional sportsmen and sportswomen to excel on the field.

Some philosophers might be tempted to ask what is authentically philosophical, if anything, in an investigation on the cognitive architecture underlying the conscious experience of expert athletes. Philosophical inquiry has often tended to investigate consciousness exclusively as a metaphysical category, a fundamental property or principle identified by the conditions of its actual or counterfactual existence within the logic of possible worlds, conditions that in turn are to be inferred through *a priori* reasoning (for example: Lewis, 1980; Searle, 1995; Chalmers, 1996). However, another genuinely philosophical approach is possible, one that investigates the experiential nature of consciousness in real-life scenarios, exploring how situated cognitive agents interact with their world-environments, and how the patterns of this interaction concretely shape the modes of their awareness (Zahavi, 2005; Gallagher and Zahavi, 2008; Marbach, 2010; Thompson, 2014). The reflection offered by this chapter is inspired by the general terms of such an approach: instead of *assuming* consciousness as the mysterious fulcrum of transcendent forces or as an illusion to be dispelled, the goal is to *discover* consciousness as a mundane phenomenon coherent with scientific accounts of the mind, with the naturalistic view that the human psyche is immanent and incarnated, and with the evidence that mental activity is intrinsically intentional, relational, and social.

This approach presupposes a deep entanglement between one's cognitive performances and the normative dimension of one's conscious experience, and searches for a systematic correspondence between the phenomenological and the functional dimension of the mind. This correspondence appears with particularly compelling evidence when we observe the negative modulations of one's consciousness, i.e. the distinctive experiences of unawareness or distraction that, in a specific range of situations, predict a decrease (or, more surprisingly, as we shall see, an increase) of one's cognitive abilities. Real-life cases (in particular, those related to the challenges faced by professional athletes) show that manipulating one's consciousness state modifies her cognitive performances too. This suggests that between phenomenal consciousness and mental function there is actually deep causal codependence, and not only extrinsic correspondence.

Why is the psychology of sportsmen and sportswomen particularly significant for a general inquiry into consciousness? In so far as our

cognitive systems are most heavily tested by challenging sport competitions and demanding psychophysical training, sport and performance psychology—the scientific discipline that studies the cognitions, the emotions, and the motivations of athletes—can tell us a few important things about the fundamental architecture of the mind, and especially about the power of its non-conscious components. Sport became a strikingly fecund domain of inquiry for both philosophical psychologists and cognitive scientists: a privileged perspective to study the limits and the potential of the mind ‘in action’ is offered by the empirical examination of the feats of ability accomplished by sportsmen and sportswomen. During these performances, humans are required to make the most out of cognitive faculties like perception, motor control, action planning, and decision in complex, possibly fast changing, real-life scenarios. Sport performance allows us to observe these cognitive faculties under a magnifying lens because it exposes them when they are both extraordinarily developed, nearing the edge of human perfection, and extraordinarily under stress, nearing the edge of exhaustion and collapse. Cognitive science can use the lens provided by sport psychology to infer the nature of the mental resources needed by well-trained humans to accomplish their superior feats of ability, but its goal is to generalize these observations in order to model the fundamental normative dimensions that define the skilful abilities of all humans.

It is remarkable that the psychologists and the cognitive scientists who work on sport seem naturally inclined towards embodied cognition theory and tend to embrace at least its most general and fundamental tenet (Varela, Thompson and Rosch, 1991; and Wilson and Foglia, 2011): that the processes involving body and environment importantly contribute to shape and scaffold higher forms of cognition. More specifically, this means that perception and action, considered as dynamic forms of coupling with the environment, participate in judgment and decision no less than the information internally stored by or produced in isolation within the central nervous system (see Beilock, 2008; Holt and Beilock, 2006; Beilock and Holt, 2007; Gray, 2014). These tenets resonate deeply with the ecological doctrines that distinctively contributed to the theoretical foundation of sport psychology (Gibson, 1979).

In fact, the performances of expert athletes, in revealing the deep integration and mutual presupposition of perceptual experience, action, and adaptive intelligence, illustrate in the most compelling way the embodied and situated dimension of the cognitive mechanisms that govern skill learning and ability. With regard to the theme of this chapter, referring to the embodied view of cognition is somehow

inevitable for an inquiry into the powers of the non-conscious mind, not only by reason of the huge influence that embodied theory exerted on philosophy of mind and cognitive science during the past thirty years, but also because this perspective has inspired the most influential philosophical approaches to skill and performance, some of which are characterized by a frankly anti-intellectualist (Fridland, 2014) and anti-representationist stance (Dreyfus, 2002; Hutto and Sanchez-Garcia, 2007).

The characterization of skill and expertise provided by the first theorists of embodiment (Dreyfus and Dreyfus, 1980) was in substantial accord with the most classical works in sport psychology (Fitts and Posner 1967): experts paradigmatically deliver their skilful performances in the form of unreflective action (i.e. through automatic routines), and expertise itself can be developed only when complex action patterns are automatized (i.e. when they become less demanding in terms of cognitive control). It is only when conscious control and explicit decision are replaced by well-trained habitual responses that the most effective and fastest results become possible. It is in these circumstances that, according to the aforementioned theorists of embodiment, the body shows its own intelligence, a skilfulness that is more primordial than intellectually and verbally educated intelligence: it is when a pre-reflective and prenoetic intelligence takes over that the body shows its inherent capability to effectively attune and respond to worldly contingencies without relying on rules of thumb, conceptual models, rational inferences, explicit instructions, or intellectual representations. Importantly, sport psychologists have ascertained that automated behaviours, despite being produced without explicit control or reflection, are nonetheless adaptive, i.e. richly flexible, dynamic, and not stereotyped (e.g. Beilock *et al.*, 2002a; 2004; 2008; Beilock and Gonso, 2008). Correspondingly, various philosophers—interested in the notions of ‘habit’ or ‘pre-reflective intelligence’—have characterized skill as an intelligent form of unaware or semi-aware responsiveness to the material and perceptual contingencies in which the body is embedded (Dreyfus, 2002; Kelly, 2002; Rietveld, 2008; Brownstein, 2014).

According to these theorists, what makes the live action of an expert fluid and quickly adaptive is the fine-grained arrangement of its dynamic elements and its transient dynamics into a highly coherent and temporally unitary system (Silva *et al.*, forthcoming; Araújo, Davids and Hristovski, 2006). The principles that guide and shape skilful action are complex, holistically interconnected, and truly intelligent because they are situated against a background of tacit know-how that the experts can *use* skilfully but can hardly be explicitly *aware of*

(Radman, 2012). According to this view, the cognitive preconditions of skilful action sit on this irrepresentable background and are infinite: a bottomless abyss of implicit knowledge that is ingrained within the very structure of the living body but that exceeds human verbalization and conceptualization (Wheeler, 2008; Dreyfus, 2012; Cappuccio and Wheeler, 2012). Is this background the real source of the power of the non-conscious? In order to understand if this view is correct, it is useful to discuss the theoretical principles of embodied cognition and their applications to the practical cases addressed by sport psychology.

2. Flow, Choke, and the Phenomenology of Trying Too Hard

I have already mentioned that one of the questions most often debated by sport psychologists is whether the best sport performances can be achieved through conscious or (at least partly) non-conscious modes of perception and action (e.g. Breivik, 2013). In other words, the question is whether sport skills are facilitated or rather disrupted by explicit attention and aware control. This question seems to pose a dilemma: on the one hand, many sport practitioners find it obvious that acting while simultaneously reflecting on their ongoing performance is detrimental and disrupts both action and reflection, probably because they require two different and, to some extent, competing or even mutually exclusive forms of engagement; on the other hand, certain sport practices also suggest that an athlete has to be completely present to her actions, remaining focused and avoiding distracting thoughts, if she wants to perform at the best of her capabilities. Is it the *excess*, then, or rather the *lack* of awareness that damages our expert abilities the most? Is it convenient for an athlete to try to dedicate more or less attentional resources to the execution of her actions in order to succeed? From the point of view of the aprioristic definitions of consciousness often favoured by metaphysicians, these two options are antithetical and mutually exclusive, and that is why philosophers have often tended to oppose them categorically. In truth, a careful examination of the actual relation between conscious awareness and performance, as it emerges in the concrete experience of the athletes, allows for other, more nuanced, solutions.

Optimal performance is usually associated with a particular psychological condition called 'flow' or, colloquially, 'the zone' (Jackson and Csikszentmihalyi, 1999). Flow is described by practitioners as a sustained state of absorption into their skilful performance, a condition of moderate arousal and enjoyment in which challenging tasks are accomplished without apparent effort (Swann, 2016). Some studies suggest that the experience of flow actually

correlates with an objective increase of the performance level (Swann *et al.*, 2016). Although not universally endorsed, the conviction that flow is more easily achieved under conditions of implicit or semi-conscious attention is a very old and influential one (often associated with east-Asian doctrines, see Slingerland, 2014). In fact, the belief that the 'effortless' phenomenology of flow is typically accompanied by non-conscious cognitive processing is linked to two somewhat common-sensical notions: that conscious awareness is involved whenever we are 'trying hard' to achieve optimality, and that 'trying too hard' tends to reduce one's capability to act skilfully.

For a long time, the notion of 'trying too hard' has been characterized as 'overthinking', and it was often associated with the folk psychological assumption that 'thinking' and 'acting' are reciprocally exclusive types of mental acts (Dreyfus and Kelly, 2001). Recently, this contrastive dualism has been criticized for being too vague to help scientific inquiry (see Cappuccio, 2015; in particular Christensen, Sutton and McIlwain 2015). On the other hand, 'trying hard' can be understood in a very different way, as 'over scrutinizing one's own performances', i.e. an excessive focus on movement execution (Carr, 2015). This connotation resonates with a different view that is well represented in sport psychology, and that has received support from a large set of bodies of evidence, including both anecdotal testimonies (Swann *et al.*, 2012), experimental data (Beilock and Carr, 2001), and comprehensive theoretical models (Wulf, 2013).

The current debate on skill acquisition and disruption is informed by the urge to combine the applied know-how of the practitioners with the scientific knowledge of cognitive systems, and this is why today, more than ever before, the study of skill and expertise is a multidisciplinary endeavour, involving competing coalitions of sport psychologists, philosophical psychologists, and various types of cognitive scientists. It doesn't focus on the opposition of reflection (thought) and performance (action), but invests in more sophisticated distinctions to critically evaluate the role played by non-conscious processing in peak performance. The progress in this field of research inevitably led to various diverging interpretations of this role.

The competition between different interpretative options becomes fierce when they try to explain the debilitating phenomenon commonly called 'choking under pressure', or 'choking effect', or just 'choke' (Beilock, 2011), which in many respects represents the antithesis of flow, and a key element for understanding flow itself. Choke, feared by many sportsmen and sportswomen as one of the most dangerous threats to their professional activity, is usually construed as the expert athletes' incapability to successfully perform actions that would

otherwise be routine due to psychological factors such as stress and tension. According to Baumeister (1984), it is associated with a paradox many practitioners are familiar with: the situations that overemphasize the expectation to produce excellent results are those that most likely bring about below-standard performances. This phenomenon is related to a multifaceted complex of cognitive, motivational, and emotional dynamics that have been studied widely, but ultimately without consensus, by sport psychologists (Hill *et al.*, 2010; Mesagno, 2013). Even though the scientific definition of choking may vary significantly across different schools (reviewed by Christensen, Sutton and McIlwain, 2015), and there is no universal agreement about its nature and causes, researchers have tended to focus on the allocation of attentional resources and on the role played by conscious and non-conscious processing (Jackson and Beilock, 2008).

While other explanatory narratives certainly exist (e.g. Carr, 2015; Sutton *et al.*, 2015; Jordet, 2010; Mesagno, Harvey and Janelle, 2012; Mesagno and Hill, 2013), three competing models of choking are particularly relevant to our discussion on consciousness: *distraction* (Wine, 1971), *overload* (often conflated with distraction), and *self-monitoring* (Beilock and Carr, 2001) theories. They impute the negative effects of choking to the intrusion of external competing attentional stimuli, the overflowing of internally available computational resources, or the disruption of fluid habitual motor routines, respectively. Note that, while the causes indicated by these models may tend to overlap, at least in part, it is indispensable to distinguish them conceptually: in fact, distraction is not necessarily caused by overload, whose origin and phenomenology are, in turn, different from self-monitoring.

While certain qualitative studies have been influential (Sutton, 2007; Hill and Shaw, 2013), most of the scientific debate of the last fifteen years orbits around the lines of experimental research that Sian Beilock, Tom Carr, Gabriele Wulf, and their fellow colleagues have conducted on the habitual skills that choking disrupts. The empirical results obtained by in-lab experimentation have provided some compelling arguments in favour of the third model of choking, a broad constellation that comprises all the self-monitoring theories, including pre-eminently 'self-focus', 'execution focus', and 'explicit control' theories. These empirical results have exposed the limits of the distraction and overload models (without entirely ruling them out), and have corroborated the claim that poor performances in experts are associated with conscious processing, or at least the particular form of consciousness involved in the explicit control of one's own component movements.

The most important arguments favouring self-monitoring theories of choking over the competitors are fourfold.

First of all, various authors (following Beilock and Carr, 2001) maintain that distraction theories are inconsistent with the fact that choking typically occurs in conditions of focused attention (when the athletes try to maintain total concentration on their task), while it normally doesn't affect experts in conditions of split attention (as in concurrent dual tasks, such as performing a golf putt while being asked to simultaneously recognize a given tone).

Secondly, they argue that cognitive overload theories can't account for the fact that choking doesn't normally affect experts who perform in fast action conditions (under speed instructions of the kind 'putt as quickly as possible'), while self-paced action conditions (under attention instructions such as 'putt only when you think you are ready') typically increase their chances of choking (Beilock *et al.*, 2004). In turn, self-monitoring theories account for poor performances by experts in dual-task conditions in which the secondary task demands attentive analysis of the primary motor tasks (e.g. 'tell me when the swing movement is finished'), or under attention instructions that stress the component processes of the performance (e.g. 'keep the knees bent').

Thirdly, self-monitoring theories can better explain the skill-level specificity of choking: split attention and fast action conditions have a negative effect only on expert athletes that perform well-practised sensorimotor skills, while novices typically exhibit an inverted pattern of skill disruption, as they experience a drop in their performance under exactly the conditions that don't damage the experts (i.e. dual tasks that involve neither self-monitoring nor speed instructions). On the contrary, distraction and heavy cognitive loads affect novices as much (or possibly more than) the experts.

Last but not least, theories based on distraction and overload struggle to explain why distractors and multiple tasks are consistently and successfully used in therapeutic interventions on athletes that are particularly prone to choke. Despite the higher cognitive burden, the performance of athletes is improved by any expedient that forces them to stop overscrutinizing the details of their own movements, redirecting their attention to external foci (Jackson *et al.*, 2006; Wulf and Su, 2007).

Taken together, these correlations can be used to argue that the disruption of the fluid course of familiar automatized routines is due to self-monitoring (or, more precisely, to 'execution focus', as elucidated by Carr, 2015), which is the reason why choking is often defined also as 'paralysis by awareness/analysis' (Jackson and Beilock, 2008). The

main assumption underlying their theory is that performance anxiety typically solicits self-monitoring in the athletes during competition; self-monitoring, in turn, is likely to disrupt the continuous flow of automatic action by breaking its holistic and highly organic structure into an unnaturally segmented, mechanic array of inflexibly juxtaposed, discrete kinematic components.

However, this explanation is not universally accepted. Various objections have been raised against self-monitoring theories of choking: to begin with, not everybody sees a necessary causal link connecting anxiety to self-monitoring and self-monitoring to a consistent decrease in performance (e.g. Montero, 2014; 2016; in press). Second, it has been pointed out that the small decrease in performance artificially produced in the laboratory has little or nothing to do with the catastrophic effects engendered by choking during real-life competitions, and that the data collected by the experimenters lack ecological validity, as the laboratorial setting inevitably fails to reproduce the pressure-filled environment in which choking typically occurs (Jordet, 2010; Christensen, Sutton and McIlwain, 2014). Third, it has been noted that explicit control is not only possible but indispensable to successfully completing many sport tasks, and that attention to one's own movements is necessary to effectively strategize and take appropriate decisions during complex interactive sport activities (Sutton, 2007).

These objections, in particular the third one, suggest that the self-monitoring theory of choking assumes a simplistic notion of expertise, one that emphasizes the velocity and fluidity of routine skilful actions executed in familiar situations but doesn't recognize that expert skills are also characterized by superior capabilities of decision making, problem solving, and creative improvisation in problematic or unfamiliar situations (Carr, 2015; Cappuccio, 2015). In order to assess the applicability of these objections, it is useful to identify and circumscribe the specific contribution offered by non-conscious processing to embodied cognitive performances.

3. The Role of the Non-Conscious

Whatever opinion one has about flow or choke, it should be quite uncontroversial that a significant component of the cognitive activity necessary for expert athletes to carry out their sport tasks is done at a non-conscious level. The stimuli subconsciously processed by the athletes on the field play a very important role for their sport tasks. There are cases in which subconscious stimuli affect performance level more than if they were explicitly perceived.

For example in soccer, during penalty kicks, the position of the goal-keeper on the goal line affects the penalty taker's choice of the direction

in which he will send the ball (penalty takers typically choose to send the ball in the direction of the goal area that is least reachable by, i.e. furthest from, the goalkeeper). Masters, van der Kamp and Jackson (2007) have shown that the preference for one side or the other is stronger when the penalty taker is not consciously aware that the goalkeeper's position is displaced on one side of the goal line, i.e. when the length of displacement from the midpoint is far enough to be subconsciously processed by the penalty taker as significant for his task, but not sufficiently far enough to trigger awareness. The length of the displacement must be small and must be perceived subliminally in order to affect the decision of the penalty taker. One reason could be that, if the displacement of the goalkeeper on the goal line were so salient that the penalty taker became aware of it, then the penalty taker would have reason to think that the goalkeeper is intentionally assuming that position in order to influence the penalty taker's direction choice, strategically soliciting him to send the ball in a predictable direction. This kind of conscious consideration may trigger in the penalty taker explicit inferences capable of neutralizing the otherwise subliminal effect of the solicitations provided by the cue.

But what happens when subliminal cues compete with more supraliminal ones? To clarify whether implicit information exerts an intrinsic influence on the decisions of experts, a follow-up study (B. Noël, van der Kamp and Memmert, 2015) ascertained that subconsciously accessed perceptual information about the asymmetric disposition of the goalkeeper on the goal line can affect the penalty takers' decision of the goal side in a 'rather pervasive' way also when 'many supraliminal sources of information' are available. In these cases, the key variable is whether the expert is acting automatically or deliberately monitoring his own actions. In this experiment, 'implicit influences of the goalkeeper's position on goal side selection were consistently overridden by the (conscious) perception of the direction of the goalkeeper's dive... only if the penalty takers deliberately monitored the goalkeeper's action and the goalkeeper committed early enough for penalty takers to respond.' This means that the actions of experts can be flexible and adjust to context when decisions about the environment are taken non-consciously, as experts have a remarkable capability to extract rich and useful information from implicit cues when they are performing without self-monitoring.

This result is confirmed by another study, by Bocanegra and Hommel (2014), which infers that explicit monitoring is useful only within a circumscribed scope: cognitive control is beneficial for challenging tasks or in unpredictable situations, but is overall detrimental for routine activity, which tends to be more flexible and

adaptive if executed automatically. The authors maintain that 'top-down control can impair and interfere with the otherwise automatic' actions, making behaviour less efficient. This is especially true of contextual scenarios that are appropriate to the execution of routine actions, i.e. environmental circumstances that provide the agent with sufficient familiar information to easily predict the consequences of his actions and hence 'behave on autopilot based on automatic processes alone'.

To test this hypothesis, the subjects of this study were asked to perform a decision task responding to various visual stimuli that varied in a number of features, including shape, size, and colour. Unbeknownst to the volunteers, the colour feature always predicted the correct response. The task was replicated in easy (familiar) conditions in which decisions tended to occur automatically, without explicit reflection, as well as in challenging/unfamiliar scenarios that prompted explicit attention. The statistical distribution of successful and unsuccessful trials showed that the subjects subconsciously picked the useful information from the colour and unconsciously used it to solve tasks, but only when the subjects were not dedicating too much attention to the decision, namely in the 'easy/familiar' condition of the task. On the contrary, when the subjects were 'exerting mental control in a predictable situation, when automatic response was enough', their performances were impaired, suggesting that 'the unneeded mental effort appeared to interfere with what is a perfectly adequate automatic performance.'

This result seems to support the belief that automaticity, the capacity to perform effectively and take decisions without explicit control or reflection, not only statistically correlates with expert performance, it actually causally contributes to the mechanisms of expert behaviour, enabling more accurate and efficacious performances. But this leaves open a question of a normative kind (stressed for example by Gottlieb, 2015), concerning what behaviour can legitimately be considered 'skilful': is adaptive automatic behaviour just a prerogative and a distinctive mark of skilful expertise (i.e. is it only a sufficient condition to be considered an expert, a feature precluded to inexperienced agents); or is it the case that adaptive automatic behaviour is also a necessary and constitutive condition of skilful expertise, in the sense that the capability to subliminally manipulate implicit information offered by predictable task environments is one of the prerequisites of expert action? In order to answer this question I will examine the notion of consciousness involved in skilful performances, first introducing and then following the interpretative grid proposed by Birch *et al.* (in press).

To counter the intellectualist view that sporting skills are not cognitive, Birch *et al.* (in press) defend the claim that sensorimotor activities involving a deep dynamic integration of neuronal and bodily processes, like playing basketball, are not less cognitive and intelligent than a purely contemplative and static task such as solving an equation. Two traditional dogmas must be rejected in order to rediscover the unjustly underestimated embodied dimension of cognition: 'the first is the belief that sporting skills are not knowledge; the other is that sporting skills go on (or should go on) without consciousness' (*ibid.*). Against these dogmas, Birch and colleagues argue that many embodied activities are properly cognitive and intelligent in so far as they involve *knowledge* and *consciousness*. Knowledge itself, according to Birch and colleagues, can be characterized as a defining characteristic, or even a particular form, of consciousness. As such, it pervasively informs the performances of the experts.

4. Defining Consciousness: Nine Dimensions

Birch *et al.* (in press) consider the view that sees skilful activity as necessarily non-conscious misleading and they reject it: while some self-paced closed disciplines put a higher stress on the perfect repetition of standardized sensorimotor schemata, without involving high levels of active control and decision, in all sport disciplines consciousness is a key element for a perfectly executed performance. Building importantly on the philosophical reflection advanced by Breivik (2007; 2013), Birch *et al.* defend this claim analytically, arguing that all the defining features of consciousness are actually preserved in skilled action and even during repetitive routine performance. Skilful action involves both *access* (or 'psychological') *consciousness* ('a mental state which plays a causal/functional role in the production of behavior') and *phenomenal consciousness* ('how it feels' and 'what is it like' to have a certain experience), in accord with the famous distinction introduced by Block (1995). Phenomenal consciousness is the distinctive characteristic of the subjective qualitative mental states experienced by all beings having a body and inhabiting a world (as opposed to 'robots or zombies').

The key functional features of psychological consciousness, on the other hand, are captured by eight dimensions: *awakeness* (which demarcates conscious states from unconscious ones like sleep and coma); *introspection* (the capability to monitor and interrogate both one's own internal states, such as emotions, and physical states, like body posture and motion, including proprioception and interoception); *reportability* (representing in declarative terms events and features of one's experience); *self-consciousness* (the sense of one's personal identity,

as distinct from others’); *attention* (the active and selective allocation of cognitive resources for processing online stimuli; it can be wider or narrower, focused or split, etc.); *awareness* (i.e. the sensibility and responsiveness to trigger stimuli; unlike attention, it is a passive and general, i.e. non-selective, form of promptness); *voluntary control* (one’s sense of agency, involving both a sense of initiation and a sense of control); and finally *knowledge* (which includes the capabilities of adaptation, explanation, and prediction that allow the agent to efficaciously deal with familiar and unfamiliar situations based on stored information and previous experience).

Each of these nine dimensions is represented to some degree in the typical skilful performance. Most, if not all, are indispensable to sporting activities.

1) Athletes certainly are *phenomenally conscious* when they perform. Their first-hand experience involves qualia, i.e. the private mental states that characterize as intrinsically subjective our percepts, feelings, sensations, and emotions during intense physical training or performance exactly like during any other moment of our lives. In particular conditions, athletes are not fully aware of how the environment affects them (e.g. subliminal cues), or they are not capable of fully reporting on tasks they were immersed in; and yet, this doesn’t prove they are not phenomenally conscious, as their performance was always subjectively oriented and coloured by a distinctive ego-centred experience (e.g. certain moods or emotions such as arousal, fear, discomfort, fun, or a complex blend of them). The performance certainly ‘felt’ in some particular way, qualitatively irreducible to its verbal description.

The study of skilful performances offers good arguments in favour of the real (as opposed to epiphenomenal) status of qualia: not only do qualia colour the athlete’s personal experience in a particular manner; they also modify the subpersonal cognitive processes in the athlete’s mind. As reminded by Birch *et al.* (in press), this view is well represented in cognitive neuroscience: Edelman (1992, pp. 135–6) doesn’t just claim that phenomenal consciousness accompanies automated action; it also contributes to its success, exerting an effective causal impact on ‘discriminating sensory perceptions’ and subsequently on ‘action and behavior, whether we are able to report it or not’.

The intrinsically qualitative dimension of the athlete’s experience is not a supplementary but a constitutive and causally relevant component of the cognitive processes that allow optimal attunement of an agent with her world. This explains why all of the eight functional dimensions of psychological consciousness are involved during skilful performances. Their systemic dynamic integration would be harder to explain if we didn’t factor in phenomenal consciousness, in that the

functioning of the athlete's mind as a highly coherent unitary system is unintelligible without the unifying factor represented by his personal experience.

2) Revisiting the features of psychological consciousness previously mentioned, Birch *et al.* (in press) note first of all that athletes are certainly *awake* when they perform, as they don't seem asleep or in a trance.

3) Also, athletes often need to *introspect*, as they must be capable of monitoring their body, for example comparing their posture with a pre-given model of their movements (or a choreography) to correct the position and the kinematic of the body parts; also, they may need to control their internal states and attend to their own psychophysical condition in order to detect when emotions, feelings, and somatic sensations deviate from an optimal standard, enacting appropriate compensatory or regulatory responses (for example when they are reaching the dangerous edge of physical exhaustion, sensorial overload, or emotional meltdown).

4) Definitely, athletes can *verbally report* about their performances (for example, football players interviewed after a match), providing elaborate descriptions of specific aspects of their behaviours, decisions, and the circumstances that motivated them.

5) They appear *self-conscious*, as they interact with inanimate objects, the other athletes, and the people in the audience in ways that suggest the stability of their personal identity and the structural persistence of their self.

6) As for *attention*, Birch *et al.* admit that it can be modulated in very different ways through different sporting disciplines ('focused, selective, spotlight, integrating, divided, alternating, etc.'): the type, the amount, and the robustness of the needed attentional resources may vary significantly between tasks; however, the authors exclude that any sporting discipline could ever be successfully practised by players affected by severe lack of attention, probably because this would prevent concentration.

7) *Awareness* is also inalienable from the experience of an expert player because, regardless of the depth of her absorption in the task at hand, and her capability to isolate herself from external contingencies, she is still sensitive and responsive to a broad perceptual horizon contextually dependent on her action possibilities. Such tight relation to the context is key to staying tuned into the continuous changes in the experiential flow while remaining in 'the zone'.

8) Regarding the *voluntariness* of the actions executed by sport people, the authors remark that they 'are certainly not random or lucky'; rather, they are the result of well-formed intentions. Even when

an agent is not entirely aware of the action she is going to perform, its movements spawn from an act of volition (as opposed to being merely reflexive).

9) Last but not least, Birch *et al.* argue that sporting skills embody a rich form of knowledge characterized in procedural and not declarative terms, i.e. an operative know-how that informs the body's flexible adaptivity to the environment. Unlike traditional cognitivist approaches, the embodied view of cognition maintains that this knowledge is less accurately characterized in propositional terms than in terms of *motor intentionality*: this concept, developed at first by phenomenologist Maurice Merleau-Ponty, refers to a non-representational form of directedness, or 'aboutness', of the mind that makes it possible for a situated agent to directly perceive the world as a rich landscape of meaningful possibilities of interaction (Merleau-Ponty, 1945; see Dreyfus, 1996, for a discussion of its contemporary relevance). Unlike other forms of cognition, the knowledge incarnated by motor intentionality is not normed by conditions of satisfaction, in that this knowledge is not about possessing a more or less accurate representation of the external world, but conditions of improvement, which define how our actions deviate from an optimal standard, i.e. a certain implicit balance with the world that is only tacitly sensed, never fully represented (Dreyfus, 2002).

While this kind of embodied knowledge is usually associated with automaticity, Birch *et al.* (in press) find that this association needs to be clarified: the know-how enacted by the body through its motor intentional attunement with the world relies on subpersonal cognitive mechanisms (like those localized in our premotor cortex) that work in an automatic way, i.e. without explicit decisions being possible or necessary; but that doesn't mean the agent is compelled to act in a certain way when she is guided by motor intentionality, as motor intentionality (unlike the reflexes that are pre-programmed by genes or hardwired in the system through associative mechanisms) is a relational and dynamic knowledge involving all the aforementioned elements of consciousness. This is fundamentally different from an automatic predisposition to deliver fixed responses.

5. The Other Side of the Story:

Non-Conscious Modulations of Conscious Awareness

I believe this analysis is essentially accurate in revealing how the main components or features of consciousness, conceptually identified by philosophers, are present in the experience of athletes. But the conclusion presented by Birch *et al.* invites an additional interpretative effort

to recognize the specific role played by legitimately non-conscious dynamics during skilful activities.

Birch *et al.* seem to reduce subconscious or semi-conscious experiences, *qua* personal experiences, to the positive realm of consciousness: the point maintained by the authors is that, even when we drive home on 'automatic pilot' (to use the famous example by Armstrong, 1981), we are *not entirely* (and certainly *not reflectively*) *aware*, but we clearly live the qualitative circumstances of driving as a subjective experience. We are still—at some level—aware of the situation, and more or less ready to respond to sudden changes; consciousness of the road conditions may be in its minimal form, but it never completely abandoned us.

While I agree with this positive account, I wonder whether it would be satisfactory if we didn't complement it with an empirically and phenomenologically accurate account of the negative modulations that specifically characterize consciousness during automatic skilful action. Consciousness is present in both reflective and autopilot driving, but is it equally present in both of them? Conscious engagement in automatic skilful action involves a different, negative or privative, dimension. It is a particular form of absence and disengagement that makes it possible to characterize the distinctive phenomenology of this experience, which is both qualitatively (it feels differently) and functionally (it works differently) irreducible to the experience of positive presence associated with fully aware and reflective driving.

The expert's experience and the novice's are paradigmatically different in this respect, as the novice's certainly includes the fully aware and reflective dimension of action, but hardly involves the automated and reflective one. Could we ever capture the specificity of the minimally articulated form of consciousness associated with the absorbed performances of expert athletes, if we excluded negative attributes and privative notions such as diminishment, absence, or disengagement? As I believe the answer to this question is negative, my proposal is to stop treating automatic action as a zombie-like mechanical procedure, considering it instead as a phenomenological *chiaroscuro* in which explicit and implicit intentional acts are intertwined, forming a complex patchwork of foreground and background knowledge that can be fully, only partially, or not at all accessible at the personal level, in different degrees and at a different times (refer to the dynamic model of background knowledge first proposed in Cappuccio and Wheeler, 2012). This involves a particularly nuanced pattern of features that we usually associate either with full awareness or with unconsciousness, respectively. Their dynamic interplay can be recognized by examining

once again the nine characteristics of consciousness indicated by Birch *et al.* (in press).

To begin with, athletes absorbed in their skilful performances are certainly *awake* (1) and vigilant, but a more interesting question is whether their state of wakefulness is qualitatively and functionally the same when they are not absorbed in their performance. Without suggesting that their condition were in any significant sense comparable to a kind of 'trance', or a blind repetition of mechanical routines, we cannot underestimate the specificity of their particular way of being awake either. This specificity is attested to by the fluctuations in the performance levels that an expert athlete experiences when the modes and the targets of her attention vary.

For example, we know that the condition of flow importantly correlates with three factors: skill level (Catley and Duda, 1997; Engeser and Rheinberg, 2008; Jackson, 1996), performance, and experience, meaning that expert athletes are more likely than novices to achieve a condition of deep absorption and maximize the benefits of this condition during challenging tasks. The phenomenological specificity of this experience is the reason why many experts believe that flow consists in a peculiar modulation of consciousness: flow increases the success rates in expert sport performance because the cognitive functions recruited for the task at hand are contingent upon the sub- or semi-conscious processing typically favoured by flow. Hence, from a cognitive point of view, being awake while 'in the zone' is very different from being just generically awake.

6. Introspection and the Attentional Focus Effect

The specificity of this condition is reflected by the particular limitations typically suffered by the athletes' capability of *introspection* (2) during peak performance. Importantly, these limitations seem *an enabling condition* to achieve peak performance: obviously, athletes can introspect during their performances, in the sense that they never lose their sheer capability to monitor their postures, movements, and their internal states; but athletes don't exert their full introspective potential when they produce top performances, or when they simply want to be efficacious in their sport activity. The empirical evidence in favour of this claim is abundant.

I am referring to the 'attentional focus effect', largely documented by the experimental activity by Gabriele Wulf and her colleagues (Wulf, Shea and Park, 2001; Wulf and Su, 2007; Wulf, Lauterbach and Toole, 2009; Wulf, 2013), and replicated by several other research groups. This comprehensive body of empirical data corroborates a theory of attention that distinguishes between internal and external

foci. The attentional focus effect model goes hand in hand with the theory of choking as 'paralysis by awareness/analysis' or 'explicit monitoring' and seems to back it, although there are differences. The model essentially claims that attending to the constitutive elements of one's performance (paradigmatically observing the components of one's own actions) usually brings about substandard outcomes. Attention to 'internal foci' (typically one's own body parts, or aspects of one's movements) damages the precision and the coordination of one's actions even when they are just familiar routines. On the contrary, paying attention to 'external foci', like the immediate transformational effects of one's action, or their physical targets, doesn't reduce the quality of one's performance and in many cases can actually help improve it, offering a dynamic indicator that co-varies with one's actions without being a component of these actions.

This is useful to strategically regulate one's own action or stabilize one's posture and movements in an indirect manner, without needing to directly attend to internal elements of one's performance. The paradox captured by this theory is that the best way to adjust or correct one's action is to avoid attention to the action itself, exploiting features, processes, or events that are produced by or predicted from one's action. They provide all the information necessary to update one's movements without directly attending to them. This theory and its variants discourages coaches and trainers from using instructions that explicitly describe (or even mention!) body parts, movements, and postures; rather, it suggests focusing on the expected outcomes and the immediate effects of one performance, or on indirect instructions intended to facilitate its achievement.

When it is indispensable to refer to internal components, the internal focus theory recommends the use of metaphorical descriptions or other means of analogical representation (in accord with implicit learning theory: Master, 1992; Komar *et al.*, 2014). If the athlete's attention is recurrently – possibly unintentionally and compulsively – being drawn by internal foci, then an intervention based on distractors and secondary tasks, to direct the attention away from the internal foci, may be necessary. Such forms of intervention prove highly beneficial to mitigate the choking effect (Wulf *et al.*, 2002). According to Wulf (2015; 2016), the model can be generalized to all kinds of expert activities because of the robustness of the attentional foci effect and its very high replicability across very different sport disciplines and typologies of tasks. Moreover, Wulf maintains that this effect is independent of the individual characteristics of the agents (e.g. age, gender, disability) and of their skill level (unlike the phenomenon of choking by awareness/

analysis/self-monitoring, which clearly correlates to expertise and automatization).

While the large amount of empirical data in support of this model cannot be ignored, their universality has been questioned by psychologists and philosophers who defend a different interpretation of attention in skilful performance. Toner, Montero and Moran (2014) and Montero, Toner and Moran (in press) have questioned the internal/external focus distinction, claiming that it is ill-defined. In particular, the apparent vagueness and extreme variability of the notion of external focus would confound the domain of applicability of the theory: external focus seems to indicate alternatively the immediate effects of an action (scoring a goal), an external object or perceptual cue, a quality ('speed, tempo, going fast, swimming hard') of the action, or its metaphorical representation; however, particular features or metaphorical descriptions of an action count as internal, if they refer to a particular isolated body part or movement, as opposed to the movement of the whole body: for example, 'focusing on the angle of one's wrist and focusing on one's wrist and hand taking the shape of a candy cane, for example, would both count as internal foci since they identify isolated bodily parts' (Montero, Toner and Moran, in press). Also, Montero and colleagues impute 'contradictory' claims to Wulf (2016), such as the claim that 'adopting an external focus does not mean that the performer is not aware of her or his body movements' (p. 337), apparently contradicting the claim that an external focus is 'related to the *planning* of the movement, but *has nothing to do with the processing of intrinsic feedback or bodily awareness*'. Montero, Toner and Moran (in press) deem these claims contradictory in that they indicate that 'an external focus both allows and excludes an internal focus (bodily awareness).'

This contradiction might be only apparent. If interpreted within the framework of control theory, the two claims by Wulf seem perfectly compatible with each other and with the attentional focus theory: it is perfectly possible for a forward system to control the online action simply by trying to reduce the difference between the effects produced by the action and the prediction of these effects generated by the system according to an internal model of the action. The system doesn't use detailed information about the execution of the action itself, because this information could impair the online capability of adjustment by abnormally over-intensifying the sensitivity of the prediction-matching system, which in turn would generate a cascade of false-errors capable of inhibiting the correct execution of the planned movements. According to the model presented by Clark (2015, pp. 218–20),

predictive processing theory is consistent with the attentional focus effect.

Like the metaphors used in implicit learning, the predictions generated by the system to anticipate the effects of its own actions vicariate the role of an internal focus by replacing it with an external focus. A confirmation of this intuition is offered by the fact that neuro-bio-feedback systems are effective in accelerating training and peaking performance in tasks that require high levels of concentration (as in marksmanship; see Berka *et al.*, 2012). The reason is exactly that such a system allows the agent to adjust her actions without directly attending to them: the system, through portable dynamic EEG technology, receives information from the brainwaves of the agent and, as soon as it registers high levels of relaxation and deep concentration (as in flow), uses tactile feedback to make the agent aware that it is the best time to shoot. The shots executed in these circumstances are in fact statistically more successful. The agent is trained to try to reach and maintain the optimal mental state that the EEG system is programmed to detect; the subject modulates the responses of the system with their sheer will until the conditions for an optimal performance arise. The agent corrects his execution without introspecting (without focusing on his current internal state), simply relying on an indirect representation of the internal state calculated by the artificial system. Attention has an internal focus only in an indirect, mediated, and delayed sense: the human agent alone (unlike the EEG system that is specifically designed for this purpose) wouldn't be naturally capable of directly monitoring his own brain state. The prediction of the optimal brain states at the same time informs and is produced by the execution of his actions, and no internal focus is ever directly targeted in this closed feedback action loop.

A different objection against the attentional focus effect theory comes from Sutton (2007). The accurate analysis of the reports by cricket batsmen provided by this study highlights that self-monitoring and attention to component movements are not just possible and likely to happen on the field without apparent decrease; they are also indispensable to allow the professional players to correctly adjust their posture and maximize the effectiveness of their actions. This qualitative study contributed to undermining the myth that expert athletes don't (*and can't, qua experts*) consciously control the component processes of their actions: it provided theoretical grounds to critically discuss the maximalist interpretations of the self-monitoring theory of choking (i.e. those interpretations that see self-monitoring and explicit control as necessarily and ultimately detrimental to all kind of skilful activities, as suggested by Dreyfus, 2002).

At a first glance, this seems sufficient to question the universal validity of the attentional focus effect theory: certain sport tasks (such as swimming; see Swann, 2016), because of their closed nature, benefit from the explicit awareness of one's posture and the kinematic components, especially for those actions that, in order to be optimized, need very precise trajectories and fine-grained mechanical adjustments that can neither be automatized, regardless of the amount of experience acquired, nor controlled through indirect strategies based on external foci or metaphorical representations. While this is true, it might not be enough to exclude the existence of a trade-off (perhaps just a minor one) between self-monitoring and quality of the execution, as per attentional focus theory. It is possible, in fact, that the execution of certain types of tasks receives both a benefit and a drawback from attending to internal foci, and that the benefits happen to be higher or more visible than the costs. If that is the case, then a general disruptive effect due to the attentional focus might still be present, although it would be largely exceeded – and therefore masked – by the concurrent benefits produced by the finer control, involving a more precise coordination brought about by explicit monitoring.

The existence of a general disruptive effect on performance due to the explicit control of one's own actions cannot be excluded *a priori* if we consider only tasks that cannot be carried out without such explicit control, as there would be no valid term of comparison (control condition) to measure the corresponding performance decrease. On the contrary, various experiments show that the negative effect engendered by internal foci is perspicuous regardless of whether the tasks were carried out with or without self-monitoring (i.e. automatable tasks that don't strictly rely on body awareness; actions that don't require fine adjustments in the posture or calibration of the movements against a pre-given model). This is the kind of setting that allows us to systematically measure the decrease in performances (see Wulf, 2015; 2016).

In this kind of setting, introspective consciousness (at least the kind of introspection that targets internal foci) is still systematically correlated with substantive decreases in performance. For the purposes of our analyses, this means that the sense of body awareness of an expert athlete who performs successfully by relying on his automated embodied routines is modulated in a specific negative way – one that naturally tends (or has been exercised) to avoid or reduce introspective attention.

7. Reportability, Attention, Awareness

Limited capabilities of introspection imply limited capabilities to *verbally report* (3) on one's actions. In fact, the reports provided by experts,

in comparison to novices, suggest a less accurate and a coarser-grained capability of recollecting information about events and features experienced by experts absorbed in the execution of their tasks (even if at least some of this information was obviously accessed and processed, probably subconsciously, during the task). This effect has been documented systematically over different disciplines, is particularly evident in closed self-paced tasks, and its intensity seems positively correlated to skill level.

Beilock, Wierenga and Carr (2002) use the expression 'expertise-induced amnesia' to indicate the impoverished episodic memory experienced by skilled athletes: they display a generally reduced 'declarative accessibility', i.e. a lower 'openness to introspection and report'. This phenomenon justifies the inference that experts don't primarily rely on declarative forms of cognition to execute their routine tasks, especially when the sporting task doesn't require the use of working memory, i.e. the kind of memory that specifically manipulates declarative representations: various experiments show that the cognitive system is not overloaded by secondary tasks that use working memory, if the primary task is a sport task that doesn't involve working memory either (e.g. Beilock and Carr 2005; Beilock *et al.*, 2006). This separation is accounted for by 'dual-process' theories that postulate the fundamental autonomy of the different cognitive subsystems that, respectively, use and that don't use working memory (Furley, Schweizer and Bertrams, 2015).

Expertise-induced amnesia doesn't contradict the well-known fact that experts can typically produce copious, rich, and articulated commentaries about their own performances (or others', as done by TV sport commentators): skilled players, rather than relying on the precision of their episodic memory, exploit their superior familiarity with the discipline (procedural memory) to imaginatively reconstruct the circumstances of their actions by elaborating on minimal recollections. This way, experts find it easier to compensate for the lack of precise factual information about the occurrence of particular events during the task or about the presence/absence of particular features associated with their performance.

How are *attention* (4) and *awareness* (5) specifically modulated in the experience of expert athletes? The answer is not simple, as it largely depends on the kind of task and the training done to maximize one's performance. Attention can be more or less narrowly funnelled during the sport task—depending on the nature of the sport discipline itself, and whether the task is closed or interactive, self-contained or continuous, self-paced or synchronized to other agents or event. In any case, the attentional field of the athlete is shaped by the way he

concretely inhabits the scenario of his performances, including first of all his expectations about the perceptual environment and the prospects opened by his previous performative knowledge. It is necessary to point out that such know-how, in turn, depends importantly on his direct familiarity with the typical circumstances of the performance, on the motoric and perceptual habits that he has developed through his training, and ultimately on the way his embodied adaptive dispositions (including a primordial background of purely reflexive reactions, emotions, and visceromotor sensations like pain) have been moulded by the repetition of previous trials.

Each way of engaging in interaction with the world brings about its own specific manners of attending to it and perceiving it. Different practical stances and goals literally change how athletes subjectively see, hear, and feel things: for example, the size of a baseball seen by a batter depends on how well he hits (Witt and Proffit, 2005), and the diameter of a golf hole is seen as larger by golf players when they putt successfully (Witt *et al.*, 2008). Similarly, the perception of the width of the gap between two goal posts varies with the facility that the football has been kicked between them (Witt and Dorsch, 2009). Durgin *et al.* (2012) and Firestone (2013) have questioned the interpretation of these results, claiming that these effects are not produced by an actual modification of one's perceptual patterns, but on retrospective biases. However, Witt (2015) and Foerster, Gray and Cañal-Bruland (2015) have convincingly argued that perception of the action itself, not the judgment about the performance's outcome, is specifically affected by the agent's performative expertise. These experiments manipulate perceptual feedback to dissociate the perceived outcome of one's performance from the objective quality and precision of the performance: this way, they have proven that a subject's tendency to perceive larger-than-real-size targets correlates only to the higher quality and precision of his performance (which in turn depends on the depth of his absorption into the task and, of course, on his skill level), not on the perceived outcome of his performance.

Expert athletes don't see the world like a general agent would see it; their senses are fine-tuned to pick only the information that is relevant to their tasks, information that is intrinsically characterized in an interactional, practical, and motoric way. Both Gibson's theory of affordances (1979) and Merleau-Ponty's theory of motor intentionality (1945), mentioned by Birch, Breivik, and colleagues, efficaciously account for the cognitive mechanisms underlying these empirical results: what makes the experts expert is exactly that they see opportunities for action and practical occasions that invite intervention or careful response, where average observers can only see pragmatically neutral,

objective, uncharacterized environments. Neuroscientific investigation confirms both the phenomenological intuition and the ecological principles behind this concept. For example, the discovery of canonical neurons explains why the perception of affordances in the environment (e.g. graspable objects) builds on the capability to execute the corresponding actions (precision grasp): the perceptual system recruits the competences stored in the repertoire of goal-oriented motor actions that is encoded in the premotor cortex of humans' and monkeys' brains (Fadiga *et al.*, 2000; Rizzolatti and Umiltà, 2013).

This specific calibration of the perceptual system, regulated by the motoric proficiency in particular classes of goal-oriented skills, implies that the attentional resources available to the system are recruited selectively, in accord with the ongoing modulation of one's state of conscious awareness, which in turn depends on the task at hand and the subject's practical familiarity with it (see premotor theory of attention; Craighero *et al.*, 1999; Rizzolatti and Craighero, 2010). This selectivity suggests that consciousness in expert athletic performance is characterized by a negative or privative dynamics, as it is modulated through a task-specific or goal-specific type of obstruction or redirection of the available global attentional resources. The creation of areas of disengagement and unawareness is a by-product of this filtering process: while the senses of the athletes are receptive to external, even distant, unanticipated stimuli, and ready to non-consciously process them, their zone of awareness is curved to filter out the percepts that are irrelevant to the success of their goal-oriented motor activity.

8. Self-consciousness, Volition, and Knowledge

If motoric expertise and skill level modulate perception, then they are likely to modulate self-perception too. This is documented by the diminished sense of self-consciousness (i.e. decreased awareness of self and of social evaluation) reported during flow by approximately 30% of athletes, most often combined with other perceptual modifications such as transformation of time (Swann *et al.*, 2012, p. 13; see also Jackson and Marsh, 1996). Why do athletes suppress self-consciousness? One of the reasons might be that, in order to achieve a flow state, athletes have to replace negative thoughts with positive feelings of fulfilment, and fear of the judgment of others can be one of the most disruptive thoughts (Mesagno, Harvey and Janelle, 2012).

In fact, it is well-known that the kind of awareness of oneself solicited by social evaluation (intended as explicit judgment about oneself, or reflection on the possible opinions or beliefs of others) is often detrimental to one's performance and undermines flow: 'stereotype threat', in particular, can significantly affect performances by

prompting negative thoughts that undermine confidence and weaken motivation. Even just mentioning negative stereotypes can disrupt performance (Beilock *et al.*, 2006; Beilock, 2011). This is either because recurrent thoughts overload the cognitive system (saturating the working memory) or because they prompt reflective tendencies that increase the chances of compulsive self-monitoring. In either case, successful athletes have reasons to avoid social judgment in order to prevent negative self-thoughts.

Self-consciousness is related to performance also in other ways: i.e. not just thinking of the judgments of others can be detrimental; also observing their poor executions can be too. In various sports, especially team-based ones, expert players are prone to contagion effects that can be explained—for example—in terms of emotional or motoric resonance between players. One player performance is affected negatively or positively by the unsuccessful or successful performances of the other players. Different models try to explain the cause of contagion: for example, it may depend on the (de)motivation arising from the general mood of the affiliated players (Moll, Jordet and Pepping, 2010); or on the fact that the observer's motor system is primed in specific ways when it is recruited to internally simulate the observed actions of others (Ikegami and Ganesh, 2014); or on motivational mechanisms related to the natural tendency to preserve consistency between one athlete's beliefs about himself, about the other players, and about the quality or value of their respective skills (as in Fritz Heider's Balance Theory; see Cartwright and Harary, 1956). In any of these cases, the self-conscious dimension of one athlete is directly modulated by interactions between team players. A psychological link between the actions of the subjects involved in a match exists, and produces remarkable effects during intense competitions, when attention to one's own and others' performances is dramatically increased by the psychological pressure. The boundaries of one's consciousness are not autonomously defined, but depend by the presence of others.

This idea is consistent with the 'shared manifold hypothesis' in social neuroscience (Gallese, 2007): regardless of what we are doing, our brain constantly simulates the behaviour of surrounding agents. Specific brain areas (motor, emotional) become selectively active not only when an agent executes an action or experiences a certain emotion, but also when the agent perceives other agents executing the same action or experiencing the same emotion. This strengthens the idea that team players reciprocally coordinate through an intersubjective network of resonating brain activities. When they successfully attune to one another, they create a supra-personal system of cognitive interactions.

This hypothesis is not enough to suggest that the boundaries of one's self-consciousness could be transcended by intersubjective resonating brain processes: one's subjective experience remains fundamentally, irremediably ego-centred (Zahavi, 2005). It remains, in some sense, not entirely communicable. But this hypothesis certainly invites us to abandon a solipsistic way of understanding consciousness as the isolated prison of personal identity: identity is not the effect of a self-contained system, but a relational phenomenon structurally open to the experience of others, and emerging from a network of conscious interactions. This dynamic is even more marked during challenging skilful activities, when the emotional and cognitive attunement between agents is key, for the better or for the worse, in determining the outcome of their individual actions in a collective context.

Sense of agency, another key feature in the phenomenology of self-consciousness, is also peculiarly modified in skilful activities. Sense of agency is crucial because our capability to experience our own actions as voluntary or involuntary (8) depends on it: according to control theory, the feeling of being an agent, i.e. the one who intentionally initiated a certain movement, derives from a comparison of the sensory predictions generated when a motor command is sent to initiate an action with the sensorial feedback actually produced by the execution of that action. This is how the system distinguishes between endogenous and exogenous movements, informing the agent at once about the actual origin of the ongoing changes in the perceptual flow and about the boundaries between internal and external world (e.g. it tells me whether the room seems to be spinning because my body is rotating or because the walls are actually moving around me).

Now, there is evidence that sense of agency is qualitatively modified during situations of flow involving deep absorption. While athletes' capabilities of identifying the actions that they have voluntarily started is not substantially modified (unlike in schizophrenic patients), the subjective experience associated with this function is modulated in a characteristic way. Expert athletes—especially in disciplines that emphasize fast reactivity and appropriate responsiveness to an opponent's actions, like martial arts—often describe their movements as springing spontaneously, seemingly without explicit intention or active volition, without thought or decision being necessary or possible: a situation in which the difference between activity and passivity, moving or being moved, seems to dissolve.

The essence of this phenomenon is paradigmatically epitomized by martial artist Bruce Lee's famous statement, 'I do not hit, it hits all by itself' (whose speculative meaning, clearly influenced by traditional Taoist doctrines, is developed in Lee's sport and philosophical treatises;

see Thomas, 2005). But a similar subjective characterization also emerges from self-reports produced by expert athletes when asked to describe the phenomenological specificity of their engagement in situations of flow (Swann, 2016). One of the most common phenomena, reported by almost 75% of athletes is 'action-awareness merging' (i.e. total absorption, or 'feeling completely at one with the activity' during the flow state).

A recent study with elite golf players (Swann *et al.*, 2016), documents that flow corresponds both to a feeling of calm/relaxation and a sense of total control, often described in terms of semi-automaticity of movement, effortlessness, and enjoyment. The athlete often describes her attitude in this condition in terms of 'letting it happen', as opposed to more effortful modes of engagement defined as 'making it happen', an expression that indicates an increased sense of alertness and effortful control. The complete awareness of the intense challenge faced by the athlete during flow, combined with the high level of performance typically achieved in this condition, results in a positive sense of exceeding one's expectations and 'surfing' problematic contingencies with ease. Crucially, this condition is experienced as a feeling of relief, a mitigated sense of effort, and a greater capability to rely on automatic subconscious routines that don't require voluntary decision and reflection.

Sense of agency and voluntary action don't disappear in this condition (the subject is still aware of being in charge), but their experiential valence is modified: the subject does not experience his involvement as a way to transform the situation, but as a way to respond to it, compensating the contingencies to find a balance with them. Volition (or intention) is not an antagonistic force emanating from the core of the self to oppose the world and alter it according to a plan; rather, volition is encountered during flow as an effect, a surface tension that vibrates through the dialectic of internal and external forces, emerging from the encounter of trained predispositions and challenging contingencies.

Last but not least, also knowledge (9) is distinctively modulated during flow: motor intentionality certainly constitutes a kind of knowledge, as pointed out by Birch *et al.* (1996); however, Merleau-Ponty (1945) also points out that it differs at root from conceptual knowledge. The know-how embodied by motor intentionality is never entirely accessible by aware consciousness, not even through careful introspection (Kelly, 2000; 2002). For example, Merleau-Ponty famously noted that experienced stenographers perfectly know where to find each key on a keyboard while they are typing, but wouldn't be able to indicate the overall positions of the keys on a diagram of the keyboard

if asked to do so. This dissociation between a 'spatiality of situation' (mapping the environment in terms of opportunities for action) and a 'spatiality of position' (mapping the environment in terms of objective metric relations) is confirmed by clinical neuropsychology, which has documented how brain damage that affects the capability to execute particular goal-oriented actions doesn't necessarily affect the capability to conceptually decompose those actions into neutral geometrical elements, and vice versa. This dissociation has been variously interpreted, for example through the two visual pathways theory (Milner and Goodale, 2008). The fact that two different systems superintend these two different types of knowledge can explain why we intuitively know how to manipulate objects and how to recognize the environmental circumstances that afford their manipulation, even if our capability to report on our motor-intentional dispositions is disturbed. In fact, explicit judgment through a declarative stance involves an *a posteriori* reconstruction that can and often does rely on motor-intentional competences, but that is not motor-intentional in itself.

This explains why athletes are not always capable of telling us 'how they did it', especially when their motor-intentional abilities (silently stored as bodily readiness to act) made them capable of improvising new actions and unexpected strategies that are not reducible to verbal instructions or models received during training. If these two types of knowledge are not always equally accessible (because motor-intentional schemata are not available to analysis through conscious control and introspection), this is not only because the two systems were contingently separated; it may well be that there is a radical incommensurability between the two kinds of information encoded as motor intentionality and as cognitive intentionality, respectively. While know-that is based on detached, explicit representation of facts, know-how is intrinsically interactional and non-contentful in nature, i.e. devoid of propositional or figurative content; the former's normative structure is defined by conditions of satisfaction, the later by conditions of improvement (Dreyfus, 2002).

Due to its fundamentally non-representational nature, motor intentionality cannot be reduced to a collection of facts or bits of information expressed in a declarative format; motor intentionality builds on a direct practical and interactive acquaintance with the world, and emerges on an holistic, unquantifiable background of bodily and material preconditions that are essentially 'transcendent', i.e. always exceeding our possibilities of verbalization and conceptualization. If this picture is correct, then, although motor intentionality scaffolds the higher levels of our cognitive engagement with the world (verbalization, conceptualization), and provides the

embodied root of our intellectual categories, there would be no assurance that motor-intentional knowledge could ever be entirely accessible through conscious awareness, reflection, or introspection (and, even if it were accessible, that would happen only reconstructively, *a posteriori*). That is why, even if deploying intellectual faculties can certainly help during sport performance, higher forms of cognition and explicit thought are never enough to accomplish excellent results and cannot replace embodied expertise.

9. Conclusions

I think the analysis of the components of consciousness proposed by Birch *et al.* (in press) efficaciously proves that athletes are not zombies without qualitative experiences or automata deprived of decision, reflection, or control. Both the phenomenology of consciousness (at the personal level) and its functional components (at the subpersonal level) play a crucial role in the mental life of athletes when they engage with challenging tasks. We can consider this role as legitimately 'cognitive', in that it importantly involves embodied forms of cognition and emotional and practical intelligence.

At the same time, a careful analysis of performance under pressure and skilful activity in a flow state reveals that consciousness and its functional components are modulated in highly specific ways, with a complex interplay of background and foreground elements. This specificity reveals some important features of the basic architecture of our minds, including how we acquire and deploy complex sensorimotor skills governed by embodied forms of intelligence (e.g. motor intentionality). These basic forms of cognition are intelligent because they allow rich and flexible forms of adaptive behaviour and very fine-grained capabilities of perceptual discrimination, practical categorization, and responsiveness to context; at the same time, our analysis confirms that these forms of basic cognition are largely or fundamentally automatic and pre-reflective. Not only in the sense that they operate at a subpersonal level through the work of mechanisms situated well below the level of our consciousness, as correctly maintained by Birch and colleagues; but also in the sense that, whenever consciousness is involved in skilful performances, the fundamental features of conscious experience are specifically modified or partly suppressed. When the best performances have to be achieved, some particular forms of our awareness are adumbrated or selectively filtered to minimize attention to oneself and maximize the allocation of resources to the goals or expected consequences of the agent's actions, in order to increase her readiness and boost her reactivity to the circumstances.

I hope I have been able to persuasively argue that, studying the cognitive processes underlying the performances of the athletes, it is necessary to take seriously both the fact that athletes are aware and conscious and the fact that their conscious awareness during peak performance is characterized by the selective suppression or attenuation of particular patterns of perception and attention.

Acknowledgments

This research has been sponsored by a UPAR grant by UAE National Research Foundation 'Sport and Brain Science: Technological Applications for Peaking Performances' (grant code: 31H087-UPAR (3) 2014). The author is thankful to Shaun Gallagher, Jesse Prinz, and Dan Zahavi for commenting on an early version of this paper when it was presented for the first time at the Inter-University Center in Dubrovnik, in April 2013, and to Giacomo Bertacchi, Jens Birch, Geir Jordet, and Gunjan Khera for kindly reviewing its latest version.

References

- Araújo, D., Davids K. & Hristovski, R. (2006) The ecological dynamics of decision making in sport, *Psychology of Sport and Exercise*, **7**, pp. 653–676.
- Armstrong, D. (1981) What is consciousness?, in *The Nature of Mind, and Other Essays*, pp. 55–67, Ithaca, NY: Cornell University Press.
- Baumeister, R.F. (1984) Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance, *Journal of Personality and Social Psychology*, **46** (3), pp. 610–620.
- Beilock, S.L. (2008) Beyond the playing field: Sport psychology meets embodied cognition, *International Review of Sport and Exercise Psychology*, **1** (1), pp. 19–30.
- Beilock, S.L. (2011) *Choke*, Boston, MA: Little, Brown Book Group.
- Beilock, S.L. (2015) *How the Body Knows Its Mind: The Surprising Power of the Physical Environment to Influence How You Think and Feel*, New York: Atria Books.
- Beilock, S.L. & Carr, T.H. (2001) On the fragility of skilled performance: What governs choking under pressure?, *Journal of Experimental Psychology: General*, **130**, pp. 701–725.
- Beilock, S.L., Wierenga, S.A. & Carr, T.H. (2002) Expertise, attention, and memory in sensorimotor skill execution: Impact of novel task constraints on dual-task performance and episodic memory, *The Quarterly Journal of Experimental Psychology*, **55A** (4), pp. 1211–1240.
- Beilock, S.L., Bertenthal, B.I., McCoy, A.M. & Carr, T.H. (2004) Haste does not always make waste: Expertise, direction of attention, and

- speed versus accuracy in performing sensorimotor skills, *Psychonomic Bulletin & Review*, **11** (2), pp. 373–379.
- Beilock, S.L. & Carr, T.H. (2004) From novice to expert performance: Memory, attention and the control of complex sensorimotor skills, in Williams, A.M., Hodges, N.J., Scott, M.A. & Court, M.L.J. (eds.) *Skill Acquisition in Sport: Research, Theory and Practice*, pp. 309–328, London: Routledge.
- Beilock, S.L. & Carr, T.H. (2005) When high-powered people fail: Working memory and ‘choking under pressure’ in math, *Psychological Science*, **16** (2), pp. 101–105.
- Beilock, S.L. & Holt, L.E. (2007) Embodied preference judgments: Can likeability be driven by the motor system?, *Psychological Science*, **18** (1), pp. 51–57.
- Beilock, S.L., Bertenthal, B.L., Hoeger, M. & Carr, T.H. (2008) When does haste make waste? Speed-accuracy tradeoff, skill level, and the tools of the trade, *Journal of Experimental Psychology*, **14** (4), pp. 340–352.
- Beilock, S.L. & Gonso, S. (2008) Putting in the mind versus putting on the green: Expertise, performance time, and the linking of imagery and action, *The Quarterly Journal of Experimental Psychology*, pp. 1–13.
- Birch, J.E., Breivik, G. & Fusché Moe, V. (in press) Knowledge, consciousness and sporting skills, in Cappuccio, M. (ed.) *MIT Press Handbook of Embodied Cognition and Sport Psychology*, Cambridge MA: MIT Press.
- Block, N. (1995) On a confusion about a function of consciousness, *The Behavioral and Brain Sciences*, **18**, pp. 227–287.
- Bocanegra, B.R. & Hommel, B. (2014) When cognitive control is not adaptive, *Psychological Science*, **25** (6), pp. 1249–1255.
- Breivik, G. (2007) Skillful coping in everyday life and in sport; a critical examination of the views of Heidegger and Dreyfus, *Journal of Philosophy of Sport*, **34** (2), pp. 116–134.
- Breivik, G. (2013) Zombie-like or superconscious?: A phenomenological and conceptual analysis of consciousness in elite sport, *Journal of the Philosophy of Sport*, **40**, pp. 85–106.
- Brownstein, M. (2014) Rationalizing flow: Agency in skilled unreflective action, *Philosophical Studies*, **168** (2), pp. 545–568.
- Cappuccio, M. (ed.) (2015) Unreflective action and the choking effect, *Phenomenology and the Cognitive Sciences*, **14** (2).
- Cappuccio, M. (ed.) (in press) *MIT Press Handbook of Embodied Cognition and Sport Psychology*, Cambridge, MA: MIT Press.
- Cappuccio, M. & Wheeler, M. (2012) Ground-level intelligence: Inter-context frame problem and dynamics of the background, in Radman, Z. (ed.) *Knowing without Thinking: Mind, Action, Cognition*

- and the Phenomenon of the Background, Basingstoke: Palgrave Macmillan.
- Cartwright, D. & Harary, F. (1956) Structural balance: A generalization of Heider's theory, *Psychological Review*, **63** (5), pp. 277–293.
- Catley, D. & Duda, J.L. (1997) Psychological antecedents of the frequency and intensity of flow in golfers, *International Journal of Sport Psychology*, **28** (4), pp. 309–322.
- Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.
- Christensen, W., Sutton, J. & McIlwain, D.J.F. (2014) Putting pressure on theories of choking: Towards an expanded perspective on breakdown in skilled performance, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 253–293.
- Christensen, W., Sutton, J. & McIlwain, D.J.F. (2016) Cognition in skilled action: Meshed control and the varieties of skill experience, *Mind and Language*, **31** (1), pp. 37–66.
- Clark, A. (2015) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford: Oxford University Press.
- Craighero, L., Fadiga, L., Rizzolatti, G. & Umiltà, C. (1999) Action for perception: A motor-visual attentional effect, *Journal of Experimental Psychology: Human Perception and Performance*, **25** (6), pp. 1673–1692.
- Davids, K., Araújo, D., Hristovski, R., Passos, P. & Chow, J.Y. (2012) Ecological dynamics and motor learning design in sport, in Williams, M. & Hodges, N. (eds.) *Skill Acquisition in Sport: Research, Theory & Practice*, 2nd ed., pp. 112–130, London: Routledge.
- Dreyfus, H.L. (1996) The current relevance of Merleau-Ponty's theory of embodiment, *The Electronic Journal of Analytic Philosophy*, **4**.
- Dreyfus, H.L. (2002) Intelligence without representation—Merleau-Ponty's critique of mental representation, *Phenomenology and the Cognitive Sciences*, **1**, pp. 367–383.
- Dreyfus, H.L. (2008) Why Heideggerian AI failed and how fixing it would require making it more Heideggerian, in Huskies, P., Holland, O. & Wheeler, M. (eds.) *The Mechanical Mind in History*, pp. 331–371, Cambridge, MA: MIT Press.
- Dreyfus, H.L. (2012) The mystery of the background qua background, in Radman, Z. (ed.) *Knowing without Thinking: The Background in Philosophy of Mind*, pp. 1–10, Basingstoke: Palgrave Macmillan.
- Dreyfus, S.E. & Dreyfus, H.L. (1980) *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*, Washington, DC: Storming Media.
- Durgin, F.H., Klein, B., Spiegel, A., Strawser, C.J. & Williams, M. (2012) The social psychology of perception experiments: Hills, backpacks,

- glucose, and the problem of generalizability, *Journal of Experimental Psychology: Human Perception and Performance*, **38**, p. 1582.
- Edelman, G. (1992) *Bright Air, Brilliant Fire: On the Matter of the Mind*, New York: Basic Books.
- Engeser, S. & Rheinberg, F. (2008) Flow, performance and moderators of challenge-skill balance, *Motivation and Emotion*, **32**, p. 158.
- Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. (2000) Visuomotor neurons: Ambiguity of the discharge or 'motor' perception?, *International Journal of Psychophysiology*, **35** (2-3), pp. 165-177.
- Firestone, C. (2013) How 'paternalistic' is spatial perception? Why wearing a heavy backpack doesn't and couldn't make hills look steeper, *Perspectives on Psychological Science*, **8**, pp. 455-473.
- Fitts, P.M. & Posner, M.I. (1967) *Human Performance*, Belmont, CA: Brooks/Cole.
- Foerster, A., Gray, R. & Cañal-Bruland, R. (2015) Size estimates remain stable in the face of differences in performance outcome variability in an aiming task, *Consciousness and Cognition*, **33**, pp. 47-52.
- Freeman, W.J. (1991) The physiology of perception, *Scientific American*, **264**, pp. 78-85.
- Fridland, E. (2014) They've lost control: Reflections on skill, *Synthese*, **191**, art. 2729.
- Furley, P., Schweizer, G. & Bertrams, A. (2015) The two modes of an athlete: Dual-process theories in the field of sport, *International Review of Sport and Exercise Psychology*, **8** (1), pp. 106-124.
- Gallagher, S. & Zahavi, D. (2008) Phenomenological approaches to self-consciousness, in *Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/entries/self-consciousness-phenomenological/>.
- Gallese, V. (2007) The shared manifold hypothesis: Embodied simulation and its role in empathy and social cognition, in Farrow, T.F.D. & Woodruff, P.W.R. (eds.) *Empathy in Mental Illness*, pp. 448-472, Cambridge: Cambridge University Press.
- Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*, Boston, MA: Houghton Mifflin.
- Gottlieb, G. (2015) Know-how, procedural knowledge, and choking under pressure, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 361-378.
- Gray, R. (2014) Embodied perception in sport, *International Review of Sport and Exercise Psychology*, **7** (1), pp. 72-86.
- Hill, D.M. & Shaw, G. (2013) A qualitative examination of choking under pressure in team sport, *Psychology of Sport and Exercise*, **14** (1), pp. 103-110.

- Hill, D.M., Hanton, S., Matthews, N. & Fleming, S. (2010) Choking in sport: A review, *International Review of Sport and Exercise Psychology*, **3** (1), pp. 24–39.
- Holt, L.E. & Beilock, S.L. (2006) Expertise and its embodiment: Examining the impact of sensorimotor skill expertise on the representation of action-related text, *Psychonomic Bulletin & Review*, **13** (4), pp. 694–701.
- Hutto, D.D. & Sánchez-García, R. (2015) Choking RECTified: Embodied expertise beyond Dreyfus, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 309–331.
- Ikegami, T. & Ganesh, G. (2014) Watching novice action degrades expert motor performance: Causation between action production and outcome prediction of observed actions by humans, *Scientific Reports*, **4**, 6989.
- Jackson, R.C. & Beilock, S. (2008) Attention and performance, in Farrow, D., Baker, J. & MacMahon, C. (eds.) *Developing Elite Sports Performers: Lessons from Theory and Practice*, New York: Routledge.
- Jackson, R.C., Ashford, K.J. & Norsworthy, G. (2006) Attention focus, dispositional reinvestment, and skillful motor performance under pressure, *Journal of Sport and Exercise Psychology*, **28**, pp. 49–68.
- Jackson, S.A. (1996) Toward a conceptual understanding of the flow experience in elite athletes, *Research Quarterly for Exercise and Sport*, **67** (1), pp. 76–90.
- Jackson, S.A. & Marsh, H.W. (1996) Development and validation of a scale to measure optimal experience: The Flow State Scale, *Journal of Sport and Exercise*, [Online], <http://fitnessforlife.org/AcuCustom/Sitename/Documents/DocumentItem/8983.pdf>
- Jackson, S.A. & Csikszentmihalyi, M. (1999) *Flow in Sports*, Champaign, IL: Human Kinetics.
- Jordet, G. & Hartman, E. (2008) Avoidance motivation and choking under pressure in soccer penalty shootouts, *Journal of Sport & Exercise Psychology*, **30**, pp. 450–457.
- Kelly, S.D. (2000) Grasping at straws: Motor intentionality and the cognitive science of skillful action, in *Essays in Honor of Hubert Dreyfus*, vol. II, Cambridge, MA: MIT Press.
- Kelly, S.D. (2002) Merleau-Ponty on the body, *Ratio*, **15** (4), pp. 376–391.
- Komar, J., Chow, J.-Y., Chollet, D. & Seifert, L. (2014) Effect of analogy instructions with an internal focus on learning a complex motor skill, *Journal of Applied Sport Psychology*, **26** (1), pp. 17–32.
- Lewis, D. (1980) Mad pain and Martian pain, in Block, N. (ed.) *Readings in the Philosophy of Psychology*, vol. I, pp. 216–222, Cambridge, MA: Harvard University Press.

- Marbach, E. (2010) Is there a metaphysics of consciousness without a phenomenology of consciousness? Some thoughts derived from Husserl's philosophical phenomenology, *Royal Institute of Philosophy Supplement*, **67**, pp. 141–154.
- Masters, R.S.W. (1992) Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure, *British Journal of Psychology*, **83** (3), pp. 343–358.
- Masters, R.S.W., van der Kamp, J. & Jackson, R.C. (2007) Imperceptibly off-center: Goalkeepers influence penalty-kick direction in soccer, *Psychological Science*, **18** (3), pp. 222–223.
- Merleau-Ponty, M. (1945) *Phénoménologie de la perception*, Paris: Gallimard.
- Mesagno, C. (ed.) (2013) Performance under pressure, *International Journal of Sport Psychology*, **44** (4).
- Mesagno, C., Harvey, J.T. & Janelle, C.M. (2012) Choking under pressure: The role of fear of negative evaluation, *Psychology of Sport and Exercise*, **13** (1), pp. 60–68.
- Mesagno, C. & Hill, D.M. (2013a) Choking under pressure debate: Is there chaos in the brickyard?, *International Journal of Sport Psychology*, **44** (4), pp. 288–293.
- Mesagno, C. & Hill, D.M. (2013b) Definition of choking in sport: Reconceptualization and debate, *International Journal of Sport Psychology*, **44** (4), pp. 267–277.
- Milner, A.D. & Goodale, M.A. (2008) Two visual systems re-viewed, *Neuropsychologia*, **46** (3), pp. 774–785.
- Moll, T., Jordet, G. & Pepping, G.-J. (2010) Emotional contagion in soccer penalty shootouts: Celebration of individual success is associated with ultimate team success, *Journal of Sports Sciences*, **28** (9), pp. 983–992.
- Montero, B.G. (2014) Is monitoring one's actions causally relevant to choking under pressure?, *Phenomenology and the Cognitive Sciences*, **14** (2), pp. 379–395.
- Montero, B.G. (2016) *Thought in Action*, Oxford: Oxford University Press.
- Montero, B.G., Toner, J. & Moran, A. (in press) Questioning the breadth of the attentional focus effect, in Cappuccio, M. (ed.) *MIT Press Handbook of Embodied Cognition and Sport Psychology*, Cambridge, MA: MIT Press.
- Noël, B., van der Kamp, J. & Memmert, D. (2015) Implicit goalkeeper influences on goal side selection in representative penalty kicking tasks, *PLoS ONE*, **10** (8), e0135423.

- Radman, Z. (ed.) (2012) *Knowing without Thinking: Mind, Action, Cognition and the Phenomenon of the Background*, Basingstoke: Palgrave Macmillan.
- Rietveld, E. (2008) Situated normativity: The normative aspect of embodied cognition in unreflective action, *Mind*, **117** (468), pp. 973–1001.
- Rizzolatti, G., Fogassi, L. & Gallese, V. (2001) Neurophysiological mechanisms underlying understanding and imitation of action, *Nature Reviews Neuroscience*, **2**, pp. 661–670.
- Rizzolatti, G. & Craighero, L. (2010) Premotor theory of attention, *Scholarpedia Journal*, [Online], http://scholarpedia.org/article/Premotor_theory_of_attention
- Rizzolatti, G. & Umiltà, M.A. (2013) Canonical neurons, in Runehov, A.L.C. & Oviedo, L. (eds.) *Encyclopedia of Sciences and Religions*, pp. 305–305, Dordrecht: Springer.
- Royal Institute of Philosophy Supplement, **85** (67), pp. 141–154.
- Searle, J. (1995) *The Mystery of Consciousness*, New York: New York Review of Books.
- Silva, P., Kiefer, A., Riley, M. & Chemero, A. (in press) Trading perception and action for complex cognition: Application of ecological theory to the design of interventions for skill learning, in Cappuccio, M. (ed.) *MIT Press Handbook of Embodied Cognition and Sport Psychology*, Cambridge, MA: MIT Press.
- Slingerland, E. (2014) *Trying Not to Try: The Art and Science of Spontaneity*, New York: Crown.
- Sutton, J. (2007) Batting, habit and memory: The embodied mind and the nature of skill, *Sport in Society*, **10** (5), pp. 763–786.
- Swann, C. (2016) Flow in sport, in Harmat, L., Andersen, F. Ø., Ullén, F., Wright, J. & Sadlo, G. (eds.) *Flow Experience: Empirical Research and Applications*, pp. 51–64, Berlin: Springer.
- Swann, C., Keegan, R.J., Piggott, D. & Crust, L. (2012) A systematic review of the experience, occurrence, and controllability of flow states in elite sport, *Psychology of Sport and Exercise*, **13** (6), pp. 807–819.
- Swann, C., Keegan, R., Crust, L. & Piggott, D. (2016) Psychological states underlying excellent performance in professional golfers: ‘Letting it happen’ vs. ‘making it happen’, *Psychology of Sport and Exercise*, **23**, pp. 101–113.
- Thomas, B. (2005) *Bruce Lee: Fighting Words*, Madison, WI: Frog Books.
- Thompson, E. (2007) *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Cambridge, MA: Harvard University Press.
- Thompson, E. (2014) Dreamless sleep, the embodied mind, and consciousness, *Open MIND*, Johannes Gutenberg Universität Mainz.

- Toner, J., Montero, B.G. & Moran, A. (2014) Considering the role of cognitive control in expert performance, *Phenomenology and the Cognitive Sciences*, **14** (4), pp. 1127–1144.
- Toner, J., Montero, B.G. & Moran, A. (2015) The perils of automaticity, *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, **19** (4), pp. 431–442.
- Toner, J., Montero, B.G. & Moran, A. (2016) Reflective and prereflective bodily awareness in skilled action, *Psychology of Consciousness: Theory, Research, and Practice*, [Online], <http://dx.doi.org/10.1037/cns0000090>
- Varela, F., Thompson, E. & Rosch, E. (1991) *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MA: MIT Press.
- Wheeler, M. (2008) Cognition in context: Phenomenology, situated robotics and the frame problem, *International Journal of Philosophical Studies*, **16** (3), pp. 323–349.
- Wilson, R.A. & Foglia, L. (2011) Embodied cognition, in *Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/entries/emobodied-cognition/>.
- Wine, J. (1971) Test anxiety and direction of attention, *Psychological Bulletin*, **76**, pp. 92–104.
- Witt, J.K. (2015) Awareness is not a necessary characteristic of a perceptual effect: Commentary on Firestone (2013), *Perspectives on Psychological Science*, **10**, pp. 865–872.
- Witt, J.K. & Proffitt, D.R. (2005) See the ball, hit the ball: Apparent ball size is correlated with batting average, *Psychological Science*, **16**, pp. 937–938.
- Witt, J.K., Linkenauger, S., Bakdash, J. & Proffitt, D. (2007) Golf performance makes the hole look as big as a bucket or as small as a dime, *Journal of Vision*, **7** (9), p. 291.
- Witt, J.K. & Dorsch, T.E. (2009) Kicking to bigger uprights: Field goal kicking performance influences perceived size, *Perception*, **38** (9), pp. 1328–1340.
- Wulf, G. (2013) Attentional focus and motor learning: A review of 15 years, *International Review of Sport and Exercise Psychology*, **6** (1), pp. 77–104.
- Wulf, G. (2015) An external focus of attention is a conditio sine qua non for athletes: A response to Carson and Collins (2015), *Journal of Sports Sciences*, **34** (13), pp. 1293–1295.
- Wulf, G. (2016) Why did Tiger Woods shoot 82? A commentary on Toner and Moran (2015), *Psychology of Sport and Exercise*, **22**, pp. 337–338.

- Wulf, G., Lauterbach, B. & Toole, T. (1999) Learning advantages of an external focus of attention in golf, *Research Quarterly for Exercise and Sport*, **70**, pp. 120–126.
- Wulf, G., Shea, C.H. & Park, J.-H. (2001) Attention in motor learning: Preferences for and advantages of an external focus, *Research Quarterly for Exercise and Sport*, **72**, pp. 335–344.
- Wulf, G., McConnel, N., Gärtner, M. & Schwarz, A. (2002) Enhancing the learning of sport skills through external-focus feedback, *Journal of Motor Behavior*, **34** (2), pp. 171–182.
- Wulf, G. & Su, J. (2007) An external focus of attention enhances golf shot accuracy in beginners and experts, *Research Quarterly for Exercise and Sport*, **78**, pp. 384–389.
- Zahavi, D. (2005) *Subjectivity and Selfhood: Investigating the First-Person Perspective*, Cambridge, MA: MIT Press.
- Zahavi, D. (2012) Mindedness, mindlessness and first-person authority, in Shear, J. (ed.) *Mind, Reason and Being-in-the-World: The McDowell-Dreyfus Debate*, London: Routledge.

Support Publication #6 [SP6]

Cappuccio M.L., Gray R., Hill D.M., Mesagno C., Carr T.H. (2019),
“The Many Threats of Self-Consciousness:
Embodied Approaches to Choking Under Pressure in Sensorimotor Skills”
in Massimiliano L. Cappuccio (ed.), *Handbook of Embodied Cognition and
Sport Psychology*, the MIT Press, 2019.

Reproduced with permission from the MIT Press. All rights reserved.

5 The Many Threats of Self-Consciousness: Embodied Approaches to Choking under Pressure in Sensorimotor Skills

Massimiliano L. Cappuccio, Rob Gray, Denise M. Hill,
Christopher Mesagno, and Thomas H. Carr

1 Introduction

The phenomenon of choking under pressure (hereafter called “choking”) has continued to fascinate sport practitioners, researchers (Beilock 2011), and people from many other backgrounds and pursuits. Each year seems to bring a new well-publicized example labeled a “choke,” with perhaps the most recent being Jordan Spieth surrendering a five-stroke lead on the final nine holes of the 2016 US Masters golf tournament (O’Connor 2016). In this chapter we offer a critical review of the scientific research and the philosophic evaluations that bear on this very phenomenon.

According to Baumeister (1984), choking epitomizes a paradox many practitioners are familiar with: the situations that overemphasize the expectation to produce excellent results are those that most likely create below-standard performances. Choking seems to occur exactly when one desires the most to succeed, and that is why it is often feared by expert athletes, like a curse or a malicious spell. This phenomenon is usually construed by psychologists as the expert athletes’ incapability to successfully perform actions that would otherwise be routine because of psychological factors such as increased desire to do well, increased anxiety and tension, and the changes in performance processes resulting from such increases (Carr 2015). However, the exact definition of choking varies significantly across different schools (reviewed by Christensen, Sutton, and McIlwain 2015), and there is no universal agreement about its nature and causes (e.g., Mesagno and Hill 2013), which relate to a multifaceted complexity of cognitive, motivational, and emotional dynamics that have been studied largely, but ultimately without consensus, by sport psychologists (Hill et al. 2010a; Mesagno, Guekes, and Larkin 2015).

This does not mean that no progress was ever achieved by psychological research or that our knowledge of choking has not improved. In fact, over the past thirty years,

—1
—0
—+1

researchers have made several advances, including designing clever manipulations for creating pressure in the lab, developing theories about the cognitive processes that underlie choking, and using multi-method approaches that assess gaze behavior, movement kinematics, or brain activity along with performance outcomes (reviewed in Beilock and Gray 2007; Hill et al. 2010b; Gray 2011).

On the surface, these research efforts have led to a broadly accepted explanation for choking in sports (and sensorimotor skills more generally), an explanation that permeates both research journals and popular culture: namely, that choking occurs when an experienced athlete or skilled performer becomes “self-conscious” (Baumeister 1984). Elaborating on Baumeister’s idea, this chapter aims to describe various levels and forms of self-consciousness (perceptual, emotional, motivational, psychological), exploring how they causally link to performance disruption and its underlying psychological dynamics.

Our account does not hide or minimize the difficulties associated with a reductive interpretation of the causal relationship between self-consciousness and performance: although broadly accepted, the explanation of choking based on Baumeister’s notion of self-consciousness has also attracted a number of important and well-argued criticisms. If this chapter combines different views and discourses about choking, it is exactly to demonstrate that most of these objections apply only to the “one size fits all” approach that for too long has dominated choking research, propagating a reductive interpretation of the notion of self-consciousness.

We contend that such interpretation is misleading and does not capture the real potential contained in Baumeister’s key intuition. Self-consciousness is not a monolithic, universal causal factor of performance disruption (because self-consciousness comes in different forms and degrees), and certainly is not the kind of factor that could *directly* cause performance disruption (because strategic thought and various regulatory processes are interposed between self-consciousness and the performance outcome). In other words, the way in which self-consciousness affects expert skills does not depend on a single linear causal chain, but is articulated at multiple parallel levels and can take several causal pathways depending on various possible interactions between intermediate variables.

A pluralistic and ecumenic approach is therefore indispensable to recognizing how multifaceted and layered the causal relationship between self-consciousness and performance disruption is. Ignoring this complexity, the bulk of previous research has failed to account for individual traits, perceptions, capabilities, and limitations of performers when understanding how they respond to pressure. Our complaint is best exemplified by how researchers identify whether “a choke” has occurred: namely, identification

-1—
0—
+1—

is based mainly, often purely, on extrinsic, objective criteria (there is some form of pressure present and the performance level is below the “typical” non-pressure level) without taking into account the individual characteristics of the performer, the phenomenology of his or her performative experience, or the situated circumstances of her or his actions (Mesagno and Hill 2013; Mesagno and Beckmann 2017).

The reductive, linear interpretation of self-consciousness as a universal trigger mechanism for choking stands in stark contrast to the rapidly growing “embodied” approach to perception and cognition in sport (reviewed in Gray 2014). The embodied approach to cognition (Varela, Thompson, and Rosch 1991) utilizes notions and methods that sport psychology research inherited from the traditions of ecological psychology (with the Gibsonian concepts of “affordances” and “direct perception”; see Gibson 1966, 1979) and developmental psychology (with the preeminence attributed to “enactive learning” by the Piagetian theories of early cognitive development, e.g., Bruner 1966). Embodied cognition applies these notions to explain how the reach and the very nature of mental functions largely depend on the details of their physical implementation and by the conditions of engagement and interaction offered to a situated agent by her or his environment. This means emphasizing the constitutive role played by fundamental forms of cognition (i.e., perceptual and motoric habits based on the consolidation of sensorimotor feedback loops) in scaffolding and shaping higher, more intellectual forms of cognition (i.e., propositional and conceptual competences based on explicit judgment and inference, logico-symbolic and linguistic representation, etc.)

In sport psychology, the embodied approach to cognition highlights that even “low-level” perceptual processes, like perceiving the size of a ball, seem to influence—and, reciprocally, to be influenced by—the individual capabilities and current state of the athlete: the perceptual, physical, emotional, and motivational condition of the athlete, as well as his epistemic, experiential, and cognitive situation.

We believe an embodied approach to choking can help see how the mechanisms that disrupt performance through self-consciousness are not just linear and top-down (as if choking was always originated by erroneous meta-representations and unnecessary beliefs accumulated in the mind of the athlete), but tend to emerge contextually, from the reciprocal interaction of complex and heterogeneous contingencies that depend largely on the athlete’s perception of her own body and the environment in which the athlete with her body is situated (DeCaro et al. 2011 offers a similar suggestion). These considerations extend, for example, to how the current condition of the performer—on a roll or in a slump (Gray and Allsop 2013), tired or full of energy (Schnall, Zadra, and Proffitt 2010), physically burdened or feeling free and light (Bhalla and Proffitt 1999)—might matter to performance.

—1
—0
—+1

“This is embodied cognition in the raw,” say Christensen, Sutton, and McIlwain (2015, 254), referring to a dramatic and unexpected episode of performance impairment in cricket, “with emotion and personality, kinesthesia and physiology, discipline and drives all affected together and on show” (254). In this spirit, we propose it is time to “embody” choking research by considering not just the perceptual and attentional dimension of self-consciousness that a performer brings to sensorimotor activity, but also the emotional, motivational, and narrative dimensions of self-consciousness and the particular ways in which each of these dimensions separately, and all of them as a whole, influence how the performer perceives and handles pressure.

For a long time, the research on choking has been dominated by the scientific approaches that impute performance disruption to the misallocation of the attentional resources available to the system (for a review, see Jackson and Beilock 2008). In the last few years, however, various explanatory narratives have focused on individual experiential circumstances, team dynamics, motivational backgrounds, and personal psychological profiles (e.g., Jordet 2009, 2010; Mesagno, Harvey, and Janelle 2012; Mesagno and Hill 2013; Carr 2015; Geukes et al. 2017). We recognize the complementarity of these approaches, and the deep interconnection between the causal factors that they study.

That is why, after some preliminary clarifications (part 2), the first half of the chapter (parts 3 to 7) reviews the experiments suggesting that performance disruption is engendered by perceptual and motoric self-consciousness (i.e., the condition in which the athlete overscrutinizes certain aspects of his own movements and actions, and their control, in the attempt to improve them), and the theories that try to make sense of these empirical results. The second half of the chapter (parts 8 to 14) explores a different aspect of self-consciousness, linked to how the psychological, motivational, and emotional backdrop of the athlete’s experience under pressure influences self-presentation and self-esteem—an influence that can negatively affect performance through the mediation of disruptive personal narratives and social constructions of identity.

2 Some Preliminary Clarifications about Consciousness and Expertise

Baumeister (1982, 1984) identified a correlation between the performers’ tendency to choke and their “dispositional self-consciousness.” This full-fledged notion of self-consciousness aims primarily to describe the general inclination to see oneself as a socially evaluable object, or as psychological susceptibility to public judgment (see Mesagno, Harvey, and Janelle 2011). But Baumeister’s theory can be applied to and account for multiple, partial dimensions of self-consciousness, including not only the sphere of personal identity and the intersubjective dynamics underlying narrative

-1—
0—
+1—

self-construction, but also more fundamental mechanisms of emotional awareness and reflective control of action. We recognize the mark of self-consciousness in every situation in which a cognitive system engaged in a challenging task seeks to gain explicit personal-level knowledge about its own psychophysical processes, states, and components: this happens when the system closely monitors its operations to access on-line information considered necessary to improve the precision and volume of its behavioral output.

This notion is a valid guide for a comprehensive study of performance disruption. Any notion of self-consciousness less multifaceted and articulated would be unable to capture all the nuanced meanings of skill and would inevitably fail to account for a complex phenomenon like choking. That is why rationalistic and aprioristic notions of “self-consciousness,” mesmerized by the apparent contrast between abstract categories such as “thought” and “action,” have recently attracted objections and criticisms. In particular, Dreyfus’s model of skill acquisition (Dreyfus and Dreyfus 1980, often likened to Fitts and Posner’s own model, 1967) recently became the polemical target of an interdisciplinary controversy concerning whether or not reflective awareness is a necessary and/or sufficient condition for choking, and whether expert, skillful execution of action allows explicit control or not (see Cappuccio 2015 for a collection of critical views; see also Breivik 2007, 2013; Montero 2010; Sutton et al. 2011; Toner, Montero, and Moran 2014, 2015).

For a long time, Dreyfus’s model has represented the most influential phenomenological account of expert action as an unreflective, semiconscious, embodied routine. According to Dreyfus’s model, a practitioner becomes an expert if and only if the explicit cognitive efforts necessary to control her actions decrease, when the execution of complex movements is automatized. Dreyfus has mainly been criticized for his nonrepresentational theory of unreflective action, or “absorbed coping,” which is a corollary of his view on skill development as progressive automatization of motor habits (Dreyfus 2002). Absorbed coping explicitly states that proceduralized skillful activities not only *can* (characteristically), but *must* (necessarily) be “mindless” and “non-conscious,” and that a state of decreased self-awareness is indispensable to deliver expert performance. That is why absorbed coping theory is associated with the categorical claim that conscious control (“reflective thought”) and expert performance (“skillful action”) are mutually exclusive. Despite the sophisticated anti-representationalist argumentation provided by Dreyfus, and his original phenomenology of embodiment (Dreyfus 2012), this claim is considered unpersuasive by many practitioners, psychologists, and philosophers, as it seems to deny the reflective dimension of expertise, which typically includes skillful capabilities of strategic planning and creative improvisation.

—1
—0
—+1

The debate concerning absorbed coping¹ is relevant to our discussion of choking insofar as some of the criticisms raised against Dreyfus's theory have been extended to Baumeister's, which represents the common denominator of the analyses conducted in this chapter. In fact, both Dreyfus's and Baumeister's theories can be categorized as "self-focus" approaches to performance disruption, as they agree that some forms of awareness of one's own condition and processes can damage performance. However, the two theories agree neither on the particular forms of self-consciousness responsible for choking, nor on the underlying cognitive mechanisms that cause skill disruption. Appreciating these nuanced differences is fundamental to correctly construing the complex relationship between peak performance and self-consciousness. What Christensen, Sutton, and McIlwain (2015) wrote about this relationship is insightful, and should be agreed on by all the supporters of Baumeister's theory:

Skill experience is rich and various, and ... there are many forms of self-focus and self-awareness. Some of these can be damaging while other forms may be essential to good performance for some people in some contexts.... There are forms of "self-focus" which are distinct from body-focused attention, from technique-focused attention, and from skill or task focus. Likewise, ... not all forms of embodied self-awareness need involve full-scale self-consciousness. (255)

Accordingly, the primary goal of the researchers investigating the causes of choking should be to concretely distinguish between various types and degrees of self-awareness in skillful activity and empirically identify which attentional foci are detrimental to optimal action, when, and why. Such concrete analyses, more than abstract inferences based on an aprioristic definition of consciousness, can shed light on the complex, articulated involvement of explicit reflective knowledge in the impairment of skillful performance.

These analyses start with sensorimotor control: that is, the most fundamental embodied dimension of self-consciousness. That is why in the next sections we introduce an approach to skill disruption inspired by Baumeister's theory that—we believe—can effectively account for how motoric and perceptual self-consciousness typically causes underperformance. This approach seems immune from many of the criticisms that absorbed coping theory had to face because, unlike Dreyfus's theory, it presupposes neither that expert performance is necessarily "mindless" (or the performer unaware of her actions), nor that proceduralized execution necessarily excludes all forms of reflection and conscious control. On the contrary, we will see that the automatic execution of proceduralized action routines can be implemented during peak performance with the concurrent involvement of reflective strategic control, consistently with a dual cognitive architecture governed by virtually independent causal and informational systems (as theorized, for example, by Evans and Stanovich 2013).

3 Self-Consciousness as the Root of Sensorimotor Skill Disruption: Attentional Focus Theory and Execution Focus Theory

Let us investigate more closely the perceptual and motoric dimension of self-consciousness (i.e., the immediate awareness of one's own actions as they are executed). This dimension is captured by two questions that are considered crucial by both practitioners and researchers: What sensorimotor aspects of one's performance are relevant to choking? And, why does monitoring them too closely disrupt one's skills? A branch of the research on choking builds on Baumeister's key intuition to answer these questions. This branch has produced at least three well-elaborated theories, which have gained wide credence among sport psychologists during the last decades.

1. According to *reinvestment theory*, proposed by Masters and colleagues (Masters 1992; Masters, Polman, and Hammond 1993), "Automated motor processes can be disrupted if they are run using consciously accessed, task-relevant declarative knowledge to control the mechanics of the movements on-line" (Masters and Maxwell 2008, 160).
2. Following the lead of Masters's seminal investigation, *execution focus theory* (EFT), proposed by Sian Beilock and colleagues, asserts that choking is caused by a "skill focus" or "execution focus" (e.g., Beilock and Carr 2001; Beilock et al. 2002; Beilock, Bertenthal, et al. 2008), which occurs when the athletes implement the cognitive goal structure of the task to be accomplished through a step-by-step level of explicit control.
3. According to *attentional focus theory* (AFT), proposed by Gabriele Wulf and colleagues, performance disruption is caused by an "internal focus" of attention (e.g., Wulf et al. 2000; Wulf and Prinz 2001), which occurs when the athletes exert explicit control over their body parts movements.

Here we focus only on AFT and EFT; reinvestment theory will be examined in part 13 (dedicated to the personality and psychology of frequent "chokers"). We must remark that, while both AFT and EFT impute performance disruption to explicit attention to one's own performance, some subtle differences characterize these two theories of performance disruption, not without implications for the etiology, the diagnosis, and the prognosis of choking.

To begin with, AFT and EFT do not fully agree on what counts as "self-components" of one's performance, and what those components are to be contrasted with. For AFT, internal foci comprise one's own body's movements, one's own body parts, and in general anything "near" the performing subject and close to the production of the movement; in turn, external foci comprise any unintentional effect of the movement,

—1
—0
—+1

the movement's target or intended transformation in the environment, the tools or effectors moved by one's body to achieve one's goal and, in general, anything that is "far" from the performing subject and beyond the direct production of the movement, including elements in the environment that are not affected by one's actions. Note that, in some cases, the subject's whole body counts as an external focus, while singled-out body parts count as internal focus.

EFT, in contrast, claims that what damages performance is attention to execution of the component movements: that is, any form of explicit control that, analyzing a complex action into its constitutive sensorimotor elements, hinders the fluidity and automaticity of well-practiced routines. These two definitions often match, but not always: for example, according to AFT, whereas attention to wrist and shoulder movements during a golf-putting task is definitely an internal focus of attention (and as such is likely to decrease performance), attention to the tool (putter) used during the same task counts as an external focus, because the swing movement of the club is just an effect of one's body movements (Wulf et al. 2000; Wulf 2013). On the contrary, according to EFT, both attention to the body parts and attention to the club swing represent an execution focus, as they both imply an analysis of the constitutive elements of the familiar putting routine.²

Another difference between EFT and AFT is that they apply to partially different sets of activities and performers. For AFT, internal focus is universally detrimental to performance (Wulf 2015), as it systematically disrupts one's actions: (1) during both practice/learning and peak-performance conditions; (2) in both novices and experts (regardless of whether the activity has been fully automatized or not); and (3) independent of the degree of involvement of working memory.

EFT is more nuanced, as it claims that an execution focus is more detrimental than a distracting cue or secondary task only to experts in peak-performance conditions (not to novices rehearsing during practice or trying to perform as well as they can in a test), and only for automatized sensorimotor activities that do not heavily rely on working memory. EFT distinguishes between cognitive systems that do and do not utilize working memory to produce their output: after the proceduralization of the related sensorimotor skill (e.g., golf putting), the production of highly automatized sensorimotor routines involves no working memory; on the contrary, concurrent secondary tasks (e.g., memorizing/recognizing certain tones in a series) might involve a high usage of working memory. Regardless of their respective difficulty, the two tasks recruit two virtually independent systems that do not interfere with one another, as they utilize heterogeneous types of cognitive resources.

-1—
0—
+1—

Accordingly, per EFT, the notion of choking as self-consciousness does not apply in the same way to skilled activities that necessarily require working memory to function properly, such as mathematical computation (Beilock, Bertenthal, et al. 2004; Beilock 2008b): in the specific case of dual tasks involving *only* activities of this kind, performance can be damaged more by distraction or cognitive overload than execution focus, because both tasks (whether involving automatized routines or not) draw cognitive resources from the same basin and will subsequently suffer from its depletion.

4 The Self-Monitoring/Execution-Focus Approach

In spite of these differences, the explanatory stories provided by AFT and EFT overlap significantly. Both theories (as well as reinvestment theory) can be considered varieties of an overarching approach to choking that we call “self-monitoring/execution-focus.” This approach, broadly construed, maintains that control of performance execution by a well-practiced and highly integrated procedure (i.e., an automated motor program that can do the job on its own, without explicit intentional intervention on its operations) should be preferred to both internal focus and execution focus—when such a procedure is available: a well-practiced procedure is generally more fluent and reliable, presents fewer choice points at which the flow of performance can be interrupted or misdirected, and hence is less likely to lead to choking.

The mechanisms deemed important by the self-monitoring/execution-focus approach to choking are illustrated by Castaneda and Gray (2007) in a study of baseball batting. This study documented how, in a batting simulation, the performances of college baseball players were disrupted by four different attentional strategies, corresponding to four dual-task conditions: two strategies involve attention to skill execution (attention to one’s own hand movements and attention to one’s bat’s movements, respectively), while the other two strategies involve attention to the external environment (attention to a series of tones, unrelated to the batting task, and attention to when the ball leaves the bat, respectively). The performance of expert players was damaged significantly more by the first two kinds of instructions, while the novices suffered more in the other two conditions. According to the authors, the study proves “that the optimal focus of attention for highly skilled batters is one that does not disrupt proceduralized knowledge and permits attention to the perceptual effect of the action, whereas the optimal focus of attention for less-skilled batters is one that allows attention to the step-by-step execution of the swing [including both attention to body movements and attention to the tool being moved]” (60).

—1
—0
—+1

Thus, for an experienced athlete attempting to perform his or her absolute best when executing a highly practiced act of sensorimotor skill under pressure, attention to self/execution (control and movement of the specific body parts involved in carrying out the action, or specific motions of a tool being used to implement the action) is less detrimental to performance than attention directed beyond the body parts moving and the tool being used (like the movement of the water generated by one's swimming). That is, if attention is directed toward the perceptual consequences of executing the action (what changed in the external environment that is relevant to my ultimate goal, and did it change for the better or for the worse?), then the damage to performance is smaller or less likely to occur. And (perhaps more surprising), even attention directed toward something completely irrelevant simply to occupy the conscious, intentional mind (e.g., singing a song to oneself, monitoring through headphones for a target in a stream of tones or words), thereby keeping conscious interventions and attempts at control occupied and out of the way of the proceduralized motor program, is less detrimental than attention to one's own bodily movements or movements of the tool one is using (as examples, see the experimental evidence collected by Mesagno, Marchant, and Morris 2009, testing for musical competences, and Gray 2004, testing for tones recognition).

5 An Opposite Explanation of Choking: The Distraction/Overload Approach versus the Self-Monitoring/Execution-Focus Approach

The self-monitoring/execution-focus approach is not the only cognitive theory of choking based on misallocation of attention: the main alternative to it within the same domain of explanation is represented by the distraction/overload approach to choking. These two approaches represent almost completely antithetical paradigms, as they respectively characterize choking as a form of *self-monitoring* (and a resulting focus on step-by-step *execution*: see Baumeister 1984; Beilock and Carr 2001; Wulf and Prinz 2001; Beilock, Kulp, et al. 2004) or a form of *distraction* (typically associated with *overload* of cognitive capacities, see Wine 1971). As previously mentioned, the self-monitoring/execution-focus approach imputes choking to the disruption of habitual motor routines that would be fluid and spontaneous if they were not hindered by the explicit, reflective control of the details of the execution. The distraction/overload approach, in contrast, imputes the negative effects of choking to the intrusion of external competing attentional stimuli or the overflow of demands on cognitive resources (caused, for example, by recurrent negative thoughts or worries). These two models differ remarkably because they identify opposite causes for choking: that is, either the

-1—
0—
+1—

excess or the lack of attention to action execution, respectively. But they also differ because they recommend contrary remedies to choking, presuppose discordant phenomenologies of skillful experience, and generate conflicting predictions on performance disruption.

One of the reasons both approaches are still popular is that each of them seems capable of telling sport scientists at least a part of the whole story about choking. The predictions generated by each approach within its own frame of reference have been tested empirically and confirmed in controlled laboratory settings using quantitative methodologies. We focus mainly on the studies supporting the self-monitoring/execution-focus model of choking, as they are particularly numerous and compelling and have been representing a significant stimulus for the whole scientific debate of the last fifteen years. Most of this debate revolved around the repeated attempts to either confirm or disprove the laboratory research that Beilock and her colleagues have conducted to investigate the formation and the application of the habitual skills that—per their theory—choking disrupts (for a review, see Jackson, Beilock, and Kinrade 2013).

In particular, three major empirical results collected by the researchers corroborate the self-monitoring/execution-focus model while simultaneously exposing the limits of the distraction/overload theory. These results suggest that performance disruption of highly practiced sensorimotor skills correlates more with the extent of attempts to explicitly control and/or monitor one's own component movements than with the degree of cognitive effort or working memory usage requested by the task.

1. Skill-focus instructions (i.e., attention instructions that emphasize the component processes of the performance, such as “keep the knees bent”), as well as skill-focused dual tasks (i.e., dual-task conditions in which the secondary task demands attentive analysis of the primary motor task, as under instructions like “keep the club head straight as you swing and say ‘straight’ when the club hits the ball” or “say when the swing movement is finished”) damage the performance of experts much more than other kinds of instructions or dual-task conditions that do not solicit execution-monitoring (e.g., golf putting while trying to identify each occurrence of a target tone or word in a series played over headphones, see, e.g., Beilock and Carr 2001). This effect is not accounted for by distraction/overload theories of choking: such theories justify the opposite prediction, as they suggest that athletes are more likely to underperform when external stimuli compete with the task at hand, preventing the subject from concentrating on it (as in split-attention dual-task conditions with an extraneous target). That is why only the execution-focus theory, not the distraction/overload theory, explains why performance is damaged by conditions of focused

—1
—0
—+1

attention (when the athletes try hard to maintain total concentration on their task, as during a competition or tournament—i.e., the kind of situation in which choking most often occurs).

2. Expert performers are better at making skill-focused judgments when under pressure (e.g., Gray 2004). Namely, expert baseball players who bat under skill-focused dual-task conditions experience a more accentuated decrease in their primary task (motor performance) and a concurrent improvement in their secondary task (judgment about their primary task), when they perform in conditions of stress. This suggests that in situations of psychological distress and worry for one's own performance (prompted by social pressure and monetary incentives), attention tends to target the procedural components of one's movements (step-by-step analysis): consequently, the fluid execution of one's habitual action is disrupted while the judgment on those very actions becomes more accurate. The existence of this causal correlation is consistent with the self-monitoring/execution-focus theory and contributes to complete its explanatory paradigm, as it links the functional effects of choking to its psychological and contextual background. On the other hand, the same correlation does not find any justification within the distraction/overload theory, which supports the inverse type of prediction (i.e., higher attention to one's own component movements is supposed to increase the overall motor performances while reducing one's judgment capabilities for extraneous stimuli, not for task activities).
3. Distraction/overload theories struggle to explain why distractors and concurrent secondary tasks are consistently and successfully used as therapeutic interventions on athletes that are particularly prone to choke. Despite the heavier cognitive burden, the experts' performance is improved by those expedients that prevent them from attending to the details of their own movements (Jackson and Beilock 2008). This is achieved by redirecting their attention to external foci (Jackson, Ashford, and Norsworthy 2006, consistent with the attentional focus theories; see Wulf and Su 2007) or reducing the time they have available to perform (Beilock, Bertenthal, et al. 2004; Beilock, Bertenthal, et al. 2008). The fact that experts do not display a speed-accuracy trade-off in spite of the manipulations of the time constraints—indeed, they display the reverse—is particularly significant because it once again frustrates the predictions of the distraction/overload theories: the performance of experts is damaged much more by self-paced action instructions (“putt only when you think you are ready”) than by fast action instructions (“putt as quickly as possible”). Distraction/overload theories cannot account for this result, as they support the opposite prediction: they assume that the performance level decreases when the

-1—
0—
+1—

cognitive load increases, and that the cognitive resources are exhausted faster when the same task has to be processed in a shorter time. Self-monitoring/execution-focus theories, in turn, are well positioned to explain this result because they assume that the performance level depends on automatic routines, and automatic routines are damaged by attention instructions (which explicitly solicit execution-focus), not by speed instructions (which, in turn, facilitate automatic action routines; see Beilock, Bertenthal, et al. 2004).

Taken together, these results suggest that the disruption of the fluid course of familiar automatized routines is due to self-monitoring and/or execution focus, not to the exhaustion of working memory (as confirmed by the fact that even secondary tasks that heavily rely on working memory do not significantly affect the sensorimotor performances of the experts). Self/execution focus impairs the normal course of skillful actions not because it increases the computational burden, but because it decomposes the habitual action routines into disconnected segments and clumsy step-by-step procedures, preventing fluid and adaptive on-line control. That is why choking is occasionally characterized as “paralysis by analysis” (Jackson and Beilock 2008).

It is useful to remark that the execution focus component of self-monitoring/execution-focus theory has another advantage over both distraction/overload theories and also attentional focus theory taken by itself: execution focus theory seems in a better position to explain why choking is skill-level specific, or—in other words—why it threatens expert sport people while sparing novices (e.g., Beilock, Bertenthal, et al. 2004).³ Split attention (dual tasks that do not involve execution focus) and fast action conditions (speed instructions) do not significantly damage the performances of expert athletes performing well-practiced sensorimotor skills, whereas novices experience a drop in their performance exactly under such conditions. On the contrary, skill-focus conditions (attention to movement execution or accuracy instructions) do not significantly decrease the performance of novices, while they can severely damage the performance of experts. In other words, distraction and heavy cognitive loads negatively affect the novices more than the experts, while self-monitoring and execution focus negatively affect the experts more than the novices. Self-monitoring/execution-focus theory can account for this inverted pattern of skill disruption in novices and experts because it assumes that only experts have automatized their motor routines, which is why only experts can experience a significant decrease in their performance when their motor routines are disrupted by execution focus. Distraction/overload theories cannot rely on an explanation of this kind, as they presuppose that choking depends on the exhaustion of the general cognitive resources available to the system (working

—1
—0
—+1

memory), although there are no reasons to believe that experts have less cognitive resources than novices.

6 Objections and Challenges for the Self-Monitoring/Execution-Focus Approach

Despite its many merits, dissatisfaction with the self-monitoring/execution-focus approach to choking has been expressed on multiple levels. First, although the bulk of quantitative evidence from laboratory research supports the self-monitoring/execution-focus theory (Beilock and Gray 2007; Beilock 2011; Gray 2011), this explanation seems to be highly inconsistent with studies of athletes' experiences using qualitative methods (e.g., Hill et al. 2009; Oudejans et al. 2011).

Second, doubts have been raised (e.g., Toner and Moran 2015; Toner, Montero, and Moran 2016; Montero, Toner, and Moran, this volume) as to whether attention can be clearly dichotomized into internal states (i.e., exclusively focusing on the movements of one's own body or on the steps of how one is using a tool) versus external states (i.e., exclusively focusing on the goal-directed outcome of the movement or whether a goal is being achieved). The partial mismatch discussed earlier between internal focus and execution focus in describing the conditions that trigger choking reflects this difficulty.

Third, the ecological validity of laboratory studies that try to extrapolate results from the pressure induced by a small monetary reward to that from winning a major sporting event has been questioned (e.g., Christensen et al. 2015; Mesagno, Harvey, and Janelle 2011). Skeptics claim that laboratory experimentation inevitably fails to reproduce the pressure-filled environment of real competition in which choking typically occurs (Christensen, Sutton, and McIlwain 2015; Hill et al. 2010a). If this attribution to laboratory experimentation is correct, then the decreases in performance artificially produced under laboratory conditions might have little or nothing to do with the emotionally catastrophic effects engendered by choking in real-life competitions (or at best imitate them in only a very pale fashion).

That is why (fourth) it has been proposed that much of the supporting research has investigated "underperformance" rather than "choking," and that additional definitory efforts will be necessary to specifically characterize the intrinsic psychological features of choking to distinguish this phenomenon from other forms of performance disruption (Mesagno and Hill 2013; see also Carr 2015, for a discussion).

Finally, and perhaps most importantly, despite some efforts to relate individual personality traits to choking (e.g., Masters 1992; Otten 2009; Geukes et al. 2012; Geukes et al. 2017), in many ways we are apparently no closer to answering a central question:

-1—
0—
+1—

Are there really “chokers” and “pressure players,” and if so, why do some athletes fail in high pressure situations while others thrive?

These objections reach into the theoretical core of the self-monitoring/execution-focus approach to choking. The main inference underlying this theory is that performance anxiety typically solicits internal focus and/or execution focus in athletes during competition. Internal focus and/or execution focus, in turn, are likely to disrupt the continuous flow of automatic action by breaking its holistic and highly organic structure into an unnaturally detached, mechanical array of inflexibly juxtaposed, discrete kinematic components. This segmentation slows performance, makes it disjointed, and introduces many transition points from one component to the next at which planning and implementation errors can originate.

But this inference is not universally accepted; some researchers explicitly deny that anxiety is causally linked to self-monitoring, internal focus, or execution focus and argue that none of these is related to a consistent decrease in performance (e.g., Montero 2014, 2016; Montero, Toner, and Moran, this volume). This objection is accompanied by phenomenological accounts and qualitative reports suggesting that explicit control is not only possible but necessary to successfully complete many sport tasks, insofar as body awareness (including attention to one’s own posture and movements) is needed to optimize action efficacy, effectively strategize, and make appropriate decisions during complex interactive sport activities (Sutton 2007; Toner and Moran 2015; Montero 2016; Birch, Fusche Moe, and Breivik, this volume), including not just game performance but training and repair after mistakes or slumps (Carr 2015).

Overall, these objections blame the self-monitoring/execution-focus theories of choking for assuming a simplistic notion of expertise, one that emphasizes the velocity and fluidity of routine skillful actions executed in familiar situations but does not recognize that expert skills are also characterized by superior capabilities of decision making, problem solving, and creative improvisation in problematic or unfamiliar situations (Sutton et al. 2011; Cappuccio 2015; Carr 2015; Rucinska 2014b; discontent with the dominant reductive picture of skill had already been raised by Csikszentmihalyi 1996).

Whether this characterization of the self-monitoring/execution-focus theories is correct or not, it reveals a theoretical polarization in the current scientific debate on the cognitive precursors of choking. While distraction/overload theories emphasize awareness, reflection, and explicit decision as the distinctive components of skillful expertise, the self-monitoring/execution-focus theories emphasize the habitual and adaptive components molded around familiar situations and repeated features of commonly encountered sporting environments. Neither provides a complete account of the many

—1
—0
—+1

facets of expertise and its development. Can these two accounts be reconciled and a more complete and successful theory produced from their synthesis, or do they build on ultimately contradictory theoretical assumptions that are irreconcilable?

If possible, a unification of the two misallocation-of-attention approaches would certainly be desirable, as it would provide a fuller and more articulated understanding of choking. Important attempts to overcome this traditional dichotomy have already been made (e.g., *mesh* theory, Christensen, Sutton, and McIlwain, this volume). The success of a truly unified approach to choking should be measured by its capability to account not only for the experimental results that support the attentional theories in laboratory research, but also for the large number of studies, both quantitative and qualitative, that focus on the experiential circumstances of individual athletes, their motivational backgrounds, and their personal psychological profiles (e.g., Jordet 2010; Mesagno, Harvey, and Janelle 2012; Mesagno and Hill 2013; Geukes et al. 2017).

The key to understand choking seems to be the complex interrelation between self-perception, motivation, emotion, sensorimotor control, and attention; that is why these components cannot be studied in isolation. A theoretical picture capable of systematically integrating all these components is still missing. This chapter does not aim to offer such theory, but provides at least some of the theoretical tools necessary to develop it, showing that multiple threats affect performance at different levels, and that all of them relate to some extent to some form of self-consciousness. Embodied cognition, broadly construed, provides the right kind of foundational narrative to appropriately recognize self-consciousness as a multifaceted and complex causal principle of performance disruption. Three notions offered by embodied cognition theory are particularly important:

1. Performance is affected by motor, perceptual, and judgment processes that are systematically integrated (they reciprocally influence one another and/or share the same neurocognitive resources). This is demonstrated by how even higher forms of decision and categorization are scaffolded by basic sensorimotor processes modulated by skill level.
2. Performance disruption cannot be understood without considering the role of emotions and feelings (primarily anxiety and its biophysical expressions) in governing choice and pursuit of goals, directing attention, and, more generally, shaping the cognitive landscape that determines better or worse performances.
3. Performance is not only affected by impersonal mechanisms, but also largely depends on the unique psychophysiological features of each performer as a situated agent with his or her anatomical configuration, motivations, beliefs, identity and self-presentation, and personality traits (e.g., narcissism and perfectionism).

-1—
0—
+1—

These three notions are explored in the parts that follow, which cover particular aspects of self-consciousness and their undesired effects: part 7 focuses on the intertwinement of perceptual, motoric, and judgment processes; parts 8, 9, and 10 focus on the disruptive role played in performance by the appraisal of emotions like anxiety; and parts 11 to 14 deepen the motivational and psychological dimension of skill, examining how self-presentation and personality traits correlate to performance.

7 How Does Pressure Influence Performers' Perceptions of Their Environment?

In sport, it is common for athletes to describe how their perception of the environment can change from moment to moment and game to game. On the one hand, when an athlete is performing well, baseballs can look like “grapefruits” (Will 1990) and basketball hoops can look like “oceans” (Nobles 1995); “things slow down and you feel like you have more time” (McEnroe and Kaplan 2002, 57), and vision is “tunneled in” on the target (Swann et al. 2016). On the other hand, when an athlete is struggling, playing baseball can involve “swinging at aspirins” (Bradley 2003) and athletes notice “spectators and the presence of television” (Oudejans et al. 2011, 67). Are these comments just hyperbole or can perception really change so dramatically?

Embodied perception theory (EPT; see Proffitt 2006) proposes that perception is determined by the relationship between the physical characteristics of objects in a perceiver’s environment and the perceiver’s ability to act on them. EPT is a direct descendant of Gibson’s (1966, 1979) ecological theories. EPT pursues the idea that our senses evolved into an integrated perceptual system to serve action (what is out there in the world, how can I with my own capabilities act in and on that world, and how do these affordances for action enable my goals or stand in their way). As Proffitt puts it in his 2006 paper, “Perception informs people about the opportunities for action and their associated costs” (110). To this end, explicit awareness of spatial layout varies not only with relevant optical and ocular-motor variables, but also as a function of the costs associated with performing intended actions. Although explicit awareness is mutable in this respect, visually guided actions directed at the immediate environment are not. When the metabolic costs associated with walking an extent increase—perhaps because one is wearing a heavy backpack—hills appear steeper and distances to targets appear greater. When one is standing on a high balcony, the apparent distance to the ground is correlated with one’s fear of falling. Perceiving spatial layout combines the geometry of the world with behavioral goals and the costs associated with achieving these goals.

EPT is highly consistent with the subjective impressions of athletes, described earlier. Furthermore, recent laboratory experiments have essentially re-created the athletes’

—1
—0
—+1

experiences illustrated in the quotes above. For example, researchers have shown that the perceived size of objects in a sporting environment is larger and the perceived speed is slower for more skilled compared with lesser skilled performers (for a review, see Gray 2014).

While the link between performance level (e.g., number of hits in a baseball game) and perception (e.g., size of the ball) is clear in these studies, an issue investigated less extensively is how pressure might alter athletes' perceptions. As we have seen, pressure can influence an athlete's ability to perform his or her sport (e.g., Baumeister 1984; Masters 1992; Beilock and Gray 2007; Mesagno, Harvey, and Janelle 2011; Carr 2015). But less evidence exists for the following questions: Does the presence of pressure or anxiety also change the perception of the size and speed of objects in the environment as would be predicted by EPT? Does this happen for all performers to the same extent, or might we be able to predict individual differences in performance or its control from individual differences in perceptual judgments? More generally, do any perceptual changes relate to and predict attentional control and/or performance outcomes? Below, we consider recent studies that have addressed these issues.

Cañal-Bruland, Pijpers, and Oudejans (2010) initially investigated how anxiety influences perception in a sporting context. In this experiment, participants threw darts at a target and produced a judgment of target size. Two conditions were compared: a low-anxiety condition in which participants threw darts while standing on the ground and a high-anxiety condition in which they threw while suspended 3.6 meters above the ground on a climbing wall. The size judgment task involved selecting one of nine circles that matched the size of the target on the dartboard. Results indicated that in the low-anxiety (on-the-ground, normal throwing posture) condition, a significant negative relationship between the mean radial error in dart throwing and the matched target size occurred. That is, participants that performed more accurately in the dart-throwing task estimated the target size to be larger. In the high-anxiety condition, there was no significant relationship between dart-throwing performance and judged target size. Cañal-Bruland et al. (2011) argued that the lack of effect in the high-anxiety condition was because the performer's attention was drawn away from the target.

One might well imagine that this conclusion is correct. People trying to cling to a climbing wall might be anxious, and either component of that situation—clinging to a climbing wall or feeling anxiety as a result—might influence performance. But whether this finding has anything to do with performance under pressure and choking or doing well is debatable. First, one might wonder whether the anxiety produced at heights while clinging to a wall is comparable to the anxiety produced by pressure to succeed in competition. Second, the anxiety manipulation did not actually significantly affect dart-throwing performance, though it did influence judged target size.

-1—
0—
+1—

The distinction we just drew among various sources of anxiety and whether they work the same way is crucially important to understanding performance under pressure. Such a possibility can be seen in an incidental finding reported by Beilock, Bertenthal, et al. (2004), who found that while a self-report measure of “I felt pressure to perform” correlated with performance in their pressure condition, a widely used measure of situational anxiety did not. But this study was not intended to determine whether “pressure” and “anxiety” are the same or different, and it had very low power to discriminate. Several years later, Mesagno, Harvey, and Janelle (2011) found that different ways of imposing “pressure” on high-level field hockey players affected performance outcomes quite differently, and this finding, we believe, should figure heavily in how to think about pressure. Whereas the presence of an audience of peers and the prospect of being evaluated by a coach led to reduced performance relative to unpresurized practice, the prospect of earning money improved performance. Corroborative data from a widely used measure of anxiety showed that while reports of physiological symptoms (such as sweaty palms, racing hearts, shaky hands) *did not* predict performance, reports of worried thoughts about success or failure and their personal and social consequences—so called cognitive anxiety—*did* predict performance, and cognitive anxiety was higher when there was an audience or the prospect of evaluation. Finally, in the results of Mesagno, Harvey, and Janelle (2011) the negative impacts of social/evaluative pressure and the positive impact of monetary reward were approximately additive—as if they might have influenced different sets of processes to create effects that were somewhat independent (Carr 2015).

More recently, in light of these issues, Gray and Cañal-Bruland (2015) investigated the relationship among perceived target size, performance, and pressure in a golf-putting task. In this experiment, twenty-five experienced golfers completed three tasks on every trial: (1) putt a golf ball to a circular target on an indoor putting green from a distance of 2.5 m; (2) estimate the size of the target using drawing software on a computer screen; and (3) perform one of two secondary tasks (a “hole” task in which participants indicated whether a sound originated to the left or right of the hole during their putting stroke, and a “club” task in which participants indicated whether the sound occurred closer to the beginning or end of their backswing). The primary dependent variables were the final distance of the ball from the target, the percent correct for the two secondary tasks, and the perceived hole size. Performance was compared for three phases: (1) a low-pressure pretest; (2) a high-pressure phase, which involved both competitive (monetary reward) and social (results posted in public area) pressure applied together; and (3) a low-pressure posttest. Gray and Cañal-Bruland (2015) initially noticed large individual differences for several of the dependent measures including heart rate, putting kinematics, perceived hole size, percent correct in the two secondary tasks,

—1
—0
—+1

and final distance from the target. Therefore, although there seemed to be links between the different dependent variables, these individual differences needed to be taken into account before traditional analyses could be done.

As we have been discussing, a challenge for researchers who investigate choking is how to take into account between-participant differences in responses to performance pressure. Should we lump all data together and analyze whether or not there was a “group choke” effect (e.g., Beilock and Carr 2001), or should we dichotomize the data and label all participants that showed significant performance decline as “chokers” and those who did not as “clutch” (e.g., Gray, Allsop, and Williams 2013), or should we treat the effect of pressure on performance as a continuum? Or perhaps there is yet another alternative: that laboratory impositions of pressure do not rise to the level that creates “choking” and hence these studies are examining a tendency to perform worse when the stakes are higher, but true “choking” is a different beast with different or additional mechanisms (see Mesagno and Hill 2013). Gray and Cañal-Bruland (2015), in their putting study, chose an option that falls between an attempt to find a bimodal distribution or a continuum in laboratory studies, which we describe next.

To make sense of their data, Gray and Cañal-Bruland (2015) performed a *k*-means cluster analysis using the pressure-manipulation check (heart rate) and two club-head kinematic variables (time to peak speed, TTPS, and velocity at impact, VI) as classification variables. The main goal of the study was to investigate the relationship among putting performance, perceived hole size, and attentional focus (as assessed by the secondary tasks); thus, these variables were not used in the classification. Similar to previous results (Beilock and Gray 2012), there were two distinct clusters: a group of eleven golfers, labelled the “choke” group, for which there were significant changes in HR, TTPS, and VI from the pretest to pressure phase, and a group of fourteen golfers, labelled the “clutch” group, for which there were no significant changes. Treating the “choke” and “clutch” groups as separate groups in the subsequent analyses, Gray and Cañal-Bruland found significant group differences for all three tasks the participants performed. Specifically, compared with the “clutch” group, the “choke” group putted worse (by 3 cm on average), estimated the target size to be smaller (by 3.8 cm on average), had lower accuracy for the “hole” secondary task (by 21 percent on average), and had higher accuracy for the “club” secondary task (by 23 percent on average).⁴

Why did the perceived size of the target decrease for the “choke” group under pressure, and what does this tell us about pressure-induced failures in performance? Based on the data from the secondary tasks, Gray and Cañal-Bruland (2015) proposed that this change in perception was essentially an artifact of changes in the performer’s attention control. Consistent with self-monitoring/execution-focus theory, strong evidence exists

of an inward shift of attention (i.e., inward toward skill execution per se, and away from the external environment and the targets of the execution) in the “choke” group because they performed worse at making the “hole” judgment and better at making the “club” judgment. When a person directs their attention to an object they intend to act on (e.g., a target they are aiming at or a ball they are about to strike), the task-relevant object becomes accentuated so that it stands out from other task-irrelevant objects, often making it appear larger (Bruner 1957). Conversely, shifting attention away from a target toward one’s own body should make the target appear smaller (e.g., Cañal-Bruland et al. 2011). From this interpretation, it is unclear whether the change in perceived target size itself has any direct effect on performance under pressure or whether the changes in putting accuracy can all be explained by the effects of attentional control (e.g., disruption of automaticity with an internal self/execution focus of attention).

An attempt to simulate the effect of pressure through a contrast between speed instruction versus accuracy instruction in a putting task manipulated target size, but while speed-accuracy instructions affected performance considerably, the variation in target size had no significant effect (Beilock, Bertenthal, et al. 2008). However, a possible direct effect of perceptual changes on choking can be seen in a study of baseball batting (Gray 2013). In this study, experienced baseball batters performed two directional hitting tasks (hitting the ball to either the left or right side of the field—called “pull” or “opposite-field” hitting, respectively) in a batting simulator and made judgments about ball size between swings. The primary finding (as shown in figure 5.1, solid symbols) was an interaction between perceived ball size and the batting task. That is, when the pitch location was ideal for the batter’s task (e.g., inside pitches for pull hitting and outside pitches for opposite-field hitting), perceived size was larger.

As first proposed by Cañal-Bruland and van der Kamp (2009), this effect can be explained by the attentional accentuation hypothesis, which claims that when a person intends to act on an object and directs their attention to it, the task-relevant object becomes accentuated so that it stands out from other task-irrelevant objects. The open symbols in figure 5.1 show unpublished data from a condition in which batters performed the same tasks under a combination of evaluative (being videotaped and evaluated by a coach) and competitive pressure (being entered into competition with a cash prize for performance). Consistent with Gray and Cañal-Bruland (2015), pressure seemed to eliminate the difference in perceived size for task-optimal and nonoptimal pitch locations. The data shown in figure 5.2 provide insight into how these perceptual effects might influence performance. Batters appeared to be using an effective hitting strategy of initiating more swing for task-optimal pitch locations in the low-pressure situation, while these advantageous behaviors essentially disappeared under pressure.

—1
—0
—+1

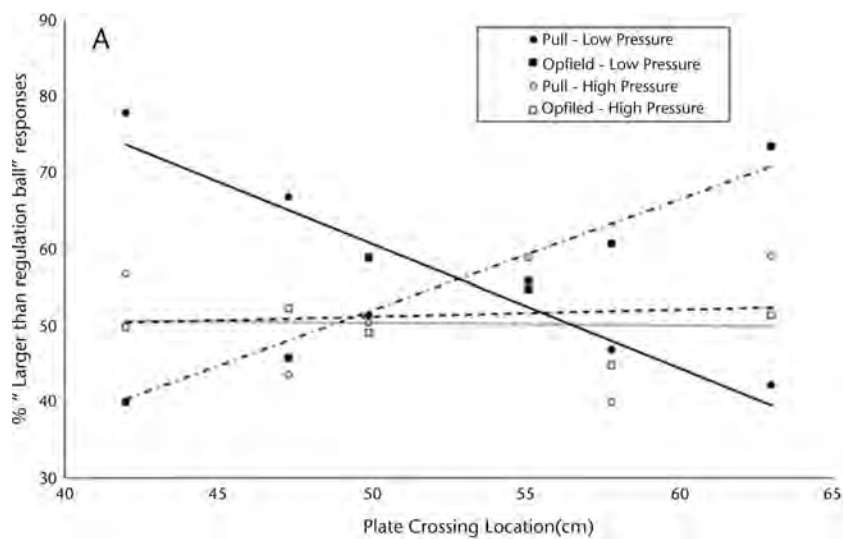


Figure 5.1

Mean percentage of “ball larger than a regulation baseball” response as a function of pitch crossing location. Data are shown for different combinations of batting task (pull hitting vs. opposite field hitting) and pressure (low vs. high).

Source: Gray (2013).

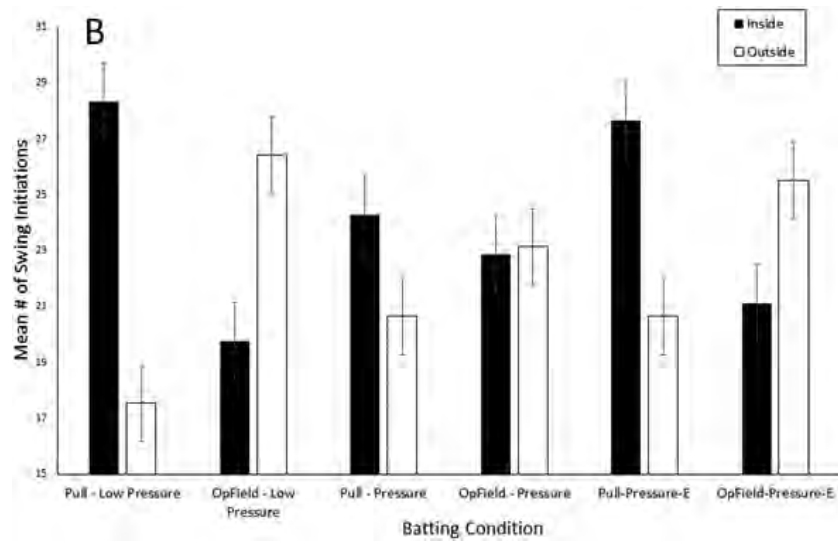


Figure 5.2

Mean number of swing initiations for different combinations of hitting task (pull hitting vs. opposite field hitting) and pressure (low vs. high). Solid bars show data for pitches on the inside (i.e., close to the hitter’s body) while open bars show outside pitches.

Source: Gray (2013).

Again, one might argue that these effects could be explained entirely at the level of attention control and are unrelated to the perceived ball size. That is, under low pressure the batter is attending to the optimal locations and is more likely to swing, while under pressure an inward shift in attention (focusing on the movement of the putter) eliminated this effect. However, the other two conditions shown in figure 5.2 (designated by an E at the end) show results from a condition in which the simulated diameter of the ball in the simulation varied as a function of pitch location. Specifically, when the batter's task was pull hitting, pitches with trajectories that crossed the inside of the plate had a simulated ball diameter of 8.4 cm, while for outside pitches it was the regulation size of 7.4 cm. The opposite was true when the batter's task was opposite-field hitting. These values were chosen based on the changes in perceived size observed for different pitch locations in low-pressure conditions (see Gray 2013 for details). Another way to think of this manipulation of varying the simulated ball size as a function of pitch location is that it is an attempt to cancel out the changes in perceived size resulting from pressure (shown in figure 5.1) by artificially manipulating the ball size. As can be seen in figure 5.2, this manipulation of simulated ball size effectively restored the batters' tendency to swing at pitches in locations that are task optimal. This in turn suggests that at least part of the change in behavior observed under pressure was due to the change in the perceived size of the ball. These results appear to be consistent with the proposed action selection role for embodied perception (Proffitt 2006). That is, perceiving objects as larger when action capability is higher makes one more likely to act.

In sum, recent research has shown that pressure can lead to reductions in the perceived size of target objects in a sporting environment. Whether these perceptual changes are directly related to how well a performer handles the pressure situation (e.g., by changing the likelihood they will act or act accurately) or are simply an artifact of changes in attentional control is an issue that needs additional future research. It will also be interesting to determine if pressure influences the perception of other aspects of the sporting environment, such as the perceived speed of objects and the perceived passage of time. Finally, because there again seems to be substantial individual variation in the perceptual effects of pressure, it is possible that changes in perception could be used as another objective, behavioral index of pressure effects. And, given extant arguments that the mechanisms of choking differ across task domains and levels of expertise (e.g., Beilock, Kulp, et al. 2004; Beilock and DeCaro 2007; DeCaro et al. 2011; Carr 2015; in general, the groundbreaking work done by Proffitt and colleagues deserves a particular mention in this context), research on perceptual changes as accompaniments and predictors (or even possibly causes) of performance changes under pressure needs to be pursued beyond the bounds of sporting environments.

—1
—0
—+1

8 How Do Emotions and Feelings Contribute to Performance Disruption?

As we have seen, there is a growing commitment to the idea that the mind must be understood in the context of its relationship to a physical body that interacts with the world (Beilock 2008a). According to this principle, in order to understand fully the phenomenon of choking in sport, it is essential to examine the athlete's psychological, emotional, and physiological response to the pressurized environment as a holistic and interacting unit. Consequently, a more detailed understanding of the causes, moderators, and consequences of choking may be gained that can inform evidence-based intervention designed to alleviate choking. Through the lens of embodied cognition, we next focus on the role of emotions in the experience of choking in sport. More specifically, we review the interplay among the athletes' emotions, cognitions, psychophysiological response, and the environment during acute sporting failure.

Emotions are subjective feelings that are a conscious response to the athlete's cognitive appraisal of an event (Fredrickson 2001, 2003). That is, by appraising the significance of a particular situation in terms of personal harm and benefits (Lazarus 2000), the athlete will experience an emotional response that triggers a physiological change (i.e., arousal, muscular tension, heart rate, and blood pressure) and action (i.e., approach and /or avoidance behaviors). Thereafter, such emotional and physical responses will be subjected to further appraisals (see Neil et al. 2011). Thus a continuous cycle of interaction and integration takes place among cognitions and emotions.

Emotions are classified according to their hedonic tone and functionality—with hedonic tone referring to whether the emotions are pleasant (e.g., happy) or unpleasant (e.g., anxiety), and functionality denoting whether the emotion has a positive (i.e., functional) or negative (i.e., dysfunctional) impact on the performance (Hanin 2000, 2007). Both pleasant and unpleasant emotions can have a functional or dysfunctional effect on performance, depending on the athlete's cognitive appraisal of those emotions (e.g., Totterdell 2000; Lane et al. 2011). For example, while the pleasant emotion of happiness often leads to optimal performance, it can also encourage performance failure if the athlete perceives the emotion is associated with satiated action tendency (Erez and Isen 2002). In this instance, athletes may lower effort when they are happy because they no longer feel the need to work hard. Conversely, the unpleasant emotion of anger can have a functional effect on performance, if the athlete believes the emotion is energizing, whereas anger can cause performance failure if it increases the athlete's perceptions of feeling out of control (Woodman et al. 2009). Thus, the appraisal of a pressurized situation will encourage an emotional response that affects attention

(e.g., Eysenck et al. 2007), motor control (Coombes et al. 2009), perception (e.g., Valleurand and Blanchard 2000), decision making (Tenenbaum and Bar-Eli 1993), motivation (e.g., Kerr 2014), and, ultimately, performance (Gucciardi et al. 2010; Hill et al. 2010a; Neil et al. 2011).

With specific regard to choking in sport, anxiety has been identified as a central emotion (Hill and Shaw 2013; Mesagno and Hill 2013; Carr 2015), which is likely the principal emotion experienced by athletes in response to perceived pressure. Indeed, extensive evidence from experimental studies supports the role of high anxiety levels within the choking process (see Beilock and Gray 2007; Hill et al. 2010a for a review). Of the limited qualitative research available, anxiety has also been reported by athletes as being the dominant emotion they associate with their choking experience (Gucciardi et al. 2010; Hill et al. 2010b; Hill and Shaw 2013; Hill and Hemmings 2015). Critically, this work provides a further indication that athletes' appraisal of anxiety may be what determines performance outcomes, rather than the mere presence of the emotion. Hence, athletes who experience choking normally perceive their anxiety as debilitating, while those who appraise it as facilitative tend to excel under pressure (Gucciardi et al. 2010; Hill et al. 2010b; Hill and Shaw 2013). Furthermore, appraising their physiological responses to anxiety (i.e., muscular tension, increased heart rate, a tight chest, and palpitations) as debilitating also appears to encourage the likelihood of choking (Gucciardi et al. 2010).

While anxiety is an important determinant of choking in sport, it is concerning that researchers have failed to consider fully the potential influence of other emotions. This is despite evidence indicating that a broad range of emotions are likely to determine the outcomes of athletic performance (e.g., Totterdell 2000; McCarthy 2011; Uphill, Groom, and Jones 2014). Through qualitative research, many unpleasant emotions have been noted by athletes before and/or during their choking experience, which include (but are not limited to) embarrassment, shame, humiliation, fear, hate, inferiority, negativity, despondency, annoyance, anger, distress, or feeling overwhelmed, uncontrolled, or hopeless (see Gucciardi et al. 2010; Hill et al. 2010a; Hill et al. 2011; Mesagno, Harvey, and Janelle 2012). Since such emotions were not directly explored, it remains unclear which situational factors triggered each of the emotional reactions, how each emotion specifically affected the performer/performance, and whether any individual differences existed in terms of the emotional response and its effect. The only exception to date is Mesagno, Harvey, and Janelle (2012), who found that fear (of negative evaluation) led to choking in their study of basketball players (consistent with the findings reviewed earlier from Mesagno, Harvey, and Janelle 2011). Accordingly, while it cannot be doubted

—1
—0
—+1

that anxiety can encourage choking in sport, researchers should also examine whether other emotions elicited as a result of perceived pressure can moderate choking.

9 Emotion, Motivation, and Choking in Sport

When placed in a meaningful situation, an athlete will experience an emotional response that enables them to prioritize their goals, generate a psychophysiological state of readiness, and become energized to act according to that situation (see Baumeister et al. 2007). While evolutionary psychology might presume emotions should have a benign and/or adaptive effect on human beings, evidence indicates that many self-defeating behaviors are fostered by emotion (Baumeister et al. 2007; Baumeister and Lobbestael 2011), including choking under pressure (Baumeister and Scher 1988). Emotions provide feedback regarding how well the athlete is moving toward a pertinent goal, with positive emotions tending to be experienced when appropriate progress is being made and negative emotions signaling progress is slower than expected (Carver and Scheier 1981; Carver, 2003). Thereafter, individuals are more likely to adjust their goals in order to avoid negative emotions and increase positive emotions. Consequently, if the athlete is moving effectively toward their goal and experiencing positive emotions, he or she is likely to become increasingly motivated to pursue that goal until it is reached. With choking in sport, however, athletes perceive they are unlikely to reach a pertinent goal, and so experience an intense negative emotional response (Hill et al. 2009). Initially, if the goal is important (e.g., winning a tournament), such negative emotions may initiate increased effort/motivation. Paradoxically, such efforts can lead to the misregulation of attention, whereby the athlete attempts to consciously control/monitor the explicit components of the skill (i.e., execution focus) and experiences choking as a result (see Baumeister and Heatherton 1996). Whereas, if the athlete continues to perceive they are not capable of reaching their goal, then the associated negative emotions will eventually cause them to cognitively disengage from the goal (Wrosch et al. 2003) and adopt avoidance motivation, as they remove themselves from the situation causing the distress. Therefore, such avoidance-goal intentions and behaviors are also acknowledged determinants of choking in sport (Jordet and Hartman 2008; Jordet 2009; Carr 2015; Hill and Hemmings 2015).

Emotions can also direct the individual's attention to relevant aspects of the situation, allowing the opportunity to identify the appropriate action and/or reflect on lessons that can be learned for the next time they find themselves in the same situation (Baumeister et al. 2007). However, athletes with high trait anxiety, who are therefore

-1—
0—
+1—

prone to choking in sport (Baumeister and Showers 1986; Wang et al. 2004), are known to experience a narrowing of their attention and a sensitivity to threat stimuli in their environment when exposed to perceived pressure (Eysenck et al. 2007). As a result, when under pressure and experiencing high anxiety (and other unpleasant emotions), they are unable to look beyond the threat stimuli causing their unpleasant emotions. Consequently, during pressurized performance, those athletes with high trait anxiety will focus on removing themselves from the situation causing the negative emotions rather than assimilate relevant information and explore all appropriate options. Such avoidance motivation/behavior can lead not only to poor sporting performance (i.e., choking), but can also prevent the athlete from learning lessons from that situation to inform their self-management during future pressurized situations. This may go some way to explaining why some athletes continue to choke each time they are exposed to pressure (Hill et al. 2009, 2010b, 2011).

As a final comment, when individuals know they will experience negative emotions in the future when facing certain events (i.e., performing under pressure), they often begin to engage in proactive coping (i.e., efforts undertaken in advance of a stressful event to prevent or modify it) to minimize those emotions (Aspinwall and Taylor 1997). Such proactive coping efforts can include preventing the aversive event from occurring and/or minimizing its eventual impact. This hopefully can encourage the athlete to learn and develop a repertoire of coping strategies that can be employed to alter their appraisal of future pressurized situation and help them manage more effectively their emotional response (Hill et al. 2011; Hill and Hemmings 2015). However, in certain cases of choking, proactive coping has consisted of self-handicapping behavior, whereby the athlete attempts to lower his or her expectations of success by subconsciously undermining preparation for the pressurized event (see Hill et al. 2011). Indeed, such a coping approach may even cause the athlete to avoid high-value situations altogether (Hill et al. 2009).

Collectively, the above demonstrates that, for certain athletes, emotions can be counterproductive and lead to choking under pressure. Accordingly, the experience of choking is one in which the pressurized situation causes the athlete to suffer emotional distress, which in turn elicits goal intentions and motivational actions that prevent effective management of the self and situation. More work is needed to identify just which athletes might be more or less susceptible to the negative effects of emotions and consequent changes in motivation, strategy, and goal direction, and on how to intervene in the patterns to help athletes avoid or overcome those effects. Next we address one approach to such identification and intervention.

—1
—0
—+1

10 Choking under Pressure and Challenge/Threat Appraisals

A useful framework for understanding the interactions among an athlete's motivation, emotions, and performance outcomes that can lead to choking is the theory of challenge and threat states in athletes (TCTSA; see Jones et al. 2009). The TCTSA draws on the biopsychosocial model (Blascovich and Tomaka 1996), the model of adaptive approaches to competition (Skinner and Brewer 2004), and the control model of anxiety (Jones 1995), by proposing that when engaged in goal-driven performance events, athletes will appraise the situation as either a challenge or a threat (Blascovich 2008). Cognitive appraising of the event as a challenge occurs if they perceive they have sufficient resources to meet the situational demands, whereas they will appraise the event as a threat if they perceive insufficient resources to meet the situational demands (Blascovich and Mendes 2000).

Critically, each appraisal will lead to a challenge or threat motivational state (Jones et al. 2009), which have different associated physiological and emotional responses. Physiologically, a challenge state is characterized by an increased activation of the sympathetic-adreno-medullary (SAM) system and the release of epinephrine, which causes higher cardiac output and vasodilatation and decreased systemic vascular resistance. Combined, this leads to increased blood flow and the efficient mobilization of energy, which can facilitate action and coping responses (Blascovich et al. 2000). A threat state is also associated with SAM activity, yet there is an increased activation of the pituitary-adrenocortical system, which initiates the release of adrenocorticotrophic hormone. Therefore, while cardiac activity increases, it is not accompanied by vasodilatation or decreased systemic vascular resistance. Accordingly, blood flow remains static and the athlete does not benefit from the efficient mobilization of energy as seen in a challenge state (Blascovich and Tomaka 1996).

Positive emotions are normally experienced by athletes in a challenge state, and negative emotions during a threat state. Nevertheless, it is accepted that the athlete's interpretation of their emotions will moderate this position. As an example, negative emotions (e.g., anxiety) can be interpreted as facilitative and therefore will exist in a challenge state (Mendes et al. 2008), while positive emotions (e.g., contentment), interpreted as debilitating for performance, can be experienced within a threat state.

The TCTSA also proposes that an athlete will appraise an event as a challenge or threat depending on their self-efficacy, perception of control, and goal orientation (Jones et al. 2009). Thus, athletes with high self-efficacy will perceive they have the capability to manage successfully the demands of the situation and will appraise the situation as a challenge. Conversely, those low in self-efficacy will appraise the event as

-1—
0—
+1—

a threat, for they are unsure whether they have the capability to meet those demands (Williams, Cumming, and Balanos 2010). Perceived control has a similar impact on athletes' appraisals, whereby if they believe they have sufficient control over themselves and the environment to achieve their performance goals, they will experience and benefit from a challenge appraisal. In contrast, those athletes lacking perceived control will appraise the situation as a threat. Finally, in terms of goal orientation, those who adopt approach goals (in particular, mastery approach) will view the upcoming event as a challenge, and those who hold avoidance goals will tend to have a threat appraisal. This is likely to be the result of self-referenced achievement goals increasing the athletes' self-efficacy and perceived control (Jones et al. 2009; Meijen et al. 2013; Turner et al. 2012).

There is an extensive evidence base to support the relationship between challenge and threat states, their associated cardiovascular indices (i.e., physiological response), and performance outcomes, which has been corroborated across tasks including perceptual-motor (Vine, Freeman, et al. 2013), cognitive (Turner et al. 2012), and sensorimotor skills (e.g., Moore et al. 2012). Hence, a challenge appraisal and motivational state has been linked to improved sport performance under pressure, with athletes in a threat state found to be more vulnerable to performance decrement and choking than those in a challenge state (Moore et al. 2012; Moore, Wilson, et al. 2013; Turner et al. 2012). By manipulating an athlete's appraisal and motivation state toward challenge (e.g., through manipulating gaze/reappraisal training), athletes are likely to experience enhanced performance under pressure and may, in turn, become increasingly resistant to choking (Moore, Vine, et al. 2013; Moore et al. 2015).

The precise mechanism through which challenge and threat appraisals affect performance remains somewhat unclear, although the most likely explanation is emotional and attentional control. Athletes in a challenge state experience facilitative emotions (Williams et al. 2010; Moore et al. 2012; Moore, Wilson, et al. 2013), which have an established relationship with optimal sport performance (Neil et al. 2011; Thomas, Maynard, and Hanton 2007). However, while challenge and threat states lead to differential emotional responses, those differences do not fully and consistently explain why athletes in a challenge state perform better than those in a threat state (see Moore et al. 2012). With regard to attentional control, individuals within a threat state are less able to remain focused on task-relevant information because they become distracted by threat stimuli in their environment (Vine, Freeman, et al. 2013). Furthermore, a longer quiet-eye duration (i.e., the final fixation located on an object for a minimum of 100 milliseconds, Vickers 2007, 280) has been noted with athletes in a challenge state (Vine, Moore, et al. 2013; Moore et al. 2012; Moore, Vine, et al. 2013), which is

—1
—0
—+1

considered by some to benefit pressurized performance as it extends the critical period during which the motor response is selected and programmed. Finally, athletes within a challenge state experience less “reinvestment” (Moore, Wilson, et al. 2013), which is the attempt to consciously control a well-learned skill under pressure (Masters, Polman, and Hammond, 1993), similar to Beilock’s execution focus (the concept of “dispositional reinvestment” is discussed in part 13); that is why they are thought to be less likely to focus on the explicit components of their skill and are less vulnerable to choking through self-focus (Masters and Maxwell 2008). However, studies have failed to find a consistent link between challenge/threat appraisal, attention control, and performance across groups (Moore, Wilson, et al. 2013), meaning further research is required before conclusions can be drawn.

Other mechanisms have been proposed to explain how challenge and threat appraisals affect performance, though they have received less research attention to date. They include the use of problem- and emotion-coping strategies during a challenge state (Allen, Frings, and Hunter 2012), which may be possible because of the associated cardiovascular changes enhancing the athlete’s available energy resources, which can then sustain such effortful approach-coping. Additionally, an increased efficiency of kinematics has also been suggested as a mechanism through which appraisals can impact performance. As an example, Moore et al. (2012) observed efficient forearm muscle activity during a putting task for those athletes who were in a challenge state, when compared with their threat counterparts.

11 How Do Personality Characteristics Influence Proneness to Pressure?

Choking and Self-Presentation Theory

The two dominant choking models as recounted earlier in the chapter differ substantially in the roles they attribute to attention and the nature of the shifts of attention that they envision to occur during performance under pressure. As we have described, there is evidence for both sets of processes and for interactions and changing relations between their effects as level of skill increases. Even more complex, however, is how personality characteristics may interact with these sets of cognitive processes and the processes and behaviors of self-presentation. These interactions can lead to increased pressure and decreased performance, and so should be integrated into models of performance under pressure.

Self-presentation is the process by which people attempt to monitor and control what they do in public and hence how they are perceived and evaluated by others (Schlenker 1980). Mesagno and colleagues (Mesagno 2009; Mesagno, Harvey, and Janelle

-1—
0—
+1—

2011) proposed that concerns about self-presentation may be the origin of the increased state anxiety for choking-susceptible athletes (those more likely to experience choking). Mesagno's model was conceptualized from qualitative research and grounded in Leary's (1992) suggestion that competitive anxiety is a sport-specific class of social anxiety. While competitive anxiety might not be limited to sports—it might arise in many arenas of life where people compete with one another—sports provide a clear avenue for study. In an initial investigation of the self-presentation model of choking in sporting environments described earlier, Mesagno, Harvey, and Janelle (2011) found that self-presentation pressure manipulations (such as having an audience of knowledgeable and talented teammates, or knowing that performance would be evaluated by their coaches) increased anxiety and choking tendencies more than did a motivational pressure-comparison condition involving monetary reward, with cognitive state anxiety mediating the relationship between self-presentation and performance. These results provided initial support for a linkage among self-presentation, increased state anxiety (at least as measured by self-reports of the content of thoughts—rather than physiology—as discussed earlier), and choking.

Mesagno and colleagues (Mesagno 2009; Mesagno, Harvey, and Janelle 2011) argue that the interaction of high athletic identity (described below), choking susceptibility through certain personality characteristics, and situational factors that affect perceived pressure are likely to increase the likelihood of elevated perceived pressure and choking. Baumeister (1984) defines pressure as “any factor or combination of factors that increases the importance of performing well on a particular occasion” (610). Personality characteristics and situational factors can interact to influence the perception of pressure in the context. One epitome of being an elite athlete is showing publicly that you have the ability to perform successfully under pressure. One key facet of the self-presentation model is athletic identity. Athletic identity is the degree to which individuals define themselves as athletes, and, as argued by Brewer, Van Raalte, and Linder (1993), such definition is central to many athletes' self-concepts.

Brewer and colleagues divided athletic identity into three components: social identity, exclusivity, and negative affectivity. Social identity refers to the degree that individuals view others' perceptions of them as athletes. Exclusivity is how strongly people rely on their athletic identities and whether they identify themselves weakly (or strongly) with other important life roles. Negative affectivity refers to the degree individuals negatively respond as a result of being unable to participate in sport (Martin, Eklund, and Mushett 1997). Researchers have found that a stronger and more exclusive athletic identity may lead to negative psychological and social concerns including decreased self-esteem (Crook and Robertson 1991), emotional difficulties (Lavalley, Gordon, and

—1
—0
—+1

Grove 1997), and mood disturbances (Brewer 1993). Based on these findings, Mesagno and colleagues argue that a stronger athletic identity could be a factor that increases perceived pressure, especially if the athlete views himself or herself exclusively as an athlete; yet, empirical evidence on this link is still to be determined. Nevertheless, personality characteristics that lead to choking susceptibility may influence the perception of pressure. In some ways this notion maps onto the similarities Beilock et al. (2006) identified between stereotype threat and choking, given that stereotype threat is widely regarded as dependent on how important are the links between a task to be performed and one's sense of self, and the large impact of the presence of an audience on the magnitude of stereotype threat that might call one's sense of self to the fore or participate in evaluation of the task's performance.

12 What Personality Characteristics Might Make Someone Choking-Susceptible?

Just as a full and detailed consideration of the relations among emotions and choking was beyond the limits of this chapter, we can provide only a brief overview of how personality characteristics might relate to choking (but see Jackson, Beilock, and Kinrade 2013; Mesagno, Geukes, and Larkin 2015 for recent reviews). To date, researchers have found support that self-consciousness, trait anxiety, fear of negative evaluation, fear of failure, perfectionism, dispositional reinvestment, and narcissism are likely characteristics of choking-susceptible athletes.

Self-consciousness is defined as the tendency to direct attention either inward, focusing on covert aspects of self related to one's inner thoughts and feelings, or outward, focusing on aspects of self that are public to others (Fenigstein, Scheier, and Buss 1975; Scheier and Carver 1985). Initial investigations (e.g., Baumeister 1984) indicated that individuals low in self-consciousness perform poorly under pressure. However, other researchers (e.g., Masters, Polman, and Hammond 1993; Dandy, Brewer, and Tottman 2001; Wang et al. 2004) have provided evidence that highly self-conscious athletes are more likely to experience choking. Geukes et al. (2013) suggested that the mixed findings may be explained by the relevance of private versus public self-consciousness in a situational context. That is, Geukes and colleagues found that private self-consciousness is negatively associated with performance under pressure in laboratory-based high-pressure situations (i.e., one to seven spectators), while public self-consciousness is positively related to performance under public pressure (i.e., in front of a large audience of 1,500 to 2,000 spectators). Nevertheless, more evidence indicates that highly self-conscious individuals generally are more likely to experience choking.

-1—
0—
+1—

Trait anxiety is the disposition to perceive a wide range of situations as threatening and to respond with state anxiety (Spielberger 1966) which appears, as discussed earlier, to play a major role in choking. Test/examination anxiety researchers have confirmed that individuals high in trait anxiety react to pressure situations with greater levels of state anxiety than individuals low in trait anxiety (Spielberger, Anton, and Bedell 1976). Furthermore, Calvo, Alamo, and Ramos (1990; Kurosawa and Harackiewicz 1995) found that high-trait-anxious individuals perform poorer under pressure than low-trait-anxious individuals. Sport anxiety researchers have also confirmed that trait anxiety is a strong predictor of state anxiety (e.g., Marchant, Morris, and Andersen 1998; Williams and Krane 1992). Baumeister and Showers (1986) posited that individuals high in trait anxiety would be more susceptible to choking than individuals low in trait anxiety, and Wang et al. (2004) found that trait anxiety (specifically, somatic trait anxiety) was negatively correlated with basketball free-throw performance under pressure. Thus, although a dearth of research has been focused on the effects of trait anxiety on performance, trait anxiety has links to choking effects, but further research is necessary to understand the underlying mechanisms regarding trait anxiety as a predictor of choking.

For many athletes, elite-level competition is a stressful and anxiety-provoking environment. When athletes experience pressure in sport, their ability to deal with being in the spotlight of competitive sport and achieve optimal performance may be substantially influenced by their coping skills. Two major classes of coping skills are discussed in the choking literature, and each may have a different effect on how performers deal with anxiety and pressure. Approach coping involves focusing on a particular concern by using direct cognitive effort (Crocker and Graham 1995), whereas avoidance coping typically entails directing cognitions away from the threat-related stimulus (Anshel and Weinberg 1999). Wang, Marchant, and Morris (2004) found that basketball players using approach coping were less accurate on a basketball free-throw shooting task under high pressure than those using avoidance coping. Based on a multiple regression analysis, Wang et al. (2004) found that approach coping accounted for 7 percent of the explained performance variance under pressure. From the resultant outcome and although largely speculative and without direct evidence, Wang et al. suggested that approach “copers” ruminate about anxiety-related cognitions during performance, which may increase perceived threat as they search for explanations for the increase in anxiety and actively seek to reduce the anxiety symptoms and experience. This irrelevant attentional focus minimizes task-relevant cognitions, which leads to poorer performance outcomes—at least in situations in which explicit task-relevant cognitions are needed (an issue already much-discussed in earlier parts of the chapter).

—1
—0
—+1

13 Fear of Negative Evaluation, Fear of Failure, Dispositional Reinvestment, Perfectionism, Narcissism

Since anxiety, impression management, and self-presentation seem to play important roles in performance under pressure, researchers have suggested that fear of negative evaluation may be a predictor of choking. Carleton et al. (2006) define fear of negative evaluation (FNE) as “apprehension and distress arising from concerns about being judged disparagingly or hostilely by others” (297). As a follow-up to the Mesagno, Harvey, and Janelle (2011) study of the impact of different kinds of pressure on professional field hockey players taking penalty shots, Mesagno, Harvey, and Janelle (2012) found that participants high in fear of negative evaluation increased anxiety and decreased basketball free-throw shooting percentage under high pressure, whereas participants low in fear of negative evaluation maintained similar anxiety levels and increased free-throw shooting percentage from the low- to high-pressure situation. Mesagno and colleagues suggested that athletes who rate high in FNE may be susceptible to choking, and as discussed earlier, this notion figures heavily in the self-presentation-related theoretical approach to choking.

Atkinson (1957) defined fear of failure (FOF) as a “disposition to avoid failure and/or a capacity for experiencing shame or humiliation as a consequence of failure” (13). Choking researchers (e.g., Jordet and Hartman 2008; Gucciardi et al. 2010; Hill et al. 2010b) have found both quantitative (i.e., behavioral indicators in World Cup soccer penalty shoot-outs) and qualitative evidence for the positive link between choking and FOF. In qualitative studies, Gucciardi et al. (2010) and Hill et al. (2010b) separately interviewed choking-susceptible golfers and found FOF as categories from their interviews. Apparently, fear (of negative evaluation or failure) may play a role in predicting choking. Nevertheless, only limited quantitative studies have focused on FNE, and primarily qualitative studies have identified that FOF is a theme, with minimal quantitative, laboratory-based studies having been conducted. To further develop this area, studies need to be conducted to tease out how these types of fear affect anxiety and performance under pressure.

Dispositional reinvestment is an athlete’s overall tendency to attempt to consciously control a well-learned skill under pressure (Masters, Polman, and Hammond 1993). Reinvestment theory “argues that the propensity for consciousness to control movements on-line is a function of individual personality differences, specific contexts and a broad range of contingent events that can be psychological, physiological, environmental or even mechanical” (Masters and Maxwell 2008, 160). Sport psychology researchers (e.g., Masters, Polman, and Hammond 1993; Maxwell, Masters, and Eves 2000;

-1—
0—
+1—

Masters and Maxwell 2004; Poolton, Maxwell, and Masters 2004; Jackson, Ashford, and Norsworthy 2006; Kinrade, Jackson, and Ashford 2010) have repeatedly confirmed that individuals high in dispositional reinvestment experience choking in sensorimotor skills, compared with those low, at least when the performers being tested are experienced rather than novices. To our knowledge, this measure has not been applied to non-sensorimotor working-memory-intensive skills, but the prediction is clear: reinvestment might benefit performance rather than hurt it (see Beilock, Bertenthal, et al. 2004; Carr 2015).

Perfectionism reflects a person's compulsive pursuit to set and attempt to reach excessively high standards and a tendency to engage in harsh, overly critical self-evaluation (Flett and Hewitt 2005). High standards are important and beneficial to sport performance, but when perfection is the only acceptable standard and non-perfection is common, a negative self-concept and fear of failure are increasingly likely to develop (e.g., Williams and Leffingwell 2002). Perfectionists, compared with non-perfectionists, rate task importance as significantly higher across different evaluative threats and tasks (e.g., Frost and Marten 1990) and have high social or trait anxiety (e.g., Alden, Ryder, and Melling 2002; Frost and Henderson 1991). Researchers (e.g., Clark and Wells 1995; Schlenker and Leary 1982) have also incorporated interpersonal and self-presentation components of perfectionism into their models of social anxiety. Recently, qualitative researchers (e.g., Gucciardi et al. 2010; Hill et al. 2010b) have identified perfectionism as a theme (or psychological attribute) associated with athletes who experience choking. Flett and Hewitt (2005) also illustrated the link between perfectionism and self-presentation by explaining that perfectionistic self-presentation (i.e., striving to create a flawless public image) may be common within sport and could be considered as an antecedent to increases in competitive anxiety. Thus, it appears that perfectionism may be linked to many of the other personality characteristics within the choking literature and could be an important predictor, but additional research is needed determine how important.

Narcissism is characterized as an addiction to self-esteem and an obsession with demonstrating personal superiority or greatness over others (Baumeister and Vohs 2001). Only limited research has focused on narcissism and choking but, when it has been studied, researchers (e.g., Geukes et al. 2012, 2013; Wallace and Baumeister 2002) have found that narcissism decreases the likelihood of choking. For example, Wallace and Baumeister (2002, experiment 3) asked participants to perform a dart-throwing task in both a low-pressure (with only the researcher present) and a high-pressure condition (where the experimenter offered a monetary incentive) and found that individuals high in narcissism were more accurate under pressure than those low in narcissism.

—1
—0
—+1

However, the findings of Mesagno, Harvey, and Janelle (2011) that monetary reward may actually increase performance, whereas the presence of an audience and the prospect of evaluation decreases performance, makes this finding difficult to interpret. A bit of light might be cast by the findings of Geukes et al. (2012, 2013), who also found empirical support for a positive relationship between narcissism and performance under pressure on a handball penalty-taking task. However, the narcissism results were found only when pressure was manipulated in a public manner in front of a large crowd of 1,500 to 2,000 spectators. Consistent with earlier discussions of comparisons between real-world and laboratory pressure situations, narcissism did not significantly correlate with high-pressure performance in a laboratory-based pressure manipulation (Geukes et al. 2012, 2013). Thus, further research should tease out the reasons for these contextual explanations of narcissistic tendencies on performance under pressure.

14 Situational Factors in Perception of Pressure and Occurrence of Choking

Situational factors of choking are often changing, and depend on environmental and psychological circumstances that differentially affect performance. Perceptions of these internal and external factors, however, can be easily modified to help individuals manage the situation. Baumeister and Showers (1986) provided an extensive explanation of these factors in their review of literature that relate (but are not limited) to: self- and others' expectations, the importance of performing well, audience and video camera presence, and monetary incentives.

If the situation is threatening and/or important to the athlete, then increases in perceived pressure occur, which usually result in higher state anxiety. The more an athlete is susceptible to increases in perceived pressure through personality characteristics, the more likely the athlete may develop self-presentation concerns. Sport provides a social situation where real and imagined self-presentation concerns are abundant and perceptions of threat increase competitive anxiety. This is especially true considering sport anxiety researchers (e.g., James and Collins 1997; Wilson and Eklund, 1998; Williams, Hudson, and Lawson 1999; Bray, Martin and Widmeyer 2000; Hudson and Williams 2001; Lorimer 2006) have consistently found a positive relationship between state and trait anxiety and self-presentation concerns. For example, researchers (Wilson and Eklund 1998; Hudson and Williams 2001) have found that self-presentation concerns are more strongly related to cognitive than somatic trait anxiety. Furthermore, Bray et al. (2000) also found that evaluative concerns were significantly, albeit modestly, related to precompetitive state anxiety among athletes, with nonspecific and performance-specific evaluative concerns both related to cognitive anxiety. The

-1—
0—
+1—

link between competitive anxiety and self-presentation has also been identified in “choking-susceptible” athletes, with researchers (e.g., Gucciardi et al. 2010; Hill et al. 2010a) explaining that self-presentation concerns and social evaluation were themes that led to increased anxiety and choking experiences.

Leary (1992) argued that, during competition, athletes risk conveying a variety of negative images of themselves to an array of evaluative others including (but not limited to) spectators, teammates, coaches, and opposing team members. Those who are prone to choking susceptibility want others to view them in a desirable way; they attempt to control inferences made about them by others, often behaving or thinking in ways that might convey a positive image about themselves to others, to perhaps minimize social anxiety. In their self-presentation model of social anxiety, Schlenker and Leary (1982) explained how social anxiety increases when people are motivated to make a desired impression on others but doubt they will be successful. Leary (2001) expanded the original self-presentation theory by suggesting that social anxiety increases if there is potential for relational devaluation to occur; when people believe that the impressions made will lead others to devalue, avoid, or reject their relationships with them. For the choking-susceptible athlete, if the public self (the image of oneself that exists in the minds of others; Baumeister 1982) is discredited, a negative self-presentation and relational devaluation may occur. Particularly for the athlete who is high in athletic identity, being portrayed as an unsuccessful athlete under pressure or, more drastically, a “choker,” can clearly lead to self- and relational devaluation. When athletes are concerned about what others think of them, their behaviors, cognitions, and emotions in competitive sport settings may change. Thus, during performance under pressure, choking-susceptible athletes may draw on cognitions within choking models (i.e., distraction and self-focus approaches) to help them (counterproductively) make a positive impression in the public setting so a negative image of them as an elite athlete does not occur. The effects of the audience-presence and evaluation-by-coaches conditions in Mesagno, Harvey, and Janelle’s (2011) experimental study of performance under pressure in field hockey penalty shots are certainly consistent with this idea.

15 Conclusion

Our review of the scientific study of choking has specifically focused on the domain of sensorimotor skills. The research shows that when one is a novice at a new sport skill, concentration on the step-by-step performance of that skill is required in order to do one’s best, while distracting thoughts about the audience, possible failure, or other irrelevant tasks imposed from the outside are heavily disruptive.

—1
—0
—+1

Highly practiced experts appear to be somewhat the opposite, when performing a very well-practiced task in their domain of expertise, as they benefit from letting go of conscious control. Their performance improves if they rely on what has been called “automaticity” from a skilled-performance perspective (e.g., Fitts and Posner 1967; Anderson 1982; Brown and Carr 1989; Beilock and Carr 2001; Beilock et al. 2002), “flow” from a more humanistic or “positive psychology” perspective (Jackson and Csikszentmihalyi 1999; Shernoff et al. 2003), or “enactive creativity” from a philosophical perspective (Rucinska 2014a, 2014b; Rucinska and Aggerholm, this volume).

However, there are limits. The benefits of a state of “flow” or “enactive creativity” that allows “automatic” procedures to govern performance can arise only if the performer possesses enough well-practiced knowledge and skill to support them, can control and use his or her emotions sufficiently and effectively, and can resist the pull to shift attention toward conscious control of particular steps of performance and/or particular body parts that carry out those steps.

What governs this shift of control? Increases in arousal accompany the experience of pressure. Moderate increases help performance by increasing available energy, increasing processing capacity, and focusing attention on task-relevant information, but larger increases in arousal eventually push performers over the top of the Yerkes-Dodson arousal-performance function, leading to decreases in performance as attention narrows, concentration falters, and reward processes shift from aiming for gains to avoiding losses.

At the highest levels of arousal in the Yerkes-Dodson function, panic may begin to set in, and withdrawal from performance altogether can be the outcome. This is the sort of end state that phenomenological reports using qualitative interview methods have obtained from high-level athletes talking about experiences of choking. Athletes say they could not do anything right and their only desire was to get off the field, to escape. This is of course an extreme response, but we must consider the extremes in order to figure out how “choking” arises, what it is like, and what it entails.

The entire human organism is involved in the processes we have described, from the earliest applications of attention for gaining perceptual information to the interpretative and judgmental processes by which that information is evaluated to the working memory processes by which explicit thoughts are entertained and portions of the goal structure of the task are activated for implementation to the implementation processes and the actions themselves, whose forms, trajectories, and outcomes are shaped as much by the physiognomy of the body as by the goals and operations of the mind. All of these rather cognitive- and kinesiological-sounding processes are governed by and integrated with emotional and motivational states and processes, which

-1—
0—
+1—

can vary significantly both with situational conditions and with measurable facets of personality.

All of the above simply repeats facts that have been reviewed in this chapter. But what remains to be determined? In our view, the important questions are what are all the various states and processes occurring in real time in athletes that change when “pressure” is felt, what states and processes change in the laboratory performer and is the lab experience anything like what an athlete experiences in moments of intense competition, and what sorts of internal and external conditions can create a feeling of “pressure”—witness the differences among conditions in the comparison of reward, presence of an audience, and evaluation by Mesagno, Harvey, and Janelle (2011).

This leads us to ask whether all the ways of creating what has been called a “pressure” situation (both real-world and laboratory) operate on states and processes in the same way—the answer at present appears to be no—and what if anything can be done to help sensorimotor-skilled performers avoid the negative effects of pressure while channeling the positive effects. To date, three answers to this question have been offered from a primarily cognitive perspective: (1) distract them while they are learning (Masters), (2) distract them while they are performing (Beilock, Masters), and (3) put them under pressure while they are learning, so they get used to it (Beilock).

Another set of answers has been offered from the phenomenological and clinical perspectives. These include (1) help performers view extremely demanding competitive situations as challenges rather than threats, and (2) help performers adopt avoidance strategies for ignoring or putting aside fearful and failure-oriented thoughts and feelings, rather than engaging and magnifying them. Each of these approaches has evidence to support it, so perhaps they are all true and their implications must be appreciated together. This recalls the old story about the blind men feeling the elephant—no one story was correct, but all together they had it right.

In some ways it may seem that the list of relevant factors considered in this chapter is so broad and disparate as to defy coherent integration or concerted investigation. Our belief is that new advances in both theoretical understanding of choking and clinical intervention in its management will require coherent integration and concerted investigation across the many factors and approaches we have raised and reviewed. This in turn will require a truly interdisciplinary approach in which philosophers, sport psychologists, cognitive scientists, personality researchers, and educational researchers learn one another’s languages and techniques and collaborate together. The research undertaken so far in each of these fields has made considerable progress, as we have tried to show. The next task is to try to bring these fields together to make a sum that can be greater than its individual parts.

—1
—0
—+1

Acknowledgments

This research on choking and embodied cognition was partly sponsored by a UPAR grant by UAE National Research Foundation Sport and Brain Science: Technological Applications for Peaking Performances (grant code: 31H087-UPAR (3) 2014).

Notes

1. Breivik (2007, 2013) and Birch, Fusché Moe, and Breivik (this volume) emphasize the need to overcome the stereotypical characterization of the expert performer as a mindless automata or consciousnessless zombie, incapable of deliberation or aware control during peak performance. Their accurate phenomenology of skillful action carefully reflects how self-awareness and attention are important components of peak performance. In dialogue with their proposal, Cappuccio (2017) suggests a nuanced account of consciousness in skillful performance that is informed by the concerns raised by Birch and colleagues, while recognizing that “flow” activities (Swann et al. 2012) and highly proceduralized execution routines are characterized by peculiar adumbrations of consciousness.
2. The internal/external distinction assumed by AFT has been criticized for its vagueness (for a review, see Montero, Toner, and Moran, this volume; a partial reply to some of these critical analyses is offered by Wulf 2016). A similar criticism might also be formulated against EFT, but it would not apply in the same way. The potential ambiguity of EFT, if any exists, concerns the proper analysis of the task’s control structure (as task analysis is not a fully-agreed-upon exercise), not what is “near”/“far from” the performing subject.
3. While EFT claims that explicit control of one’s own actions is detrimental only to peak performance of experts (not novices), AFT claims that internal focus universally undermines performance, regardless of skill level and expertise type (Wulf 2013, 2015), and that external foci, in turn, benefit both experienced professionals during competition and learning novices during training (Wulf, Lauterbach, and Toole 1999; Wulf, Shea, and Park 2001; Wulf et al. 2002).
4. While as described earlier, Mesagno, Harvey, and Janelle (2011) found no impact of physiologically expressed anxiety as measured by self-report, these results show that a direct measure of physiological state can predict performance-relevant responses to pressure.

References

- Abreu, Ana Maria, E. Macaluso, R. T. Azevedo, Paola Cesari, Cosimo Urgesi, and Salvatore Maria Aglioti. 2012. “Action Anticipation beyond the Action Observation Network: A Functional Magnetic Resonance Imaging Study in Expert Basketball Players.” *European Journal of Neuroscience* 35 (10): 1646–1654.
- Aglioti, Salvatore M., Paola Cesari, Michela Romani, and Cosimo Urgesi. 2008. “Action Anticipation and Motor Resonance in Elite Basketball Players.” *Nature Neuroscience* 11 (9): 1109–1116.

- Alden, Lynn E., Andrew G. Ryder, and Tanna M. B. Mellings. 2002. "Perfectionism in the Context of Social Fears: Toward a Two-Component Model." In *Perfectionism: Theory, Research and Treatment*, edited by G. L. Flett and P. L. Hewitt, 373–391. Washington, DC: American Psychological Association.
- Allen, Mark S., Daniel Frings, and Steve Hunter. 2012. "Personality, Coping, and Challenge and Threat States in Athletes." *International Journal of Sport and Exercise Psychology* 10 (4): 264–275.
- Anderson, John R. 1982. "Acquisition of Cognitive Skill." *Psychological Review* 89 (4): 369.
- Anshel, Mark H., and Robert S. Weinberg. 1999. "Re-Examining Coping among Basketball Referees Following Stressful Events: Implications for Coping Interventions." *Journal of Sport Behavior* 22 (2): 141–161.
- Aspinwall, L. G., and S. E. Taylor. 1997. "A Stitch in Time: Self-Regulation and Proactive Coping." *Psychological Bulletin* 121 (3): 417–436.
- Atkinson, John W. 1957. "Motivational Determinants of Risk-Taking Behavior." *Psychological Review* 64 (6): 359–372.
- Baumeister, Roy F. 1982. "A Self-Presentational View of Social Phenomena." *Psychological Bulletin* 91 (1): 3–26.
- Baumeister, Roy F. 1984. "Choking under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance." *Journal of Personality and Social Psychology* 46 (3): 610–620.
- Baumeister, Roy F., and Todd F. Heatherton. 1996. "Self-Regulation Failure: An Overview." *Psychological Inquiry* 7 (1): 1–15.
- Baumeister, Roy F., and Jill Lobbestael. 2011. "Emotions and Antisocial Behavior." *Journal of Forensic Psychiatry and Psychology* 22 (5): 635–649.
- Baumeister, Roy F., and S. J. Scher. 1988. "Self-Defeating Behavior Patterns among Normal Individuals: Review and Analysis of Common Self-Destructive Tendencies." *Psychological Bulletin* 104 (1): 3–22.
- Baumeister, Roy F., and Carolin J. Showers. 1986. "A Review of Paradoxical Performance Effects: Choking under Pressure in Sports and Mental Tests." *European Journal of Social Psychology* 16 (4): 361–383.
- Baumeister, Roy F., and Kathleen D. Vohs. 2001. "Narcissism as Addiction to Esteem." *Psychological Inquiry* 12 (4): 206–210.
- Baumeister, Roy F., Kathleen D. Vohs, C. Nathan DeWall, and Liqing Zhang. 2007. "How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather than Direct Causation." *Personality and Social Psychology Review* 11 (2): 167–203.
- Beilock, Sian. 2015. *How The Body Knows Its Mind*. London: Hachette UK.
- Beilock, Sian L. 2008a. "Beyond the Playing Field: Sport Psychology Meets Embodied Cognition." *International Review of Sport and Exercise Psychology* 1 (1): 19–30.

Beilock, Sian L. 2008b. "Math Performance in Stressful Situations." *Current Directions in Psychological Science* 17 (5): 339–343.

Beilock, Sian L. 2011. *Choke*. Boston: Little, Brown.

Beilock, Sian L., Bennett I. Bertenthal, Michael Hoerger, and Thomas H. Carr. 2008. "When Does Haste Make Waste? Speed-Accuracy Tradeoff, Skill Level, and the Tools of the Trade." *Journal of Experimental Psychology: Applied* 14 (4): 340–352.

Beilock, Sian L., Bennett I. Bertenthal, Annette M. McCoy, and Thomas H. Carr. 2004. "Haste Does Not Always Make Waste: Expertise, Direction of Attention, and Speed versus Accuracy in Performing Sensorimotor Skills." *Psychonomic Bulletin and Review* 11 (2): 373–379.

Beilock, Sian L., and Thomas H. Carr. 2001. "On the Fragility of Skilled Performance: What Governs Choking under Pressure?" *Journal of Experimental Psychology: General* 130 (4): 701–725.

Beilock, Sian L., Thomas H. Carr, Clare MacMahon, and Janet L. Starkes. 2002. "When Paying Attention Becomes Counterproductive: Impact of Divided versus Skill-Focused Attention on Novice and Experienced Performance of Sensorimotor Skills." *Journal of Experimental Psychology: Applied* 8 (1): 6–16.

Beilock, Sian L., and Marci S. Decaro. 2007. "From Poor Performance to Success under Stress: Working Memory, Strategy Selection, and Mathematical Problem Solving under Pressure." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33 (6): 983–998.

Beilock, Sian L. and Rob Gray. 2007. "Why Do Athletes Choke under Pressure?" In *Handbook of Sport Psychology*, 3rd ed., edited by Gerson Tenenbaum and Bob Eklund, 425–444. Hoboken, NJ: John Wiley.

Beilock, S. L., and R. Gray. 2012. "From Attentional Control to Attentional Spillover: A Skill-Level Investigation of Attention, Movement, and Performance Outcomes." *Human Movement Science* 31 (6): 1473–1499.

Beilock, Sian L., William A. Jellison, Robert J. Rydell, Allen R. McConnell, and Thomas H. Carr. 2006. "On the Causal Mechanisms of Stereotype Threat: Can Skills That Don't Rely Heavily on Working Memory Still Be Threatened?" *Personality and Social Psychology Bulletin* 32 (8): 1059–1071.

Beilock, Sian L., Catherine A. Kulp, Lauren E. Holt, and Thomas H. Carr. 2004. "More on the Fragility of Performance: Choking under Pressure in Mathematical Problem Solving." *Journal of Experimental Psychology: General* 133 (4): 584–600.

Beilock, Sian L., Ian M. Lyons, Andrew Mattarella-Micke, Howard C. Nusbaum, and Steven L. Small. 2008b. "Sports Experience Changes the Neural Processing of Action Language." *Proceedings of the National Academy of Sciences of the United States of America* 105 (36): 13269–13273.

Bhalla, Mukul, and Dennis R. Proffitt. 1999. "Visual—Motor Recalibration in Geographical Slant Perception." *Journal of Experimental Psychology: Human Perception and Performance* 25 (4): 1076.

- Blascovich, Jim. 2008. "Challenge, Threat, and Health." In *Handbook of Motivation Science*, edited by James Y. Shah and Wendi L. Gardner, 481–493. New York: Guilford Press.
- Blascovich, Jim, and Wendy B. Mendes. 2000. "Challenge and Threat Appraisals: The Role of Affective Cues." In *Feeling and Thinking: The Role of Affect in Social Cognition*, edited by J. Forgas, 59–82. Cambridge: Cambridge University Press.
- Blascovich, Jim, Wendy Berry Mendes, Sarah B. Hunter, and Brian Lickel. 2000. "Stigma, Threat, and Social Interactions." In *The Social Psychology of Stigma*, edited by T. Heatherton, R. Kleck, M. R. Hebl, and J. G. Hull, 307–333. New York: Guilford.
- Blascovich, Jim, and Joe Tomaka. 1996. "The Biopsychosocial Model of Arousal Regulation." *Advances in Experimental Social Psychology* 28: 1–51.
- Bradley, Jeff. 2003. "The Flash." *ESPN Magazine*, May 27. <http://www.espn.com/magazine/vol5no11chiro.html>
- Bray, S. R., K. A. Martin, and W. N. Widmeyer. 2000. "The Relationship between Evaluative Concerns and Sport Competition State Anxiety among Youth Skiers." *Journal of Sports Sciences* 18 (5): 353–361.
- Breivik, Gunnar. 2007. "Skillful Coping in Everyday Life and in Sport: A Critical Examination of the Views of Heidegger and Dreyfus." *Journal of the Philosophy of Sport* 34 (2): 116–134.
- Breivik, Gunnar. 2013. "Zombie-Like or Superconscious? A Phenomenological and Conceptual Analysis of Consciousness in Elite Sport." *Journal of the Philosophy of Sport* 40 (1): 85–106.
- Brewer, Britton W. 1993. "Self-Identity and Specific Vulnerability to Depressed Mood." *Journal of Personality* 61 (3): 343–364.
- Brewer, Britton W., Judy L. Van Raalte, and Darwyn E. Linder. 1993. "Athletic Identity: Hercules' Muscles or Achilles Heel?" *International Journal of Sport Psychology* 24 (2): 237–254.
- Brown, Tracy L., and Thomas H. Carr. 1989. "Automaticity in Skill Acquisition: Mechanisms for Reducing Interference in Concurrent Performance." *Journal of Experimental Psychology: Human Perception and Performance* 15 (4): 686–700.
- Bruner, Jerome S. 1957. "On Perceptual Readiness." *Psychological Review* 64 (2): 123–152.
- Bruner, Jerome S. 1966. *Toward a Theory of Instruction*. Cambridge, MA: Harvard University Press.
- Calvo, Manuel G., Leopoldo Alamo, and Pedro M. Ramos. 1990. "Test Anxiety, Motor Performance and Learning: Attentional and Somatic Interference." *Personality and Individual Differences* 11 (1): 29–38.
- Cañal-Bruland, Rouwen, J. R. Pijpers, and Raoul R. D. Oudejans. 2010. "The Influence of Anxiety on Action-Specific Perception." *Anxiety, Stress, and Coping* 23 (3): 353–361.
- Cañal-Bruland, Rouwen, and John van der Kamp. 2009. "Action Goals Influence Action-Specific Perception." *Psychonomic Bulletin and Review* 16 (6): 1100–1105.

—1
—0
—+1

- Cañal-Bruland, Rouwen, Frank F. Zhu, John van der Kamp, and Rich S. W. Masters. 2011. "Target-Directed Visual Attention Is a Prerequisite for Action-Specific Perception." *Acta Psychologica* 136 (3): 285–289.
- Cappuccio, Massimiliano L., ed. 2015. "Unreflective Action and the Choking Effect." *Phenomenology and the Cognitive Sciences* 14 (2).
- Cappuccio, Massimiliano L. 2016. "Mind-Upload. The Ultimate Challenge to the Embodied Mind Theory." *Phenomenology and the Cognitive Sciences* 16 (3): 1–24.
- Cappuccio, Massimiliano L. 2017. "Flow, Choke, Skill: The Role of the Non-Conscious in Sport Performance." In *Before Consciousness: In Search of the Fundamentals of Mind*, edited by Zdravko Radman, 246–283. Exeter, UK: Imprint Academic.
- Carleton, R. Nicholas, Donald R. McCreary, Peter J. Norton, and Gordon J. G. Asmundson. 2006. "Brief Fear of Negative Evaluation Scale-Revised." *Depression and Anxiety* 23 (5): 297–303.
- Carr, Thomas H. 2015. "Strengths and Weaknesses of Reflection as a Guide to Action: Pressure Assails Performance in Multiple Ways." *Phenomenology and the Cognitive Sciences* 14 (2): 227–252.
- Carver, Charles. 2003. "Pleasure as a Sign You Can Attend to Something Else: Placing Positive Feelings within a General Model of Affect." *Cognition and Emotion* 17 (2): 241–261.
- Carver, Charles S., and Michael F. Scheier. 1981. *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*. Springer Series in Social Psychology. New York: Springer.
- Castaneda, Brooke, and Rob Gray. 2007. "Effects of Focus of Attention on Baseball Batting Performance in Players of Differing Skill Levels." *Journal of Sport and Exercise Psychology* 29 (1): 60–77.
- Christensen, Wayne, John Sutton, and Doris McIlwain. 2015. "Putting Pressure on Theories of Choking: Towards an Expanded Perspective on Breakdown in Skilled Performance." *Phenomenology and the Cognitive Sciences* 14 (2): 253–293.
- Christensen, Wayne, John Sutton, and Doris J. F. McIlwain. 2016. "Cognition in Skilled Action: Meshed Control and the Varieties of Skill Experience." *Mind & Language* 31 (1): 37–66.
- Clark, Andy. 2015. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.
- Clark, David M., and Adrian Wells. 1995. "A Cognitive Model of Social Phobia." *Social Phobia: Diagnosis, Assessment, and Treatment* 41 (68): 00022–00023.
- Coombes, Stephen A., Torrie Higgins, Kelly M. Gamble, James H. Cauraugh, and Christopher M. Janelle. 2009. "Attentional Control Theory: Anxiety, Emotion, and Motor Planning." *Journal of Anxiety Disorders* 23 (8): 1072–1079.
- Cosmelli, Diego, and Evan Thompson. 2010. "Embodiment or Envatment? Reflections on the Bodily Basis of Consciousness." In *Enaction: Towards a New Paradigm for Cognitive Science*, edited by J. Stewart, O. Gapenne, and E. Di Paolo, 361–385. Cambridge, MA: MIT Press.

- Crocker, Peter R. E., and Thomas R. Graham. 1995. "Emotion in Sport and Physical Activity: The Importance of Perceived Individual Goals." *International Journal of Sport Psychology* 26 (1): 117–137.
- Crook, Jan M., and Sharon E. Robertson. 1991. "Transitions Out of Elite Sport." *International Journal of Sport Psychology* 22 (2): 115–127.
- Csikszentmihalyi, Mihaly. 1996. *Flow and the Psychology of Discovery and Invention*. New York: Harper Collins.
- Dandy, J., N. Brewer, and R. Tottman. 2001. "Self-Consciousness and Performance Decrements within a Sporting Context." *Journal of Social Psychology* 141 (1): 150–152.
- Davids, Keith, Duarte Araújo, Robert Hristovski, Pedro Passos, and Jia Yi Chow. 2012. "Ecological Dynamics and Motor Learning Design in Sport." In *Skill Acquisition in Sport: Research, Theory and Practice*, 2nd ed., edited by M. Williams and N. Hodges, 112–130. London: Routledge.
- DeCaro, Marci S., Robin D. Thomas, Neil B. Albert, and Sian L. Beilock. 2011. "Choking under Pressure: Multiple Routes to Skill Failure." *Journal of Experimental Psychology: General* 140 (3): 390–406.
- Dreyfus, Hubert L. 1996. "The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment." *Electronic Journal of Analytic Philosophy* 4 (Spring): 1–16.
- Dreyfus, Hubert L. 2002. "Intelligence without Representation—Merleau-Ponty's Critique of Mental Representation, the Relevance of Phenomenology to Scientific Explanation." *Phenomenology and the Cognitive Sciences* 1 (4): 367–383.
- Dreyfus, Hubert L. 2012. "Introductory Essay: The Mystery of the Background qua Background." In *Knowing without Thinking*, edited by Zdravko Radman, 1–10. New Directions in Philosophy and Cognitive Science. Basingstoke, UK: Palgrave Macmillan.
- Dreyfus, Stuart E., and Hubert L. Dreyfus. 1980. "A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition." Berkeley: University of California Operations Research Center.
- Erez, Amir, and Alice M. Isen. 2002. "The Influence of Positive Affect on the Components of Expectancy Motivation." *Journal of Applied Psychology* 87 (6): 1055–1067.
- Evans, Jonathan B. T., and Keith E. Stanovich. 2013. "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 8 (3): 223–241.
- Eysenck, Michael W., and Manuel G. Calvo. 1992. "Anxiety and Performance: The Processing Efficiency Theory." *Cognition and Emotion* 6 (6): 409–434.
- Eysenck, Michael W., Nazanin Derakshan, Rita Santos, and Manuel G. Calvo. 2007. "Anxiety and Cognitive Performance: Attentional Control Theory." *Emotion* 7 (2): 336–353.
- Fenigstein, Allan, Michael F. Scheier, and Arnold H. Buss. 1975. "Public and Private Self-Consciousness: Assessment and Theory." *Journal of Consulting and Clinical Psychology* 43 (4): 522–527.

—1
—0
—+1

- Ferrari, Pier Francesco, and Giacomo Rizzolatti. 2015. *New Frontiers in Mirror Neurons Research*. Oxford: Oxford University Press.
- Fink, Philip W., Patrick S. Foo, and William H. Warren. 2009. "Catching Fly Balls in Virtual Reality: A Critical Test of the Outfielder Problem." *Journal of Vision* 9 (13): 14.1–8.
- Fitts, P. M., and M. I. Posner. 1967. *Human Performance*. Pacific Grove, CA: Brooks/Cole.
- Flett, Gordon L., and Paul L. Hewitt. 2005. "The Perils of Perfectionism in Sports and Exercise." *Current Directions in Psychological Science* 14 (1): 14–18.
- Fredrickson, B. L. 2001. "The Role of Positive Emotion in Positive Psychology: The Broaden-and-Build Theory of Positive Emotion." *American Psychologist* 56:218–226.
- Fredrickson, Barbara L. 2003. "The Value of Positive Emotions." *American Scientist* 91 (4): 330–335.
- Frost, Randy O., and Katherine J. Henderson. 1991. "Perfectionism and Reactions to Athletic Competition." *Journal of Sport and Exercise Psychology* 13 (4): 323–335.
- Frost, Randy O., and Patricia A. Marten. 1990. "Perfectionism and Evaluative Threat." *Cognitive Therapy and Research* 14 (6): 559–572.
- Gallagher, Shaun. 2005. *How the Body Shapes the Mind*. New York: Oxford University Press.
- Gallagher, Shaun. 2011. "Interpretations of Embodied Cognition." In *The Implications of Embodiment: Cognition and Communication*, edited by W. Tschacher and C. Bergomi, 59–70. Exeter, UK: Imprint Academic.
- Gallese, Vittorio. 2007. "The Shared Manifold Hypothesis: Embodied Simulation and Its Role in Empathy and Social Cognition." In *Empathy in Mental Illness and Health*, edited by Tom F. D. Farrow and Peter W. R. Woodruff, 448–472. Cambridge: Cambridge University Press.
- Geukes, Katharina, Jack T. Harvey, Alexandra Trezise, and Christopher Mesagno. 2017. "Personality and Performance in Real-World Competitions: Testing Trait Activation of Fear of Negative Evaluation, Dispositional Reinvestment, and Athletic Identity in the Field." *Psychology of Sport and Exercise* 30:101–109.
- Geukes, Katharina, Christopher Mesagno, Stephanie J. Hanrahan, and Michael Kellmann. 2012. "Testing an Interactionist Perspective on the Relationship between Personality Traits and Performance under Public Pressure." *Psychology of Sport and Exercise* 13 (3): 243–250.
- Geukes, Katharina, Christopher Mesagno, Stephanie J. Hanrahan, and Michael Kellmann. 2013. "Activation of Self-Focus and Self-Presentation Traits under Private, Mixed, and Public Pressure." *Journal of Sport and Exercise Psychology* 35 (1): 50–59.
- Gibson, James J. 1977. "The Theory of Affordances." In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, edited by Robert Shaw and John Bransford, 67–82. New York: Routledge.
- Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

- Gibson, James Jerome. 1966. *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.
- Gray, Rob. 2004. "Attending to the Execution of a Complex Sensorimotor Skill: Expertise Differences, Choking, and Slumps." *Journal of Experimental Psychology: Applied* 10 (1): 42–54.
- Gray, Rob. 2011. "Links between Performance Outcomes, Attention and Movement in Skilled Motor Action." *Current Directions in Psychological Science* 20:301–306.
- Gray, Rob. 2013. "Being Selective at the Plate: Processing Dependence between Perceptual Variables Relates to Hitting Goals and Performance." *Journal of Experimental Psychology: Human Perception and Performance* 39 (4): 1124–1142. doi:10.1037/a0030729.
- Gray, Rob. 2014. "Embodied Perception in Sport." *International Review of Sport and Exercise Psychology* 7 (1): 72–86.
- Gray, Rob, and Jonathan Allsop. 2013. "Interactions between Performance Pressure, Performance Streaks, and Attentional Focus." *Journal of Sport and Exercise Psychology* 35 (4): 368–386.
- Gray, Rob, Jonathan Allsop, and Sarah Williams. 2013. "Changes in Putting Kinematics Associated with Choking and Excelling under Pressure." *International Journal of Sport Psychology* 44 (4): 387–407.
- Gray, Rob, and Rouwen Cañal-Bruland. 2015. "Attentional Focus, Perceived Target Size, and Movement Kinematics under Performance Pressure." *Psychonomic Bulletin and Review* 22 (6): 1692–1700.
- Gucciardi, Daniel F., Jay-Lee Longbottom, Ben Jackson, and James A. Dimmock. 2010. "Experienced Golfers' Perspectives on Choking under Pressure." *Journal of Sport and Exercise Psychology* 32 (1): 61–83.
- Hanin, Yuri L. 2000. "Successful and Poor Performance and Emotions." In *Emotions in Sport*, edited by Yuri L. Hanin, 157–187. Champaign, IL: Human Kinetics.
- Hanin, Yuri L. 2007. "Emotions in Sport: Current Issues and Perspectives." In *Handbook of Sport Psychology*, edited by Gershon Tenenbaum and Robert C. Eklund, 31–58. Hoboken, NJ: John Wiley.
- Hill, Denise M., Sheldon Hanton, Scott Fleming, and Nic Matthews. 2009. "A Re-Examination of Choking in Sport." *European Journal of Sport Science* 9 (4): 203–212.
- Hill, Denise M., Sheldon Hanton, Nic Matthews, and Scott Fleming. 2010a. "Choking in Sport: A Review." *International Review of Sport and Exercise Psychology* 3 (1): 24–39.
- Hill, Denise M., Sheldon Hanton, Nic Matthews, and Scott Fleming. 2010b. "A Qualitative Exploration of Choking in Elite Golf." *Journal of Clinical Sport Psychology* 4 (3): 221–240.
- Hill, Denise M., Sheldon Hanton, Nic Matthews, and Scott Fleming. 2011. "Alleviation of Choking under Pressure in Elite Golf: An Action Research Study." *Sport Psychologist* 25 (4): 465–488.

—1
—0
—+1

- Hill, Denise M., and Brian Hemmings. 2015. "A Phenomenological Exploration of Coping Responses Associated with Choking in Sport." *Qualitative Research in Sport, Exercise and Health* 7 (4): 521–538. doi.org/10.1080/2159676X.2014.981573.
- Hill, Denise M., and Gareth Shaw. 2013. "A Qualitative Examination of Choking under Pressure in Team Sport." *Psychology of Sport and Exercise* 14 (1): 103–110.
- Hostetter, Autumn B., and Martha W. Alibali. 2008. "Visible Embodiment: Gestures as Simulated Action." *Psychonomic Bulletin and Review* 15 (3): 495–514.
- Hudson, Joanne, and Matthew Williams. 2001. "Associations between Self-Presentation and Competitive A-Trait: A Preliminary Investigation." *Social Behavior and Personality: An International Journal* 29 (1): 1–10.
- Jackson, Robin C., Kelly J. Ashford, and Glen Norsworthy. 2006. "Attentional Focus, Dispositional Reinvestment, and Skilled Motor Performance under Pressure." *Journal of Sport and Exercise Psychology* 28 (1): 49–68.
- Jackson, Robin C., and Sian L. Beilock. 2008. "Performance Pressure and Paralysis by Analysis: Research and Implications." In *Developing Sports Expertise: Researchers and Coaches Put Theory into Practice*, edited by Damian Farrow, Joseph Baker, and Clare MacMahon, 104–118. London: Routledge.
- Jackson, Robin C., Sian L. Beilock, and Noel P. Kinrade. 2013. "'Choking' in Sport: Research and Implications." In *Developing Sport Expertise: Researchers and Coaches Put Theory in to Practice*, 2nd ed., edited by Damian Farrow, Joseph Baker, and Clare MacMahon, 177–194. London: Routledge.
- Jackson, Susan A., and Mihaly Csikszentmihalyi. 1999. *Flow in Sports*. Champaign, IL: Human Kinetics.
- James, Benjamin, and David Collins. 1997. "Self-Presentational Sources of Competitive Stress during Performance." *Journal of Sport and Exercise Psychology* 19 (1): 17–35.
- Jones, G. 1995. "More than Just a Game: Research Developments and Issues in Competitive Anxiety in Sport." *British Journal of Psychology* 86 (Pt 4) (November): 449–478.
- Jones, Marc, Carla Meijen, Paul Joseph McCarthy, and David Sheffield. 2009. "A Theory of Challenge and Threat States in Athletes." *International Review of Sport and Exercise Psychology* 2 (2): 161–180.
- Jordet, Geir. 2009. "Why Do English Players Fail in Soccer Penalty Shootouts? A Study of Team Status, Self-Regulation, and Choking under Pressure." *Journal of Sports Sciences* 27 (2): 97–106.
- Jordet, Geir. 2010. "Choking under Pressure as Self-Destructive Behavior." In *Coping in Sport: Theory, Methods, and Related Constructs*, edited by Adam R. Nicholls, 239–259. New York: Nova Science.
- Jordet, Geir, and Esther Hartman. 2008. "Avoidance Motivation and Choking under Pressure in Soccer Penalty Shootouts." *Journal of Sport and Exercise Psychology* 30 (4): 450–457.

- Kerr, John H. 2014. *Motivation and Emotion in Sport: Reversal Theory*. London: Psychology Press.
- Kinrade, Noel P., Robin C. Jackson, and Kelly J. Ashford. 2010. "Dispositional Reinvestment and Skill Failure in Cognitive and Motor Tasks." *Psychology of Sport and Exercise* 11 (4): 312–319.
- Kirsh, David, and Paul Maglio. 1994. "On Distinguishing Epistemic from Pragmatic Action." *Cognitive Science* 18 (4): 513–549.
- Kurosawa, Kaoru, and Judith M. Harackiewicz. 1995. "Test Anxiety, Self-Awareness, and Cognitive Interference: A Process Analysis." *Journal of Personality* 63 (4): 931–951.
- Lane, A. M., C. J. Beedie, T. J. Devonport, and D. M. Stanley. 2011. "Instrumental Emotion Regulation in Sport: Relationships between Beliefs about Emotion and Emotion Regulation Strategies Used by Athletes." *Scandinavian Journal of Medicine and Science in Sports* 21 (6): e445–451.
- Lavallee, David, Sandy Gordon, and J. Robert Grove. 1997. "Retirement from Sport and the Loss of Athletic Identity." *Journal of Personal and Interpersonal Loss* 2 (2): 129–147.
- Lazarus, Richard S. 2000. "Cognitive-Motivational-Relational Theory of Emotion." In *Emotions in Sport*, edited by Yuri L. Hanin, 39–63. Champaign, IL: Human Kinetics.
- Leary, Mark R. 1992. "Self-Presentational Processes in Exercise and Sport." *Journal of Sport and Exercise Psychology* 14 (4): 339–351.
- Leary, Mark R. 2001. "Social Anxiety as an Early Warning System: A Refinement and Extension of the Self-Presentation Theory of Social Anxiety." In *From Social Anxiety to Social Phobia: Multiple Perspectives*, edited by Stefan G. Hoffman and Patricia Marten DiBartolo, 321–334. Needham Heights, MA: Allyn and Bacon.
- Lorimer, Ross. 2006. "The Relationship between Self-Presentational Concerns and Competitive Anxiety: The Influence of Gender." *International Journal of Sport and Exercise Psychology* 37 (4): 317–329.
- Marchant, Daryl B., Tony Morris, and Mark B. Andersen. 1998. "Perceived Importance of Outcome as a Contributing Factor in Competitive State Anxiety." *Journal of Sport Behavior* 21 (1): 71–91.
- Martin, Jeffrey J., Robert C. Eklund, and Carol Adams Mushett. 1997. "Factor Structure of the Athletic Identity Measurement Scale with Athletes with Disabilities." *Adapted Physical Activity Quarterly* 14 (1): 74–82.
- Masters, Richard S. W. 1992. "Knowledge, Knerves and Know-How: The Role of Explicit versus Implicit Knowledge in the Breakdown of a Complex Motor Skill under Pressure." *British Journal of Psychology* 83 (3): 343–358.
- Masters, Richard S. W., and J. P. Maxwell. 2004. "Implicit Motor Learning, Reinvestment and Movement Disruption: What You Don't Know Won't Hurt You?" In *Skill acquisition in Sport: Research, Theory and Practice*, edited by A. M. Williams and N. J. Hodges, 207–228. London: Routledge.

- Masters, Richard S. W., and Jon P. Maxwell. 2008. "The Theory of Reinvestment." *International Review of Sport and Exercise Psychology* 1 (2): 160–183.
- Masters, Richard S. W., R. C. J. Polman, and N. V. Hammond. 1993. "'Reinvestment': A Dimension of Personality Implicated in Skill Breakdown under Pressure." *Personality and Individual Differences* 14 (5): 655–666.
- Maxwell, Jon P., Richard S. Masters, and F. F. Eves. 2000. "From Novice to No Know-How: A Longitudinal Study of Implicit Motor Learning." *Journal of Sports Sciences* 18 (2): 111–120.
- McCarthy, Paul J. 2011. "Positive Emotion in Sport Performance: Current Status and Future Directions." *International Review of Sport and Exercise Psychology* 4 (1): 50–69.
- McEnroe, John, and James Kaplan. 2002. *You Cannot Be Serious*. London: Penguin.
- Meijen, Carla, Marc V. Jones, Paul J. McCarthy, David Sheffield, and Mark S. Allen. 2013. "Cognitive and Affective Components of Challenge and Threat States." *Journal of Sports Sciences* 31 (8): 847–855.
- Mendes, Wendy Berry, Brenda Major, Shannon McCoy, and Jim Blascovich. 2008. "How Attributional Ambiguity Shapes Physiological and Emotional Responses to Social Rejection and Acceptance." *Journal of Personality and Social Psychology* 94 (2): 278–291.
- Mesagno, C., and D. M. Hill. 2013. "Definition of Choking in Sport: Re-Conceptualization and Debate." *International Journal of Sport Psychology* 44 (4): 267–277.
- Mesagno, Christopher. 2009. "Choking under Pressure: Toward a Self-Presentation Explanation of Why Athletes Use Self-Monitoring Techniques." Paper presented at the 12th World Congress of Sport Psychology. Marrakesh, Morocco.
- Mesagno, Christopher, and Jürgen Beckmann. 2017. "Choking under Pressure: Theoretical Models and Interventions." *Current Opinions in Psychology* 16:170–175.
- Mesagno, Christopher, Katharina Geukes, and Paul Larkin. 2015. "Choking under Pressure: A Review of Current Debates, Literature, and Interventions." In *Contemporary Advances in Sport Psychology*, edited by S. D. Mellalieu and S. Hanton, 148–174. Abingdon, UK: Routledge.
- Mesagno Christopher, Jack T. Harvey, and Christopher M. Janelle. 2011. "Self-Presentation Origins of Choking: Evidence from Separate Pressure Manipulations." *Journal of Sport and Exercise Psychology* 33:441–459.
- Mesagno Christopher, Jack T. Harvey, and Christopher M. Janelle. 2012. "Choking under Pressure: The Role of Fear of Negative Evaluation." *Psychology of Sport and Exercise* 13:60–68.
- Mesagno, Christopher, Daryl Marchant, and Tony Morris. 2009. "Alleviating Choking: The Sounds of Distraction." *Journal of Applied Sport Psychology* 21 (2): 131–147.
- Montero, Barbara G. 2010. "Does Bodily Awareness Interfere with Highly Skilled Movement?" *Inquiry: A Journal of Medical Care Organization, Provision and Financing* 53 (2): 105–122.

- Montero, Barbara G. 2014. "Is Monitoring One's Actions Causally Relevant to Choking under Pressure?" *Phenomenology and the Cognitive Sciences* 14 (2): 379–395.
- Montero, Barbara G. 2016. *Thought in Action: Expertise and the Conscious Mind*. Oxford: Oxford University Press.
- Moore, Lee J., Samuel J. Vine, Paul Freeman, and Mark R. Wilson. 2013. "Quiet Eye Training Promotes Challenge Appraisals and Aids Performance under Elevated Anxiety." *International Journal of Sport and Exercise Psychology* 11 (2): 169–183.
- Moore, Lee J., Samuel J. Vine, Mark R. Wilson, and Paul Freeman. 2012. "The Effect of Challenge and Threat States on Performance: An Examination of Potential Mechanisms." *Psychophysiology* 49 (10): 1417–1425.
- Moore, Lee J., Samuel J. Vine, Mark R. Wilson, and Paul Freeman. 2015. "Reappraising Threat: How to Optimize Performance under Pressure." *Journal of Sport and Exercise Psychology* 37 (3): 339–343.
- Moore, Lee J., Mark R. Wilson, Samuel J. Vine, Adam H. Coussens, and Paul Freeman. 2013. "Champ or Chump? Challenge and Threat States during Pressurized Competition." *Journal of Sport and Exercise Psychology* 35 (6): 551–562.
- Neil, Rich, Sheldon Hanton, Stephen D. Mellalieu, and David Fletcher. 2011. "Competition Stress and Emotions in Sport Performers: The Role of Further Appraisals." *Psychology of Sport and Exercise* 12 (4): 460–470.
- Nobles, Charlie. 1995. "Basketball: Scott Fires First Shots in Magic Arsenal." *New York Times*, Sports, May 25.
- O'Connor, Ian. 2016. "Jordan Spieth's Collapse at the Masters the Most Shocking in Golf History." *ESPN Golf*, April 11. http://www.espn.co.uk/golf/masters16/story/_/id/15176844/jordan-spieth-collapse-masters-most-shocking-golf-history.
- Otten, Mark. 2009. "Choking vs. Clutch Performance: A Study of Sport Performance under Pressure." *Journal of Sport and Exercise Psychology* 31 (5): 583–601.
- Oudejans, Raoul R. D., Wilma Kuijpers, Chris C. Kooijman, and Frank C. Bakker. 2011. "Thoughts and Attention of Athletes under Pressure: Skill-Focus or Performance Worries?" *Anxiety, Stress, and Coping* 24 (1): 59–73.
- Piaget, Jean. 1971. *The Theory of Stages in Cognitive Development*. New York: McGraw-Hill.
- Piaget, Jean. 1977. "The Role of Action in the Development of Thinking." In *Knowledge and Development*, edited by Willis F. Overton and Jeanette McCarthy Gallagher, 17–42. New York: Springer.
- Poolton, Jamie, Jon Maxwell, and Rich Masters. 2004. "Rules for Reinvestment." *Perceptual and Motor Skills* 99 (3 Pt 1): 771–774.
- Posner, Michael I., and Charles R. Snyder. 1975. "Attention and Cognitive Control." In *Information Processing and Cognition: The Loyola Symposium*, edited by R. L. Solso, 55–82. Hillsdale, NJ: Erlbaum.

—1
—0
—+1

- Proffitt, Dennis R. 2006. "Embodied Perception and the Economy of Action." *Perspectives on Psychological Science* 1 (2): 110–122.
- Rizzolatti, Giacomo, and Corrado Sinigaglia. 2007. "Mirror Neurons and Motor Intentionality." *Functional Neurology* 22 (4): 205–210.
- Rucinska, Zuzanna. 2014a. "Basic Pretending as Sensorimotor Engagement?" In *Contemporary Sensorimotor Theory* 15:175–187.
- Rucinska, Zuzanna. 2014b. "Radically Enacted Creativity in Sports." First International Conference on Sport Psychology and Embodied Cognition. Al Ain UAE, February.
- Scheier, Michael F., and Charles S. Carver. 1985. "The Self-Consciousness Scale: A Revised Version for Use with General Populations." *Journal of Applied Social Psychology* 15 (8): 687–699.
- Schlenker, B. R., and M. R. Leary. 1982. "Social Anxiety and Self-Presentation: A Conceptualization and Model." *Psychological Bulletin* 92 (3): 641–669.
- Schlenker, Barry R. 1980. *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations*. Pacific Grove, CA: Brooks/Cole.
- Schnall, Simone, Jonathan R. Zadra, and Dennis R. Proffitt. 2010. "Direct Evidence for the Economy of Action: Glucose and the Perception of Geographical Slant." *Perception* 39 (4): 464–482.
- Shernoff, David J., Mihaly Csikszentmihalyi, Barbara Shneider, and Elisa Steele Shernoff. 2003. "Student Engagement in High School Classrooms from the Perspective of Flow Theory." *School Psychology Quarterly* 18 (2): 158.
- Skinner, Natalie, and Neil Brewer. 2004. "Adaptive Approaches to Competition: Challenge Appraisals and Positive Emotion." *Journal of Sport and Exercise Psychology* 26 (2): 283–305.
- Spielberger, Charles D. 1966. "Theory and Research on Anxiety." In *Anxiety and Behavior*, edited by Charles D. Spielberger, 3–20. Amsterdam: Elsevier.
- Spielberger, Charles D., William D. Anton, and Jeffrey Bedell. 1976. "The Nature and Treatment of Test Anxiety." In *Emotions and Anxiety: New Concepts, Methods, and Applications*, edited by C. D. Zuckerman and C. D. Spielberger, 317–345. New York: Erlbaum.
- Sutton, John. 2007. "Batting, Habit and Memory: The Embodied Mind and the Nature of Skill." *Sport in Society* 10 (5): 763–786.
- Sutton, John. 2008. "Material Agency, Skills and History: Distributed Cognition and the Archaeology of Memory." In *Material Agency*, edited by Carl Knappett and Lambros Malafouris, 37–55. New York: Springer.
- Sutton, John, Doris McIlwain, Wayne Christensen, and Andrew Geeves. 2011. "Applying Intelligence to the Reflexes: Embodied Skills and Habits between Dreyfus and Descartes." *Journal of the British Society for Phenomenology* 42 (1): 78–103.

- Swann, Christian, Richard Keegan, Lee Crust, and David Piggott. 2016. "Psychological States Underlying Excellent Performance in Professional Golfers: 'Letting It Happen' vs. 'Making It Happen.'" *Psychology of Sport and Exercise* 23 (March): 101–113.
- Swann, Christian, Richard J. Keegan, David Piggott, and Lee Crust. 2012. "A Systematic Review of the Experience, Occurrence, and Controllability of Flow States in Elite Sport." *Psychology of Sport and Exercise* 13 (6): 807–819.
- Tenenbaum, Gershon, and Michael Bar-Eli. 1993. "Decision Making in Sport: A Cognitive Perspective." In *Handbook of Research on Sport Psychology*, edited by R. N. Singer, 171–192. New York: Macmillan library reference.
- Thomas, Owen, Ian Maynard, and Sheldon Hanton. 2007. "Intervening with Athletes during the Time Leading up to Competition: Theory to Practice II." *Journal of Applied Sport Psychology* 19 (4): 398–418.
- Toner, John, Barbara Gail Montero, and Aidan Moran. 2014. "Considering the Role of Cognitive Control in Expert Performance." *Phenomenology and the Cognitive Sciences* 14 (4): 1127–1144.
- Toner, John, Barbara Gail Montero, and Aidan Moran. 2015. "The Perils of Automaticity." *Review of General Psychology: Journal of Division 1, of the American Psychological Association* 19 (4): 431–442.
- Toner, John, Barbara Gail Montero, and Aidan Moran. 2016. "Reflective and Prereflective Bodily Awareness in Skilled Action." *Psychology of Consciousness: Theory, Research, and Practice* 3 (4): 303–315. doi:10.1037/cns0000090.
- Toner, John, and Aidan Moran. 2015. "Enhancing Performance Proficiency at the Expert Level: Considering the Role of 'Somaesthetic Awareness.'" *Psychology of Sport and Exercise* 16 (January): 110–117.
- Totterdell, P. 2000. "Catching Moods and Hitting Runs: Mood Linkage and Subjective Performance in Professional Sport Teams." *Journal of Applied Psychology* 85 (6): 848–859.
- Turing, Alan Mathison. 1937. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society, Third Series* 2 (1): 230–265.
- Turner, Martin J., Marc V. Jones, David Sheffield, and Sophie L. Cross. 2012. "Cardiovascular Indices of Challenge and Threat States Predict Competitive Performance." *International Journal of Psychophysiology* 86 (1): 48–57.
- Uphill, Mark, Ryan Groom, and Marc Jones. 2014. "The Influence of In-Game Emotions on Basketball Performance." *European Journal of Sport Science: Official Journal of the European College of Sport Science* 14 (1): 76–83.
- Vallerand, Robert J., and Celine M. Blanchard. 2000. "The Study of Emotion in Sport and Exercise: Historical, Definitional, and Conceptual Perspectives." In *Emotions in Sport*, edited by Yuri L. Hanin, 3–37. Champaign, IL: Human Kinetics.

- Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge MA, MIT Press.
- Vickers, Joan N. 2007. *Perception, Cognition, and Decision Training: The Quiet Eye in Action*. Champaign, IL: Human Kinetics.
- Vine, Samuel J., Paul Freeman, Lee J. Moore, Roy Chandra-Ramanan, and Mark R. Wilson. 2013. "Evaluating Stress as a Challenge Is Associated with Superior Attentional Control and Motor Skill Performance: Testing the Predictions of the Biopsychosocial Model of Challenge and Threat." *Journal of Experimental Psychology: Applied* 19 (3): 185–194.
- Vine, Samuel James, Lee J. Moore, Andrew Cooke, Christopher Ring, Mark Robert Wilson, and Others. 2013. "Quiet Eye Training: A Means to Implicit Motor Learning." *International Journal of Sport Psychology* 44 (4): 367–386.
- Wallace, Harry M., and Roy F. Baumeister. 2002. "The Performance of Narcissists Rises and Falls with Perceived Opportunity for Glory." *Journal of Personality and Social Psychology* 82 (5): 819–834.
- Wang, Jin, Daryl Marchant, and Tony Morris. 2004. "Coping Style and Susceptibility to Choking." *Journal of Sport Behavior* 27 (1): 75–92.
- Wang, Jin, Daryl Marchant, Tony Morris, and P. Gibbs. 2004. "Self-Consciousness and Trait Anxiety as Predictors of Choking in Sport." *Journal of Science and Medicine in Sport / Sports Medicine Australia* 7 (2): 174–185.
- Will, George F. 1990. *Men at Work*. New York: Macmillan.
- Williams, Jean M., and Vikki Krane. 1992. "Coping Styles and Self-Reported Measures of State Anxiety and Self-Confidence." *Journal of Applied Sport Psychology* 4 (2): 134–143.
- Williams, Jean M., and Thad R. Leffingwell. 2002. "Cognitive Strategies in Sport and Exercise Psychology." In *Exploring Sport and Exercise Psychology*, 2nd ed., edited by Judy L. Van Raalte and Britton W. Brewer, 75–98. Washington, DC: American Psychological Association.
- Williams, Matthew, Joanne Hudson, and Romy J. Lawson. 1999. "Self-Presentation in Sport: Initial Development of a Scale for Measuring Athletes' Competitive Self-Presentation Concerns." *Social Behavior and Personality: An International Journal* 27 (5): 487–502.
- Williams, Sarah E., Jennifer Cumming, and George M. Balanos. 2010. "The Use of Imagery to Manipulate Challenge and Threat Appraisal States in Athletes." *Journal of Sport and Exercise Psychology* 32 (3): 339–358.
- Wilson, Philip, and Robert C. Eklund. 1998. "The Relationship between Competitive Anxiety and Self-Presentational Concerns." *Journal of Sport and Exercise Psychology* 20 (1): 81–97.
- Wine, Jeri D. 1971. "Test Anxiety and Direction of Attention." *Psychological Bulletin* 76 (2): 92–104. <http://dx.doi.org/10.1037/h0031332>.
- Witt, Jessica K., Susan C. South, and Mila Sugovic. 2014. "A Perceiver's Own Abilities Influence Perception, Even When Observing Others." *Psychonomic Bulletin and Review* 21 (2): 384–389.

- Woodman, Tim, Paul A. Davis, Lew Hardy, Nichola Callow, Ian Glasscock, and Jason Yuill-Proctor. 2009. "Emotions and Sport Performance: An Exploration of Happiness, Hope, and Anger." *Journal of Sport and Exercise Psychology* 31 (2): 169–188.
- Wrosch, Carsten, Michael F. Scheier, Gregory E. Miller, Richard Schulz, and Charles S. Carver. 2003. "Adaptive Self-Regulation of Unattainable Goals: Goal Disengagement, Goal Reengagement, and Subjective Well-Being." *Personality and Social Psychology Bulletin* 29 (12): 1494–1508.
- Wulf, Gabriele. 2013. "Attentional Focus and Motor Learning: A Review of 15 Years." *International Review of Sport and Exercise Psychology* 6 (1): 77–104.
- Wulf, Gabriele. 2015. "An External Focus of Attention Is a Condition Sine qua Non for Athletes: A Response to Carson, Collins, and Toner (2015)." *Journal of Sports Sciences* 34 (13): 1293–1295.
- Wulf, Gabriele. 2016. "Why Did Tiger Woods Shoot 82? A Commentary on Toner and Moran (2015)." *Psychology of Sport and Exercise* 22 (January): 337–338.
- Wulf, Gabriele, B. Lauterbach, and T. Toole. 1999. "The Learning Advantages of an External Focus of Attention in Golf." *Research Quarterly for Exercise and Sport* 70 (2): 120–126.
- Wulf, Gabriele, Nathan McConnel, Matthias Gärtner, and Andreas Schwarz. 2002. "Enhancing the Learning of Sport Skills through External-Focus Feedback." *Journal of Motor Behavior* 34 (2): 171–182.
- Wulf, Gabriele, N. H. McNevin, T. Fuchs, F. Ritter, and T. Toole. 2000. "Attentional Focus in Complex Skill Learning." *Research Quarterly for Exercise and Sport* 71 (3): 229–239.
- Wulf, Gabriele, and W. Prinz. 2001. "Directing Attention to Movement Effects Enhances Learning: A Review." *Psychonomic Bulletin and Review* 8 (4): 648–660.
- Wulf, Gabriele, C. Shea, and J. H. Park. 2001. "Attention and Motor Performance: Preferences for and Advantages of an External Focus." *Research Quarterly for Exercise and Sport* 72 (4): 335–344.
- Wulf, Gabriele, and Jiang Su. 2007. "An External Focus of Attention Enhances Golf Shot Accuracy in Beginners and Experts." *Research Quarterly for Exercise and Sport* 78 (4): 384–389.

Support Publication #7 [SP7]

Cappuccio M.L. (2023),
“Dreyfus is right: knowledge-that limits your
skill”, *Synthese* 202, 85, Springer Nature.
<https://doi.org/10.1007/s11229-023-04248-6>

Reproduced with permission from Springer
Nature. All rights reserved.



Dreyfus is right: knowledge-that limits your skill

Massimiliano L. Cappuccio¹ 

Received: 5 July 2021 / Accepted: 29 June 2023

© The Author(s) 2023

Abstract

Skilful expertise is grounded in practical, performative knowledge-how, not in detached, spectatorial knowledge-that, and knowledge-how is embodied by habitual dispositions, not representation of facts and rules. Consequently, as action control is a key requirement for the intelligent selection, initiation, and regulation of skilful performance, *habitual* action control, i.e. the kind of action control based on habitual dispositions, is the true hallmark of skill and the only veridical criterion to evaluate expertise. Not only does this imply that knowledge-that does not make your actions more skilful, but it also implies that it makes them less skilful. This thesis, that I call Radical Habitualism, finds a precursor in Hubert Dreyfus. His approach is considered extreme by most philosophers of skill & expertise: an agent –says Dreyfus– *does not perform like an expert* when they lack the embodied dispositions necessary to control their action habitually or when they stop relying on such dispositions to control their actions. Thus, one cannot perform skilfully if their actions are guided by representations (isomorphic schemas, explicit rules, and contentful instructions), as the know-that that they convey disrupts or diminishes the agent’s habitual engagement with the task at hand. In defence of Radical Habitualism, I will argue that only the contentless know-how embedded in habitual dispositions fulfils (i) the genetic, (ii) the normative, and (iii) the epistemic requirements of skilful performance. I will examine the phenomenological premises supporting Dreyfus’ approach, clarify their significance for a satisfactory normative and explanatory account of skilful expertise,

I am extremely thankful to the handling editors and three anonymous reviewers for their patient reading and detailed comments and to Wayne Christensen, Shaun Gallagher, Dan Hutto, Jesús Ilundáin-Agurruza, John Sutton for some conversations that inspired this project. The earlier versions of this paper have benefitted from the helpful feedback received from the investigators of the Minds in Skilled Performance ARC-funded research project, the participants of two conferences held on 15–16 February 2022 and 12–13 December 2022 at the University of Wollongong as part of that project, and the attendees of the webinar organized by Edoardo Fugali at University of Messina (26 January 2023). I am grateful to Shaun Gallagher for the Norwegian conference’s anecdote and his precious help with the references.

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

¹ Trusted Autonomy Research Group, School of Engineering and Information Technology, University of New South Wales at Australian Defence Force Academy, Northcott Drive, Room 201, Building 15, Campbell, Canberra, ACT 2612, Australia

and rebut the most common objections raised by both intellectualists and conciliatory habitualists, concerning hybrid actions guided by a mix of habitual and representational forms of control. In revisiting Dreyfus anti-representationalist approach, I will particularly focus on its epistemological implications, de-emphasizing other considerations related to conscious awareness.

Keywords Skilful expertise · Habitual disposition · Intellectualism about know-how · Radical habitualism · Enactive and embodied cognition

[...] Rules are like training wheels. We may need such aids when learning to ride a bicycle. But we must eventually set them aside if we are to become skilled cyclists. to assume that the rules we once consciously followed become unconscious is like assuming that, When we finally learn to ride a bike The training wheels that were required for us to be able to ride in the first place must have become invisible.

Hubert Dreyfus, *Overcoming the Myth of the Mental*, 2005

To see victory only when it is within the ken of the common herd is not the acme of military expertise. Neither is it the acme of military expertise if you fight and conquer and the whole world congratulates with you [...]. What the ancients called a skilful warrior is one who not only wins but excels in winning without effort.

Sun Tzu, *The Art of War*, Chapter 4

1 Introduction

1.1 A challenging legacy

In this paper, I will defend “Radical Habitualism”, a rather strong anti-intellectualist thesis: skilful expertise is grounded in practical, performative knowledge-how, not in detached, spectatorial knowledge-that, and knowledge-how is embodied by habitual dispositions, not contentful representation of facts and rules. Consequently, as action control is a key requirement for the intelligent selection, initiation, and regulation of skilful performance, *habitual* action control, i.e. the kind of action control based on habitual dispositions, is the true hallmark of skill and the only veridical criterion to evaluate expertise. Not only does this imply that contentful (e.g., propositional) knowledge does not make your actions more skilful, but it also implies that it makes them less skilful. In defence of this thesis, I will argue that only the contentless know-how embedded in habitual dispositions fulfils (i) the genetic, (ii) the normative, and (iii) the epistemic requirements of skilful performance.

Defending Radical Habitualism means clarifying and expanding some views attributed to Hubert L. Dreyfus (1929–2017), who can be considered its precursor and inspirator. The legacy that he and his key co-author (and brother) Stuart E. Dreyfus (1931–) have left to the field of philosophy of skill & expertise is as influential as it is divisive (Wrathall, 2014). Revisiting this legacy today –after amending it from certain inadequacies recurrently underscored by the philosophical literature– is paramount, inter alia, to better appreciate the theoretical solutions that it has prefigured, like those proficuously advanced during the last decade by *enactivists* (Varela et al., 1991; Di

Paolo, Buhrmann & Barandiaran, 2017; Gallagher, 2017; Hutto & Myin, 2013, 2017; Hutto & Satne, 2015; Noë, 2005; Thompson, 2014), *neo-pragmatists* (Gallagher, 2014; Caruana & Testa, 2020; Heras-Escribano, 2021; Miyahara, Ransom & Gallagher 2020; Miyahara & Robertson, 2021; Legg & Reynolds, 2022; Paolucci, 2021), and *affordance theorists* (Araújo et al., 2006; Bruineberg et al., 2019; Magri, 2019; Rietveld & Kiverstein, 2014; Rucińska, 2021).

Dreyfus (2002a, 2002b) vigorously argued that skilful expertise, the ability to perform competently in a specialised domain of practical activity (like sport tasks, trade jobs, and arts & crafts) relies on the immediate responses produced by embodied habitual dispositions (i.e., “pre-reflective” knowledge-how, or “know-how”), not the detached appreciation of facts and explicit rules (reflective knowledge-that, or “know-that”). Unsurprisingly, his approach attracted the criticisms of the philosophical views committed to the cognitivist idea that the control of skilful actions involves the manipulation of *semantic content* (typically construed as propositional and/or conceptual –albeit “practical” in nature and “motoric” in format– representation, e.g. Brozzo, 2017; Ferretti & Zipoli Caiani, 2021; Springle & Humphreys, 2021). These views notably include: *instructionist theories of action*, which claim that skilful performance is mediated by stored representations of action, such as conceptual and rule-based models (Fodor 1968, ; McDowell, 1981, 2007); *intellectualism about skill*, which emphasizes the role of reflective cognitive processes (such as attentive self-monitoring and explicit inference) in expert control (Fridland, 2013, 2014; Montero, 2010; Toner et al., 2014, 2015); and *intellectualism about know-how* (Pavese, 2019; Stanley, 2011; Stanley & Krakauer, 2013; Stanley & Williamson, 2017), which argues that all practical competences (know-how) are ultimately reducible to propositional knowledge (know-that).¹

Perhaps more surprisingly, Dreyfus’ approach is regularly antagonized by several theorists of embodiment, who –despite their post-cognitivist and non-intellectualist posture– have distanced themselves from Dreyfus’ habitualism (Sutton, 2007; Sutton et al., 2011; Gallagher, 2016, 2018; Cappuccio & Wheeler, 2010a, 2010b; Cappuccio, Miyahara & Ilundáin-Agurruza 2020). Why is it the case that Dreyfus’ perspective appears alarmingly radical not only to his most obvious opponents but also to his nominal allies? Neither did Dreyfus simply claim that *action automatization* is one of the distinctive marks of skilful performance and a prerogative of expert agents (Dreyfus & Dreyfus, 1980, 1988); nor that experts perform better when their movements are automatized, in comparison to when movements are chosen deliberately and executed

¹ The position normally referred as “intellectualism about know-how” famously dates back to Ryle (1949/2009), while the notion of “intellectualism [about] sporting skills” was first introduced by Papineau (2015) to map the opposition of “habitualists” and “intellectualists” in philosophy of skill & expertise: “on the habitualist side are Dreyfus (2005, 2007a, b), Beilock (2007, 2010) and (Beilock and Carr 2001, 2004; Beilock et al., 2002a, 2002b); intellectualism is defended in various ways by McDowell (2007), Stanley (2011), Sutton (2007), Sutton et al. (2011), Montero (2010) and Fridland (2013)” (Papineau 2015, p. 295). I find it useful to distinguish between these two philosophical varieties of intellectualism because some authors (e.g., Ellen Fridland) defend arguments in favour of the latter but refrain from endorsing the former, while other authors (e.g., Jason Stanley) wholeheartedly advocate the former without explicitly committing to the latter. Note that, unlike me, some authors, for example Springle and Humphreys (2021) and Robertson & Hutto (forthcoming), use “intellectualism about skills” simply to indicate the specific application of “intellectualism about know-how” to the domain of skill & expertise.

reflectively, in accord with explicit judgement (Dreyfus, 2002b). In fact, these two claims seem at least reasonable, if not uncontroversial, to performance scientists: most kinesiologists agree that a skill develops when the automatization of motor behaviours ensues (Fitts & Posner, 1967); and several performance psychologists maintain that self-focused and/or declarative processes can be detrimental to optimal skilful performances (Baumeister, 1984; Beilock & Carr, 2001; Masters, 1992; Wulf et al., 2002).

Dreyfus' signature view on skill and expertise is more demanding and inflammatory: an agent *cannot act like an expert* (thus, their performance *is not truly skilful*) when they lack the embodied dispositions necessary to control action habitually (e.g., due to insufficient practice), when they don't rely on such dispositions to control their actions (e.g., when reflective thoughts conveyed by instructions and rules impose self-attention, interfering with habitual control), or when the dispositions they must rely on are temporarily disrupted or inaccessible (e.g., due to anxiety undermining the agent's confidence in their own habits).

1.2 A remarkable anecdote

This view is illustrated by an anecdote that sees Dreyfus arguing, at a conference,² that downhill skiers are pre-reflectively absorbed in their performance when they skilfully negotiate a ski track during a competition (Gallagher, 2007, 2009, 2016). When faced with the evidence that professional skiers obviously *can*, typically *do* and, moreover, often *must* carefully scrutinize and deliberately adjust their posture, predict the condition of the snow, and make on-the-fly strategic decisions to maximize their skying effectiveness, Dreyfus imperturbably answered that, as soon as a skier starts to explicitly monitor the details of the situation or modify their own behaviour in accord with verbal instructions and maxims, they stop performing *as experts*, in that their reflective stance interrupts their pre-reflective absorption into the task at hand. Most authors vehemently disagree with this conclusion (Breivik, 2007; Fridland, 2014; Gallagher, 2016, 2017; Gottlieb, 2015; Montero, 2010, 2016; Sutton et al., 2011, to mention but a few).

I shall contend that Dreyfus had excellent reasons to reach this conclusion, which becomes compelling once his motives are elucidated. I have elsewhere argued that Dreyfus' approach suffers from the tendency (that, in fairness, he shares with some of his philosophical adversaries) to conflate too many and too general conceptual dichotomies: mindful/mindless, conscious/automatic, embodied/cognitive, etc. (Capuccio, 2017). Once this terminological inadequacy is addressed and overcome, we realize that the kernel of Dreyfus' controversial statement on expertise resonates with the habitualist perspective on skill that enactivism, neo-pragmatism, and affordance theory have been recently advocating and that Dreyfus' phenomenology contributed to inspire. Conversely, as I shall argue in this paper, a correct appreciation of Dreyfus' thesis urges these embodied approaches to coherently accept the most far-reaching

² The conference is "Phenomenological approaches to Moral Philosophy and Education" hosted by the Norwegian School of Sports Sciences, University of Oslo, on 6–8 June 2006 (Shaun Gallagher, personal communication).

ramifications of their own habitualist wager, finally embracing the markedly anti-intellectualist approach to skill & expertise that I call “Radical Habitualism”. My inquiry primarily aims to provide a sound and viable formulation of such an approach, not an apologia of Dreyfus, which is why I will emphasize the spirit more than the letter of his work, prioritizing hermeneutical efficacy over philological rigor.

My point is articulated through the six sections and fifty subsections that follow this Introduction. Section 2 analytically presents Radical Habitualism, illustrating its distinctive theses (Sect. 2.1), its characteristic assumptions about performance (Sect. 2.2) and control (Sect. 2.3), and the three pillars that sustain its theoretical foundation (Sect. 2.4). Section 3 presents three arguments supporting Dreyfus’ view: the first argument builds on the structure of skill development (Sect. 3.1); the second on the phenomenology of background pre-comprehension and pre-reflective control (Sect. 3.2); the third on the epistemic qualities of expert know-how (Sect. 3.3). Sect. 4 clarifies Dreyfus’ view about hybrid scenarios, i.e. performances involving a mix of automatized and reflective control, and identifies three contentious hybrid cases that are erroneously believed to disprove Radical Habitualism (Sect. 4.1). To dispel this common misunderstanding, I prove that Radical Habitualism can perfectly account for such cases because it relies on a robust “Skill-No-Skill” approach to hybrid scenarios, which represents a preferable alternative to the “Knowledge-No-Knowledge” approach typically endorsed by intellectualists (Sect. 4.2). Section 5 examines three objections that challenge the Skill-No-Skill approach: the “Scope” objection focusses on the apparent disembodiment of intellectual activities (Sect. 5.2); the “Efficacy” objection on the competitive advantages allegedly offered by contentful knowledge (Sect. 5.3); and the “Modality” objection on certain content-manipulating practices that seem linked to skilful execution (Sect. 5.4). I rebut them systematically: against the first objection, I argue that all forms of expertise originate from a constitutive agent-environment interactive coupling; against the second objection, I develop a three-step strategy to demonstrate, in the first place, that contentful knowledge is only rarely beneficial to performance (Sect. 5.3.1), in the second place, that contentful representation is causally impotent and contextually irrelevant when it comes to benefitting performance (Sect. 5.3.2) and, in the third place, that performance outcomes are not sufficient evidence of expertise (Sect. 5.3.3); against the third objection, I argue that verbal reporting (Sect. 5.4.1), creative improvisation (Sect. 5.4.2), and strategic decision (Sect. 5.4.3) can effectively contribute to skilful execution only in the form of contentless habits. Section 6 illustrates these points with the example of barbecue skills and meat thermometers. The conclusive (Sect. 7) briefly summarizes how Radical Habitualism unveils aspects of skilful performance to which other theories are largely blind.

2 Defining Radical Habitualism

2.1 A controversial thesis

Dreyfus’ Controversial Thesis [DCT] states (more or less).

DCT: The more a performance is controlled reflectively the less it is skilful, and vice versa.

To clarify the meaning of DCT we will assume that “performing like an expert” is synonymous of “performing skilfully”. From Dreyfus’ perspective, what distinguishes pre-reflective and reflective control is that the former is grounded in *embodied habitual dispositions* (thus incorporates know-how) while the latter is grounded in *contentful knowledge* of facts and rules (thus conveys know-that), the former being irreducible to the latter. The latter includes the rules of thumb, explicit verbal instructions as well as the non-verbal representations (such as tactical schemas and diagrams, step-by-step protocols and standardised recipes for action) that novices (such as junior athletes) are often asked to rehearse for acquiring their basic abilities, or for maximizing the outcomes of their performance, or for ensuring that the form of their execution complies with some conventional standards (including the sport discipline’s distinctive praxical and ethical norms, see Sánchez García, Cappuccio, and Ilundáin-Agurruza, forthcoming). Conversely, performers increasingly rely on their embodied habitual dispositions as they progress in the acquisition of skill and expertise. Dewey (1983, p. 32) famously stated that habits “predispose” the performer to intelligently perceive, interpret, decide, and act in familiar (but not stereotyped) “ways or modes” when facing familiar situations or tasks. Additionally, habits make the performer ready to cope with unfamiliar scenarios, equipping them with a “special sensitiveness or accessibility to certain classes of stimuli, standing predilections and aversions” (*ibidem*).

That said, to capture the specificity of Dreyfus’ position, we still need to differentiate Radical Habitualism from the generic habitualist view on skilful performance (Cappuccio, Miyahara, K., Ilundáin-Agurruza, J. 2020). All habitualist views, including DCT, share the same Baseline Habitualist Assumption [BHA].

BHA: embodied habitual dispositions are a necessary component of expertise and action control is skilful only if embodied habitual dispositions contribute to it.

BHA specifies that habit is a necessary non-sufficient condition for skilful control. However, BHA may or may not be accompanied by the following Disjunctive Clause [DisC].

DisC: the more the control of an action (or action component) is grounded in embodied habitual dispositions the less it is grounded in contentful knowledge, and vice versa.

BHA motivates two different strands of habitualism, distinguishable by whether or not they endorse DisC.

Conciliatory Habitualism ($BHA \wedge \neg DisC$): embodied habitual dispositions are a *necessary constituent*, but not necessarily the main or the sole one, of skilful expertise. Both embodied habitual dispositions and contentful knowledge can positively contribute to skilful control.

Radical Habitualism ($BHA \wedge DisC$): embodied habitual dispositions necessarily are the sole constituent of skilful expertise. While embodied habitual dispositions can positively contribute to skilful control, contentful knowledge cannot.

Conciliatory Habitualism is utterly popular nowadays for its ecumenic aspiration to integrate representational (contentful) and habitualist (embodied) explanations. Versions of it are endorsed by many supporters of the embodied/enactive cognition theory and some amongst the most liberal intellectualists. Regrettably, this position is vitiated by cognitivist and rationalist premises no less than the more intransigent intellectualist position from which it ineffectively tries to distance itself (Cappuccio, Miyahara & Ilundáin-Agurruza 2020).

Unlike Conciliatory Habitualism, Radical Habitualism rigorously captures the notion that a skill is an embodied ability formed through practice and based on know-how (i.e., *how* to obtain X in a Y way), not erudition formed through detached understanding and based on know-that (i.e., *that* Y is a way to obtain X). Some performances may involve elements of contemplative and non-situated knowledge about the relevant actions and their execution but acting skilfully in real life is more of an art than a science: like a performing art, skilful expertise is the ability to execute *a performance* endowed with certain positive qualities.

2.2 Defining skilful performance

A performance is a particular kind of *pragmatic action*, that is an instance of *purposive* (Butterfill & Sinigaglia, 2014; Kruglanski & Szumowska, 2020), *agentive* (Brownstein, 2014a, 2014b; Swann, 2016), and *intentional* (Fusche Moe, 2016; Cappuccio, Miyahara & Ilundáin-Agurruza 2020) engagement that (unlike, for example, epistemic actions) aims to fulfill concrete, tangible expectations by bringing about valuable outcomes (e.g., complete a task, produce some transformative effects, solicit certain reactions in the bystanders, or preserve an existing condition). Both *performances* and *instrumental actions* are subtypes of pragmatic actions. Crucially, skill pertains to the former, not the latter. That is because the value of instrumental actions prototypically consists in bringing about a desired outcome regardless of the means employed to do it. For example, compiling forms: both handwriting and typewriting can be skilful activities in their own respect, but form compilation, whether done using handwriting or typewriting, is an instrumental action, not a skilful performance, precisely because the details of the writing practice adopted by the compiler are entirely irrelevant to the goal of form filling.

Conversely, the fine-grained details of execution (including its sensorimotor and morphological contingencies) are constitutive of performances and contribute to determine their value (Beilock, 2015; Gray, 2014, 2021). Whether on the sport field, on the stage, or in the workplace, a performance is neither merely instrumental nor value-neutral because its execution can and must be evaluated qualitatively or quantitatively in comparison to executions of the same kind. All performances are not equally skilful and, in some competitive domains, such as agonistic sport, stringently regulated measuring practices are specifically adopted to normalise their comparison. Assuming sufficiently similar background conditions, our comparative assessment of the value of a performance is based on two aspects:

- (1) *the outcome* X resulting from the form Y of performance execution (e.g., a theatrical role interpreted by an actor, a handmade vase, a perfectly done steak);

- (2) *the form of execution* Y aiming at the production of performance outcome X (e.g., acting with a particular theatrical method, modelling the vase with a certain crafting technique, grilling a steak using a particular piece of equipment or procedure).

In most cases, outcome and form of execution are entirely distinguishable (e.g., the fact that a football team scored a certain number of goals during a match versus the quality of the actions performed to score those goals). In some cases (dance is a paramount example), the outcome produced by the performance coincides with the behaviours executed to produce it, hence the performance's desired outcome is entirely exhausted by the *execution of a distinctive form of behaviour* (often, the expected outcome is nothing else than a certain measurable quality of the execution, e.g. the time necessary to run a certain distance, the height of a jump, etc.).

Either way, the modality or *form* of execution, more than its outcome considered in isolation, reveals the kind of knowledge (how/that) and the modality of control (habitual/reflective) that underpin a performance (Foerster et al., 2015). Thus, that an agent “does not perform like an expert”, in Dreyfus' sense, means –primarily or essentially– that their performance is *not executed in a habitual modality* and –only secondarily or incidentally– that their performance *did not produce the outcomes normally expected from an expert*.

2.3 The qualities of habitual control

To distinguish between more and less skilful ways of performing the same action we need to look at how the execution of that action is controlled by the agent. As emphasized by the recent philosophical literature, *control* is a quintessential feature of skilful execution. Default abilities (including innate talents and knacks) are profitably exploited by skilled performers and acting skilfully means, among other things, making the most of such abilities, but no ability without control could be considered evidence of skilful expertise (Bermúdez, 2017; Carr, 1979; Beilock & Carr, 2004; Christensen & Bicknell, 2022; Fridland, 2017; Toner Montero & Moran 2014). Perfectly controlled performances display well-coordinated, purposeful configurations of movements initiated and executed with consistency, timeliness, sensitivity, and responsiveness to all the environmental solicitations that are relevant to the agent's purpose (Fridland, 2020; Mylopoulos & Pacherie, 2021).

We immediately recognize finely controlled performances because they are executed in a confident, spontaneous, richly flexible, and intelligently adaptive manner. Control is not simply the agent's power to reliably produce certain outcomes: expert performance is also characterised by certain qualities of the execution (such as accuracy, precision, velocity, efficacy, and parsimony), which are neither secondary nor merely instrumental to produce the desired outcomes but constitutive of the performative process (Fridland, 2014, 2017; Toner et al., 2014).

While everybody acknowledges the importance of control, its nature is disputed (Cappuccio & Ilundáin-Agurruza, 2020): intellectualist and cognitivist philosophers see habit as competing with control (Fridland, 2014; Wu, 2013) or consider these two

concepts orthogonal (Toner, Montero & Moran 2021; Toner & Moran, 2021)³; on the contrary, neo-pragmatists, enactivists and dynamicists recognize that *control can be habitual* because the degree and the extension of action automatization are measures of how the agent is familiar to their complex world-environment (Cappuccio & Ilundáin-Agurruza, 2020).

This kind of performative familiarity consolidates through recurrent interactions, like those based on deliberate practice, targeted exercise, and (to a lesser extent) direct observation (Araújo & Davids, 2011; Egbert & Barandiaran, 2014; Freeman, 2000). A history of rich and varied (and, for some activities, also lengthy and repeated) interactions produce a dynamical coupling between the agent and the context of their habitual activities, establishing a continuing relation of diachronic co-constitution (Kirchhoff & Kiverstein, 2020) and reciprocal co-belonging (Barandiaran, 2017) that can be characterized in terms of “metastable attunement” to multiple concurrent systems of nested affordances (Bruineberg et al., 2021). This coupling, or attunement, combined with a reward system and appropriate hedonic stimulations, solicits the formation of articulated patterns of purpose and motivated interest that bring the agent to develop fundamental capabilities like cross-context saliency detection, engaged attention, circumsperspective interpretation, and projective anticipation (Di Paolo & Iizuka, 2008; Di Paolo, Buhrmann & Barandiaran, 2017), which are constitutive of skilful control.

Ultimately, action automatization coincides with the formation of a robust agent-environment coupling that derives from the history of their adaptive interactions. Carr (2015) gestures at an analogous form of coupling when referring to “action-situation combinations” that reflect the sensitivity to contextual conditions acquired by an expert through automatization (the notion of coupling is further explored in Section 3.2.3 and 3.3.2; the notion of automatization in Section 3.1.1, 3.1.2, 3.1.3). This account captures at once the dispositional and the relational/interactive character of skills, as auspicated by Chemero (2009) and Magri (2019). The agent-environment coupling embodied by the agent’s habitual dispositions is both constitutive and irreplaceable for controlling a well-executed performance: control is skilful to the extent it reflects the rich executive know-how acquired through practice, exercise, and direct observation. The best way to construe this know-how is not as a power, entitlement, or function but as *the qualitative profile of the performance’s mode of execution*, which in turn reflects the holistic form of the performer’s attunement to the performance’s environment (Dreyfus, 1996, 2000, 2002a). By explicating the theoretical pillars of Radical Habitualism, I will progressively explain why such an attunement must necessarily be based on habit and

³ Fridland (2014, p. 2732) argues that Dreyfus lacks “an adequate account of control” and that automatization replaces control with a “passive, mechanistic” process that is incompatible with intelligence and person-level intention. My impression is that this kind of objection to Dreyfus overlooks how, far from being merely reactive, automatized actions predictively anticipate the ever-evolving context of action (Araújo et al., 2006; Hipólito et al., 2021; Robertson & Kirchhoff 2020): automatized control is, in fact, exactly the contrary of a blindly mechanical form of control because it responds adaptively to complex circumstances (Cappuccio Miyahara & Ilundain Agurruza, 2020). The idea that automatized action is blind and mechanical seems originating from the common misconception (of which falls victim, among others, Fridland 2014, p. 2720, note 1) that sees intentionality as necessarily contentful and reflective, ignoring that motor intentionality (while pre-reflective and contentless in nature) has distinctively agential and personal features (that I summarize in Sect. 3.1 and Sect. 3.2.3). I agree with Christensen that a more “sophisticated” appreciation of automaticity can be found in Fridland (2017), where it is argued that “automatic processes show complex sensitivities to cognitive states.”

thus why (in accord with DisC) only the mode of execution based on habitual control is truly skilful.

2.4 Radical Habitualism's three pillars

This analysis should suffice to a preliminary clarification of DCT. However, to characterise Dreyfus' view on skill as Radical Habitualism we need to recognize the Three Pillars that underpins his approach. The Three Pillars are introduced in the following sub-sections (Sect. 2.4.1, 2.4.2, 2.4.3) and will be further unpacked throughout the rest of the paper, also via the discussion of their most common objections (Sect. 5).

2.4.1 The first pillar of Radical Habitualism: skill and habit are underpinned by the same control system because they share a communal genesis and develop together

The terms “skill” and “habit” have different connotations, but both denote dispositional complexes formed through a rich history of direct experiential engagement (Moya, 2014), which is why Dreyfus (1996, 2014) calls “skilful coping” or “absorbed coping” the modality of expert engagement characterised by habitual control. Of course, not all habitual behaviours are skilful or intelligent: some of them are fixed, compulsive responses resulting from conditioning⁴; yet all skills originate from the acquisition of habitual control, because rich responsiveness to the context and prompt readiness to the circumstances require the kind of practical familiarity that can only be learned through the purposeful accumulation, consolidation, and refinement of direct performative experience.

Pace Pacherie & Mylopoulos (2021), training does not conceptually oppose skill to habit because, at bottom, *skills are just well-trained habits*. Habits are well-trained when they are cultivated *viapractice for a purpose*. The training of habits paradigmatically involves practices that may or may not be deliberately designed but that necessarily include, inter alia: *opportunistically diversified exercises* (aiming to expand one's repertoire of habits and avoid the plateau of skill described by Montero, 2016, p. 84); the careful observation and *imitation of models* (aiming to acquire the habits of more experienced performers); and even what may superficially look like *rote repetition* of standardized sets of movements.⁵ The purpose of training does not have to be always entirely evident to the trainee (consider the Miyagi sensei-like training routines illustrated by the Karate Kid movies and examined in Cappuccio, Miyahara and Ilundáin-Agurruza 2020) but it normally aims to the deliberate improvement of action control, inter alia. This topic is further explored in Sect. 4.1.1 and 5.2.

⁴ Ramírez-Vizcaya & Froese (2019) correctly point out that even negative habits traditionally considered merely compulsive, like addictions, are in fact governed by complex attractors that, while suboptimal, can nonetheless solicit highly creative and skilful (although not necessarily wise) behaviours (e.g., drug-seeking and drug-taking behaviours).

⁵ In fact, from a dynamicist perspective, even *prima facie* repetitive sequences of identical actions are not a mechanical re-enactment of blindly stereotyped movements but rather the target of a continuous effort of adaptation and adjustment to an ever-evolving horizon of complex micro-sensorimotor contingencies (Button et al., 2020).

2.4.2 The second pillar of Radical Habitualism: habitual control is grounded in pre-reflective intelligence

Dreyfus characterises skilful coping as embodied, dispositional and *pre-reflective* (or “absorbed”, see Dreyfus, 1996). According to Maurice Merleau-Ponty (1945/2012), pre-reflective intelligence does not necessarily prevent reflection but certainly constrains and demarcates it, as these two categories are in a background-foreground relationship: despite their reciprocity, the pre-reflective is more primordial than the reflective, because it operates before and below contentful cognition, as its transcendental precondition (Jackson, 2018; Moya, 2014; Reynolds, 2006). As such, the pre-reflective defines both the ontological boundary and the functional limit of the reflective (Dreyfus, 1996, 2000, 2008; Moe 2007; McManus, 2008; Wheeler, 2008). Dreyfus uses without apparent distinction two different characterisations of pre-reflective intelligence, in a way that is potentially ambiguous: first, as non-conscious, non-cognitive, non-agential, or even “mindless” (Eriksen, 2010; Zahavi, 2012); and, second, as non-representational, i.e. guided solely by the contentless solicitations produced by online experience.

The First Characterization of the pre-reflective is misleading insofar as it suggests that habitual control can be neither rational, self-aware, nor deliberate, while we know that –far from being blindly mechanical– habitual responses are *flexibly adaptive* (Davids et al., 2012; Di Paolo, Buhrmann, & Barandiaran, 2017), *predictive* (Christensen & Bicknell, 2019; Robertson & Kirchhoff 2020; Hipólito et al., 2021), and capable to *reliably detect and discriminate* subtle situational constraints or task-specific requirements (Hutto & Robertson, 2020; Miyahara & Robertson, 2021). These ecological features indicate that habitual dispositions can be *sensitive to context* as they are informed by *implicit reasons* (Kruglanski & Szumowska, 2020), *goals* (Martens, 2021), *preferences* (Beilock, 2008; Beilock & Holt, 2007) and *values* (Brownstein, 2014a, 2014b). Thus, I agree with other authors (Birch, Breivik & Fusché Moe 2019) that the First Characterisation tend to confound the correct appreciation of habitual control: on the one hand, characterising the pre-reflective stance of skilful performers as non-conscious ignores that competent habitual routines can be and often are initiated purposely, executed attentively, and integrated in broader decisional activities (Breivik, 2013; Brownstein, 2014a, 2014b; Gottlieb, 2015; Martens, 2021); on the other hand, reducing the scope of the pre-reflective to non-conscious movement fails to acknowledge that also patterns of strategic thinking and rational problem-solving, no less than overt muscular efforts, can be automatized, becoming cognitively effortless habitual activities (Van der Herik & Rietveld, 2021).

But Dreyfus is right about the Second Characterization: reflective control interferes with habitual control (Holt & Beilock, 2006; Jackson & Beilock, 2008) because (through the explicit representation of one’s own actions, or rules and instructional maxims, or objective facts) it evokes a self-focussed, analytic, and calculative stance. In perfectly familiar activities, habitual control does not involve representation because it is *immediately responsive*, *intuitive*, *non-analysable* and *non-vetoable* (Neemeh, 2022); also, it is selectively inattentive towards the details of execution (i.e., the attention sustaining skilful control is outcome-focused, as opposed to self or component-focused); finally, sense of agency, sense of volition, and sense of effort

happen to be specifically attenuated during episodes of flow performance, which are characterised by habitual control (Swann, 2016; Swann et al., 2016). However, as clarified in Cappuccio (2017), none of this implies that habitual control is “non-conscious” (no qualitative experience, no attentive awareness), “non-agential” (entirely devoid of sense of control, effort, commitment) or “mindless” (no cognitive elaboration, no decision).

For a correct appreciation of Radical Habitualism, it is imperative to sanitise the concept of “pre-reflective” (see Sect. 3.2.2). We must in fact urgently acknowledge that, during the last ten years at least, the disproportionate emphasis put by the critical literature on the First Characterization has regrettably derailed the philosophical debate on the role of habitual control, reducing it to an artificial, unproductive, and rather caricatural rivalry between “mindful” and “mindless” cognitions (Segundo-Ortín & Heras-Escribano, 2021), as if Dreyfus’ interest was really to ascertain whether performers act “consciously” or rather “as zombies”. The good news is that the First and the Second Characterization of the pre-reflective are entirely dissociable.

That is why, in this paper, I will methodologically ignore any portrayal of habitual control as *non-conscious*, *non-agential*, or *mindless* to focus on an aspect of Dreyfus’ phenomenology of skilful coping that I find much more important and interesting. What truly characterises the distinction between pre-reflective and reflective control is not the absence or presence of conscious awareness, attentive deliberation, or reasons to act, but *the agent’s reliance on contentful or rather contentless forms of intelligence, attention, and knowledge* (Hutto & Sánchez García, 2015; Hutto & Satne, 2015). My choice to selectively focus on the contentful-vs-contentless dichotomy (semiotic), definitively abandoning the conscious-vs-zombie dichotomy (metaphysical), is not arbitrary: first, as argued in Sect. 3.2.3, it best epitomises the Merleau-pontian account of motor intentionality that inspires Dreyfus (1996, 2000, 2002a, 2002b); second, as discussed in Sect. 5.3.1.2, it is well positioned to elucidate the complex connection between attentive self-monitoring and representational processes that Dreyfus examines phenomenologically with only scant references to the experimental literature (Dreyfus, 2005, 2007a, 2013); third, as illustrated in Sect. 5.3.2, it is the only dichotomy truly capable to challenge the representationalist dogma of cognitivism, indicating that an anti-representationalist alternative is both possible and necessary (Dreyfus & Dreyfus, 1999).⁶

2.4.3 The third pillar of Radical Habitualism: pre-reflective and reflective intelligence correlate to contentless know-how and contentful know-that, respectively

Assuming *ceteris paribus* conditions, pre-reflective control proves the expertise level of a performance, while reflective control is symptomatic of a lack of expertise. By

⁶ Another important terminological remark is necessary: occasionally, Dreyfus asserts that skilful coping involves content of a non-conceptual, non-propositional kind (e.g., Dreyfus 2002a) and, in other occasions, that “there is no place in the phenomenon of fully absorbed coping for intentional content mediating between mind and world” (Dreyfus 2013, p. 28). The contradiction is only apparent: motor intentionality is semantically contentless in the sense that it involves neither logical values (Dreyfus 2007a) nor content/attitude distinction (Kelly 2002); if it has a “content”, it is only in the phenomenological sense of hyletic contingencies, i.e. inherently qualitative, unorganized experiential material (Dreyfus & Dreyfus 2014). I believe this explains the lexical discrepancy highlighted by Thompson (2018).

capturing the specificity of DisC, the Third Pillar neatly distinguishes Radical Habitualism from the conciliatory versions of habitualism. A performance is unskilled in proportion of how it is controlled reflectively, i.e. to the extent that it is produced in absence of relevant habits or through the mediation of contentful representations (verbal commands, plans, models, rules, etc.) that explicitly instruct the agent, telling them if, when, and how certain component movements are to be executed (Dreyfus & Dreyfus, 1999). Such a reflectively controlled action is not skilful even when the outcome of said performance is unanimously considered satisfactory or successful and even in the case that the representations informing the agent's reflective control never interfered with their habitual dispositions, because in both cases the mode of execution was unfamiliar, i.e. not based on direct practice, which remains the exclusive source of expertise. Thus, the Third Pillar does not exclude that successful performances could at times rely on representational processes, but it clarifies that the varieties of performances correctly indicated as "skilful" or "expert" are never controlled by contentful representations (Dreyfus, 2002b).

Importantly, this thesis collides with "Mesh" (Christensen et al., 2016; Christensen & Sutton 2019; Christensen et al., 2019a, 2019b; Christensen & Bicknell, 2022) and similar integrative approaches to skill (e.g. Gallagher & Varga, 2020; Gallagher et al., 2022), which consider beneficial to performance a perfectly harmonized blend of pre-reflective and reflective cognition; also, Dreyfus' anti-representationalism seems *prima facie* at odds with the neo-pragmatist approaches to enactive cognition that construe semiosis as a representational process grounded in interpretative habits (Baggio, 2021; Legg, 2021). Since the early days of enactivism (Varela, Thompson, Rosch 1991), Dreyfus strongly criticized all the emerging embodied approaches to cognition, alleging their unadmitted commitment to Cartesian premises (Dreyfus, 1993, 2008).

Today, this ancient antagonism appears less interesting than the fact that Radical Habitualism shares a generic, yet very important, habitualist assumption (BHA) with enactivism and other varieties of the "4E approach" to cognition (Newen, De Bruin, Gallagher, 2018). In particular, in substantial accord with the First and the Second Pillar, all enactivists see "mental life as a habit ecology" and consider habitual control as "an agent's behaviour generating mechanism in coordination with its environment" (Barandiaran, 2017). Emerging from the sedimentation of the agent-environment coupling, habitual dispositions are the true source of the motoric and perceptual know-how, i.e. the very fabric of which embodied intelligence itself is made out. This perspective is truly relational and interactionist in the sense explored by Coeckelbergh (2015).

The strands of 4E cognition that seek a compromise with representationalism struggle to accept the Third Pillar. Radical Enactivism (which emphasizes the contentless nature of "basic" forms of cognition, see Hutto & Myin, 2013, 2017) might, on the other hand, provide a more conducive ground to defend an intransigent anti-representationalism like the one proposed by Radical Habitualism. If it will prove itself radical enough to incorporate the Third Pillar, Radical Enactivism could offer a new home to Dreyfus' idea that pre-reflective know-how is informed by a primordial kind of intentionality, a form of cognition that is both ontogenetically and phylogenetically more fundamental than, functionally irreducible to, and logically incompatible

with representational cognition (i.e., the kind of intelligence based on contentful judgment and explicit knowledge of facts, rules, principles, and instructions, see Hutto & Satne, 2015).

As the Third Pillar is probably the most polarizing aspect of Radical Habitualism, we will need to clarify its reach in Sect. 4.1 and exhaustively refute the objections that threaten it in Sect. 5.2, 5.3, and 5.4.

3 Three arguments supporting Dreyfus' Controversial Thesis

In this section, I will present three arguments that support Radical Habitualism by corroborating the DisC. The first two are directly taken from Dreyfus, while the third argument is, as far as I know, original. The first argument (Sect. 3.1) derives from the analysis of the general mechanics of expertise development according to the influential "Adult Skill Acquisition" model proposed by Dreyfus and Dreyfus (1980, 1988), asserting that a learner's skill level progresses as their contentful know-that is replaced by dispositional know-how. The second argument (Sect. 3.2) builds on the phenomenology of skilful action and the reciprocal incompatibility of the different normative conditions under which the success of reflective and unreflective action is defined. The third argument (Sect. 3.3) stems from the acknowledgment that reflective and unreflective action rely on irreducibly different forms of knowledge, and that only the kind of expert know-how embedded in embodied habitual dispositions has the qualities required to support skilful performance.

3.1 First argument: the structure of skill development

The first argument builds on the empirical analysis of the universal form of skill acquisition, a gradual process that follows a predictable pattern. Bryan and Harter (1899) noted that the development of expertise progresses at once with "the acquisition of a hierarchy of habits." Since James (1890), at least, scientific psychologists have used the technical term "automatization" to indicate the *learning process* through which the components of performance progressively become *habitual* (Logan, 1988). According to Logan's prominent definition (Logan, 1985), the sub-personal processes underpinning the control of automatized (i.e., habitual) actions distinctively are: (1) comparatively effortless⁷; (2) autonomous; (3) partly unavailable to attentive awareness (i.e., eluding explicitly directed, analytic self-monitoring); and (4) consolidated through practice. Psychologists still debate whether any of these four conditions is either necessary, sufficient, or both for a full automatization. Importantly, within the lexicon normally used by contemporary performance psychologists, the term "automatized" (2) does not imply fixed repetitiveness or blind mechanicity (Carr, 2015),

⁷ As specified earlier, the complete absence of sense of effort is not a necessary pre-requisite of automatized action (i.e., controlled habitually). A total lack of experienced effort is rare, even in the extraordinary conditions of pre-reflective control referred as "flow" or "clutch" performance. However, progressive action automatization (coinciding with control habitualization) does correlate with a diminished or transformed sense of effort, which is experienced as more tolerable, or even enjoyable, when compared to unhabitual action.

which is why it is rather remarkable that some philosophers have started using it with this connotation (e.g., Montero, 2016; Wu, 2013)⁸: “automatized” simply indicates that the performer’s control ability irreversibly moves toward a fully pre-reflective modality while progressively moving away from the reflective modality.

But how can our actions become habitual if the trainee’s practice is based on the consumption of contentful indications like rules or maxims? When they begin to learn a skill, the typical learner *temporarily relies on* the explicit instructions and representational models provided by an expert trainer/teacher *but the learning process does not consist in acquiring or retaining such instructions and models in a contentful form*. The role of content is particularly evident during the early stages of skill learning, which heavily rely on such instructions and models. The more experienced performers may rely on explicit know-that to reach the higher levels of expertise too and the exact extent of their reliance on know-that depends significantly on the discipline’s specificity, the instructor’s teaching method, the learners’ preferences, etc.

The point to be made here is not that learners initially *need* contentful representation of rules and facts to become experts (in fact, in many cases, they don’t); the point is rather that, any time contentful know-that and reflection are leveraged to learn a skill, *the role played by content is merely instrumental* to reaching the next level of know-how, without being *constitutive* of it. This role, like *that of a scaffold*, is *temporary, functionally replaceable, and not essential*. The learners reach a more mature developmental stage the very moment they prove ready to discard such a scaffold, when its assistive function finally becomes redundant. Thus, their knowledge (which is contentful only at an early stage, if it is contentful at all) is transformed (i.e., it becomes contentless) by the very same process of skill development that transforms them into experts.

This point is perfectly captured by the influential model introduced by Dreyfus and Dreyfus (1980, 1988), recapitulated by Dreyfus (1997, 2002a), and recently expanded and refined by Rouse and Dreyfus (2021). According to this model, skill development advances linearly through six⁹ distinct stages that reflect the performer’s increasing expertise level: *novice* (a learner who rigidly applies narrow sets of rules to navigate context-neutral and goal-specific situations), *advanced beginner* (a learner who still relies on general instructional maxims and applies them flexibly to a variety of familiar practical situations characterised by similar aspects), *competent* (a performer who still needs to deliberately choose the practical perspective to be adopted for reaching the goal associated with a familiar situation), *proficient* (a performer who can intuitively adopt the most appropriate practical perspective on the situation but still needs to explicitly decide how to respond to direct solicitations), *expert* (a performer who can count on a large repertoire of familiar perspectives to fluidly switch from one goal-specific situation to another, intuitively responding to the direct solicitations

⁸ Montero (2016, p. 42): “I think this conception of automaticity suffices for my purposes and more or less captures Dreyfus’s notion of automaticity, the concept is understood in myriad ways in the philosophy and psychology literature. For example, Sutton et al. (2011) argue that automatic actions themselves are thoughtful. See also Wu (2013) and Fridland (2015).”.

⁹ Originally five, as the Mastery level does not appear in all the versions of the theory.

to act), and, finally, *master* (an expert capable to actively experiment with the available perspectives, exploring new action possibilities and creatively reconfiguring the solicitations to act in innovative ways).

As summarized by the following table, the key mental functions involved in performance control (recollection of relevant actions, recognition of relevant situations, decision about the best courses of action, awareness of the action possibilities offered by the situation) are gradually automatized as the performer advances through the higher levels of expertise.

Skill level/Mental function	Novice	Advanced beginner	Competence	Proficient	Expert
Recollection	Non-Situational	Situational			
Recognition	Decomposed		Holistic		
Decision	Analytical			Intuitive	
Awareness	Monitoring				Absorbed

The table shows how the mental functions involved in performance control become progressively pre-reflective as the skill levels improves (Dreyfus & Dreyfus, 1988)

The protocols, rules, and norms relevant to the task at hand represent the first component of the performance to be automatized, followed by the capability to decide the tactical perspective on the practical context in which the agent is situated, then by the capability to appropriately select the strategic goals to be achieved in the given circumstances, and finally by the capability to track, select, and directly respond to specific action opportunities among the multitude that the situation affords.

The following remarks (Sect. 3.1.1, Sect. 3.1.2, Sect. 3.1.3) are necessary to dispel three common misunderstandings associated with the decreasing dependency on reflective thought and contentful knowledge that marks the automatization of control in experts.

3.1.1 In the performance domain, habits are trained for a specific purpose

The key aspects of a performance typically *can*, and most often *must*, become automatized via targeted training practices before they can be controlled skilfully. While not all habits are useful to an expert or beneficial for performance, expert performances always require a broad, cohesive set of well-chosen, well-diversified, well-trained habits to attune their actions to the situation. That is why, in the performance context, the development of habits via practice and exercise is purposeful and deliberate. As the performer achieves skilful control through habit formation, the depth and breadth of

action automatization prove that the performer is prepared to control the relevant performances skilfully; conversely, a superficial, narrow level of automatization proves that the performer's control is not mature enough for skilful execution.

Unlike compulsive habits developed accidentally, the actions controlled by well-trained habitual dispositions developed for a specific purpose are normally initiated and interrupted voluntarily. Thus, the automatization of well-trained behaviours does not cause the agent to suddenly lose their ability to mindfully monitor the details of their performance's geometry or deliberately decide the exact trajectory of their actions based on strategic considerations, if this is what the agent wants or what is required by the praxical norms of a specific task. Remarkably, the execution of automatized actions (unlike the execution of compulsive actions) is never experienced as an imposition. On the contrary, automatization implies a new form of freedom as it offers the possibility to execute certain actions without involving the kind of centralized, reflective processes that consume cognitive resources to manipulate contentful representations. So, an experienced performer may or may not be able, but certainly is not required, to represent to themselves the low-level sensorimotor details of their actions in order to execute them correctly (Dreyfus, 2000; Kelly, 2000).

It is to be noted that slowly practicing an action according to a rigidly repetitive exercise protocol (even a mechanical, scripted routine) is one of the possible methods to achieve automatization (although it rarely is the most effective one, cfr. Gray, 2021), but this does not imply that automatization always requires mechanical repetition or that automatized actions necessarily have the detrimental characteristics of a repetitive mechanism (Cappuccio & Ilundáin-Agurruza, 2020). Rote repetition of standardized movements, accompanied by attentive self-monitoring and explicit instructions, may or may not be beneficial during training and pre-performance routines but it is inevitably detrimental and highly disruptive during peak performance, when all the attentional resources must coherently converge toward the action's intended outcomes (Wulf, 2013, 2015).¹⁰ Ultimately, the goal proper of forming well-trained habits is making control more decentralized and thus autonomous, not indefinitely repeatable or inflexible (Dewey 1983; Hutto & Robertson, 2020).

3.1.2 Habits acquired for a purpose are neither unintelligent nor involuntary: they are "autonomous"

Dreyfus might have at times indulged in the praise of "mindless" (in the sense of "not cognitively demanding") performance to rhetorically emphasize the anti-cognitivist posture of his doctrine, but he certainly did not intend to suggest that habitual agents are insentient zombies (Breivik, 2013) or puppets mechanically manoeuvred by their own habitual dispositions (Fridland, 2014).

What Dreyfus intended to suggest is compatible with Fitts and Posner's (1967) model of attention and the dual-process models of intelligence: the progressive

¹⁰ It is also to be noted that, as remarked by Gray (2021), the repetitive training of abstract movements is not the most effective and meaningful manner to obtain automatization: on the contrary, the repetition of modularized and standardized movements abstracted from the ecological context that solicits them reduces the adaptive flexibility of the habits acquired, hence delaying the skill development process. More specific considerations apply to different kinds of skilful activity.

minimization of the cognitive resources committed to low-level control corresponds to the gradual increase of uncommitted resources available for other, higher-level aspects of the performance (Evans & Stanovich 2013, Neemeh 2022). However, as critically remarked by Christensen (2019), Dreyfus also emphasizes that the progressive automatization of the action components tends to progressively reduce, not increase, the performer's overall reliance on decisional procedures that utilise explicit inferences based on rules and instructions (that Dreyfus calls "calculative reasoning").

This needs clarification: that experts rely less on calculative reasoning does not mean that they make less intelligent or fewer decisions, or that they cannot rely anymore on self-focussed thought and contentful knowledge, especially if the situation solicits them to rely on those. Rather, it means that calculative reasoning is progressively replaced, as skill develops, by the expert's capability to exert their "deliberative rationality", which greatly grows in scope and efficacy as the experts develop their "perspectival deliberation" ability, i.e. a pre-reflective readiness to intuitively adopt the practical perspective most appropriate to the task at hand (Dreyfus, 1997). This readiness leverages either the performer's familiarity with the practical perspectives previously adopted to successfully address similar tasks or the performer's capability to improvise new perspectives based on the familiar ones (Rousse & Dreyfus, 2021).

Importantly, that an action is automatized does not mean that the agent is unable to foresee its broader implications, act purposefully and rationally, or pay attention to what they are doing; on the contrary, it means that the control of the details of the execution becomes *autonomous* from the explicit, centralized decisional processes that are based on inferential, analytic and calculative reasoning (Fitts & Posner 1967; Logan, 1988). An automatized performance (compared to a non-automatized one) does not only *allow* a greater and greater independency from reflective control and contentful instructions, but also *requires* such an independency (as demonstrated by the direct correlation between Stroop interference and skill-level, see Logan, 1988) because automatized performance relies on control routines incorporated into a background of contentless know-how. Habitual control neither needs nor can directly tap into the pool of cognitive resources specifically dedicated to representational processes, such as explicit, analytic, attention, working memory, formal inference, and logical analysis (Beilock & Carr, 2005).

That an automatized action is controlled autonomously means neither that it is repeated mechanically nor that it is executed in a rigid, scripted, stereotyped, or inflexible manner (Hutto & Robertson, 2020; Miyahara & Robertson, 2021; Robertson & Hutto forthcoming). Both clinical evidence and everyday experience attest that experts confidently rely on habitual control. This motivates our justified scepticism toward the claims that autonomously controlled actions are not "initiated by intentions" and "run to completion without regulation by goal" (Moors & De Houwer, 2006, cited by Christensen et al., 2016), as if autonomous control was a defective, substandard process: in fact, the compulsive production of stereotyped actions is a pathological condition that has little to do with the way experts normally carry out their habitual actions (Rietveld, 2012; Cappuccio & Wheeler, 2012), hence the two concepts should be kept carefully distinct; moreover, the risk of initiating inflexible, poorly controlled actions due to a scarce consideration of their goal and an

inadequate understanding of the context is even greater for the novices than for the experts who have already automatized their execution.

3.1.3 Control becomes more intelligent and reliable when it becomes autonomous

Habitual control is strictly indispensable to produce competitive performances. Making an action autonomous, via automatization, is an advantageous way, and often the only way, to deliver that action in a rapid, fluid, richly adaptive, context-sensitive, intuitive, spontaneous and thus creative and original manner (Dewey 1983; Carr, 2015; Rucińska & Aggerholm, 2019). The main advantage of automatization is that spontaneously initiated, instinct-like, and self-motivating forms of intuitive control become available to the performer like their second nature. This reduces the burden for the cognitive system, promotes a superior degree of internal coherency and, most important guarantees contextual adequacy, which is why control risks being severely slowed down, if not entirely disintegrated, when the agent decides to micro-manage the actions that have already been automatized.

Competent and proficient performers don't need to make micro-decisions about the action structure and, when their expertise is fully mature, they don't even need to explicitly reflect about the situational context or make effortful strategic decisions. Skilled control processes typically occur below the level of representation, at the margin of the expert's explicit judgment. They are incorporated within direct (non-vetoable) perceptual processes that the agent can, but does not need, to actively monitor (Neemeh, 2022).

3.1.4 Provisional conclusion: the direction of skill development

The Dreyfus & Dreyfus model corroborates DisC because it recognizes that many components of performance execution can be automatized and that the process of skill acquisition proceeds consistently and irreversibly toward the broader and deeper autonomy of actions, decisions, and thoughts. The destination of the learning process is the same whether the development of pre-reflective competences is preceded by a reflective stage or not (Hutto & Sánchez García, 2015; Masters & Maxwell, 2004). Even though pre-reflective cognition is ontologically more fundamental than reflective cognition (because non-representational intentionality is more primordial than representational intentionality), pre-reflective control can become more sophisticated and effective than reflective control insofar as it embodies the kind of intelligence that is uniquely acquired via practice (Hutto & Myin, 2017; Kiverstein & Rietveld, 2015). Thus, habitual control is not a simpler and less refined form of intelligence: rather, it is the specific form of performative, practical intelligence that derives from, and is attested by, the progressive eclipse of reflective control and contentful knowledge. Some objections to the First Argument will be discussed in Sect. 4.1.1.

3.2 Second argument: the phenomenological foundations of pre-reflective control

The second argument builds on Dreyfus' phenomenological analysis of the experiential condition of skilful performers, which is "embodied, embedded, situated and

absorbed” (Fridland & Pavese, 2020). Intellectualists and cognitivists contend that skilful behaviours must be controlled by contentful intentions, not habitual dispositions, as they assume that only the former (by conveying truthful indications about the most effective ways to perform) can enable the flexible adaptive behaviour that experts need to perform successfully (Butterfill et al., 2014; Bermúdez, 2021; Brozzo, 2017; Ferretti & Zipoli Caiani, 2021; Mylopoulos & Pacherie, 2019 and Wu, 2013 among others).

Two assumptions are typically evoked in support of this representationalist doctrine: first, flexible and adaptive control needs sensitivity to contextual conditions; second, such a context sensitivity requires explicit representation of the action goals and reasons to act. Dreyfus’ analyses (2000) validate the first assumption. However, not only does his phenomenology indicate that the second assumption is unwarranted, but it entirely flips it over, showing that representation necessarily constrains and limits our context sensitivity (Dreyfus, 1997, 2002a). To unpack this view and its key terms (background, pre-comprehension, situated familiarity) we need to introduce three observations, respectively organized as two premises of an argument and the conclusion logically deriving from them.

3.2.1 First observation: both habit and representation presuppose a background of performative familiarity

Experts confidently anticipate the implications of their actions (Christensen & Bicknell, 2019). This is possible because they are intuitively familiar with an open-ended horizon of practical preconditions that define the significance, appropriateness, and purpose of their actions (Dreyfus, 2014). Experts draw from this horizon their hermeneutical capacity to discriminate between very similar contexts and confidently navigate the familiar ones. Dreyfus, appealing to the Heideggerian notion of “Being-in-the-world” (Heidegger 1927; Taylor, 1993; Wheeler, 2005, 2008), calls this horizon “background” to indicate the incalculable number of formal and material preconditions behind every particular variety of skilful coping (Dreyfus, 2008, 2012; Fusché Moe, 2004).

The background concretely shapes and fills of operational significance the pre-reflective know-how that informs the expert’s decisions during their performance. Representational content can guide performance only when it is situated against a backdrop of concrete know-how that precedes and vitalizes know-that, being necessary to make contentful representation practically meaningful and contextually relevant (Hutto, 2012). That is why, despite the claim by Stanley and Krakauer (2013) that “motor skill depends on knowledge of facts”, both experts and neurotypical people who have not suffered brain traumas usually do not need objective knowledge of facts and rules to choose the most efficacious course of action in familiar circumstances (Masters, 1992; Masters & Maxwell, 2004; Porter, Wu & Partridge 2010; Komar et al., 2014). Even when their performance benefits from some know-that, experts can apply it in a way that is distinctively relevant and appropriate to context because they directly perceive the most appropriate opportunities of actions afforded by the circumstances, not because they were provided with rules and objective facts (Kiverstein & Rietveld, 2018; Rietveld & Kiverstein, 2014).

Dreyfus' persuasion about the fundamentally non-representational architecture of human intelligence originates from his theoretical research on general artificial intelligence (AI) and the history of its fiascos (Dennett, 1984; Dreyfus, 1992, 2008; Dreyfus & Dreyfus, 1988). The systematic failure of every known attempt to build general AIs using descriptions of actions and stored heuristics justifies the conjecture that the experiential background required to navigate a meaningful context of action cannot be grounded in detached, contentful representation. If an artificial cognitive system solely based on finite representations of the world were realizable (hypothesizing that it is conceivable) it would only offer a rigid, stereotyped, incomplete guide to action because it would lack any authentic familiarity with the relevant practical contexts (Dreyfus, 1992). Portraying the human cognitive system like a disembodied mechanism removed away from the pre-verbal and non-conceptual background of real-life experience is, from this perspective, utterly misleading.¹¹

3.2.2 Second observation: performative familiarity affords a form of comprehension that is more fundamental than representation

A deep and broad background of familiar experiences and performative know-how provides the pre-comprehension of the context that experts need to adjust their actions to the circumstances. With “pre-comprehension”, philosophical hermeneutics indicates the kind of implicit comprehension that both precedes and is presupposed by analytic understanding (Gallagher, 2004); such a tacit background of experience and knowledge is purely practical and performative in nature; it sustains and, at once, is sustained by, the pre-reflective engagement of experts in their skilled activities (Dreyfus, 2012). The pre-comprehension relevant to skilled performance originates from a primordial dimension of pre-verbal, interactive experience that is solely disclosed through embodied involvement and affective solicitations (Andler, 2000). It is accessible only in the form of a rich perceptual responsiveness to complex, fluid, multifaceted circumstances (Dreyfus, 2002a, 2008). As such, it exceeds the low granularity of propositional knowledge and conceptual analysis. The theory of autopoietic enactivism expresses a comparable concept when it argues that the fundamental *sense-making capabilities* of a certain lifeform predate representational cognition and arise from the history of embodied dealings established by that lifeform in their unique world-environment (Varela, Thompson, Rosch 1991; Di Paolo, Buhrmann & Barandiaran, 2017).

This notion builds on the principle that real-life experience necessarily transcends its own representation: an indefinite number of different representations can be created to fixate this or that set of aspects of experience, but no finite set of representations could exhaust the transcendent complexity of the life-world that animates them. Thus, representations are structurally insufficient to make the agent practically acquainted with the inherent intricacy and perspectival depth of contextual circumstances: while local, particular perspectives on the background may become representable at this

¹¹ This claim becomes ever more important in the age of generative AI and Large Language Models, as the incapability of chatbots to produce accurate, consistent accounts of practical engagement with real-life conditions suggests that the simulation of human intelligence is inherently limited and faulty when based only on disembodied, linguistic descriptions.

or that moment, the background as an interconnected whole inevitably exceeds the knowledge of particular facts. This means that contentful judgments and inferences about facts can only offer a bounded, derived, and uncertain guide to skilful action (Van den Herik & Rietveld, 2021). Independently of how large, numerous, and granular they are, representations simply do not provide an agent with the kind of *familiar* (transparent, trusted, immediately available), *inherent* (tied to the subject's unique experiential history), and *pragmatic* (actional, embodied, purposeful) comprehension that they need to make sense of the real purpose, motivation, and implications of their own actions.

Conversely, the expert is exactly the sort of agent who can count on the *holistic* (systemic, non-analysable in discrete parts) and *situated* (sensitive to an unquantifiable number of massively interconnected experiential particulars) *pre-comprehension* (non-conceptual competent familiarity) of their own actions' *practical background* (the inexhaustible horizon of pre-conditions that implicitly defines the causes and reasons of their actions, see Dreyfus, 2012; Cappuccio & Wheeler, 2012). Control is skilful only if it is sufficiently fine-grained and flexible to incorporate this contextual pre-comprehension (Wheeler, 2008; Cappuccio & Wheeler, 2010a, 2010b, 2012). Only contentless, embodied habitual dispositions, not contentful knowledge, are able to sustain a practically relevant pre-comprehension: while representation of facts is self-contained, value-neutral, merely receptive, and contextually detached, well-trained habitual dispositions allow perception to be active, intelligent, value-laden, predictive, and practically tuned to numerous contexts so to enable the fast, precise adaptivity required to perform skilfully (Beilock & Gray, 2012; Craighero et al., 1999; Gray, 2014; Rizzolatti & Craighero, 2010; Witt et al., 2007).

3.2.3 Third observation: reflective and pre-reflective forms of comprehension obey two irreducibly different normative regimes

This observation concerns the normative dimension of skilful control (Rietveld, 2008), that Dreyfus derives from Merleau-Ponty's *Phenomenology of Perception* (1954/2012). According to this phenomenology, each instance of effective action control may rely only on one of two mutually exclusive types of intentionality (Dreyfus, 2000; Fusche Moe, 2016; Jackson, 2018): *cognitive intentionality* (representational) vs *motor intentionality* (dispositional).¹² Only the former can have a semantic content and truth values thus, from a normative perspective, what distinctively characterises cognitive intentionality is that it serves an adequative (i.e., correspondence/similarity-seeking) semantic function; motor intentionality, in turn, does not distinguish contents from attitudes (Kelly, 2000), hence it serves a compensatory and homeostatic (i.e., dynamical balance/equilibrium-seeking), regulatory function.

¹² This duality of forms of intentionality hints at what, according to dual-process theorists, is a fundamental duality of types of cognitive processes, i.e. a dichotomy between irreducibly different forms of evaluation and decision: the former, i.e. "Type 1" processes, are "intuitive", "autonomous" (not explicitly monitored and unsupervised), "fast", "involve no working memory", and "cannot be rationally vetoed"; the latter, i.e. "Type 2" processes, are "reflective", "decoupled from the immediate situation", "slow", "require working memory" and can be rationally scrutinized (Evans & Stanovich 2013). Stuart Dreyfus (2015), however, rejects this analogy, asserting that pre-reflective control is not reducible to Type 1 cognition but constitutes a third Type (Type 0).

They sustain two completely different epistemic processes: know-that, which is normatively satisfied by *truth, correctness, or accuracy conditions*, and know-how, which is normatively satisfied by *far-from-balance or “improvement” conditions* aiming to bring the agent closer and closer to a desired, but implicit, goal-state (Dreyfus, 1999, 2007a). In accord with these characterisations, the success/appropriateness of an action controlled reflectively is attested by the structural isomorphism to an explicit plan, rule, or conceptual model, while the success/appropriateness of a pre-reflectively controlled action coincides with the progressive reduction of the distress to which the system is subject and the minimization of the sense of deviation from an optimal Gestalt configuration that is never entirely explicatable yet always perspicuous in terms of uncertainty or entropy (Dreyfus, 2002a; Fusché Moe, 2016).

Habitual control generates rich responses under the urge to reduce the sense of deviation from an optimal global configuration (e.g., a certain homeostatic balance or dynamical coupling with the environment, see Dreyfus, 2002a). Only the contentless, embodied readiness to act that distinctively informs motor intentionality, not the contentful knowledge of facts conveyed by representation, can provide the kind of pre-comprehension required by experts to navigate the context of action adaptively and flexibly (Ungureanu & Rotaru, 2014). This is true regardless of how accurate and structured the know-that provided by representations is. Contentful representations and habitual dispositions exist to fulfill entirely different kinds of expectations, and only the latter are directly relevant to control skilful performance: by definition, a representation is correct if and only if it truthfully corresponds to the reality it is meant to represent; in turn, a disposition is efficacious if and only if it produces the expected form of engagement with the environment. The skilful value of a performance is measured by the qualities and merits of its execution, not the veridicity of the assertions associated with it, therefore only embodied habitual dispositions are categorically relevant to skilful expertise.

3.3 Third argument: the epistemic qualities of expert know-how

The third way to defend DisC consists in showing that skilful control manifests certain epistemic qualities that exclusively pertain to contentless know-how. Skilful expertise presupposes a performative knowledge that is inherently embodied (Beilock, 2015). This implies, *inter alia*, that experts do not only simply *use* but literally *incorporate* the know-how necessary to exert online control through perceptual, motoric, and anticipatory-predictive dispositions (Dreyfus, 1996, 2002a, 2007a, 2007b; Robertson & Kirchhoff 2020). This know-how is grounded exclusively in habits because only embodied habitual dispositions can offer a specialised competency that is immediately, constantly, and consistently accessible, more like a constitutive trait of the performer's own being than an external resource instrumentally used by the performer.

Experts, in other words, do not *simply know how* to reliably produce skilful performances; they also have this knowledge in a *private, context-sensitive, and trusted* manner. Skilful control necessarily incorporates these three *epistemic qualities*, i.e. properties that differentiate the kind of knowledge mastered by expert performers from other kinds of knowledge. These qualities provide action control with the precision,

rapidity, flexibility, and robustness uniquely associated with skilful expertise: not only are they deeply ingrained in the pre-reflective control enabled by habitual dispositions; they are also at odds with the reflective control informed by representational content. This is clearly shown by the fact that only the performances informed by embodied know-how, not the actions controlled by representations of facts and rules, display the spontaneous, self-reliant, harmonious, parsimonious, and effortless execution that distinctively characterizes the control of an expert.

In what follows (Sect. 3.3.1, 3.3.2, 3.3.3), I will describe more precisely these qualities, discussing why they can be found only in the know-how embedded in habitual dispositions. I will partly utilize investigations on tool use initiated by theorists of extended/distributed cognition (Clark & Chalmers, 1998; Menary, 2010; Sterelny, 2010; Wheeler, 2012). Their analyses suggest that the practical knowledge required to perform skilfully must be privately owned (Sect. 3.3.1), context-sensitive or non-neutrally situated (Sect. 3.3.2), and trusted (Sect. 3.3.3).

3.3.1 Privately owned (characterized by a privileged access deriving from the unique constitutive relationship between the expert and their skill)

Skilled performers possess their skills in an intimately familiar manner, i.e. a manner that implies phenomenal transparency and fast and reliable access. Experts *have* their skilful expertise in a deep sense, in modalities and forms that inherently define their performative engagement: *not only do experts own this knowledge*; they also have an *exclusive privilege* over it in reason of the unrepeatable and irreversible personal history that ties together an agent and their performance. According to Bäckström and Gustafsson (2017), this idea can be traced back to Gilbert Ryle (1949/2009): performative know-how is what experts *typically do*, not some information they *store*; as such, it diachronically constitutes the expert's identity as an expert through the history of their dealings with familiar contexts of action. Before explicit representation of goals is useful or even possible, their well-trained dispositions already make them receptive of intrinsically motivating solicitations and pre-reflectively aware of the contextual relevance of the consequences of their actions (Cappuccio et al., 2019b; Miyahara & Robertson, 2021).

Contentful knowledge, in comparison, resembles a neutral commodity or a currency that performers can spend only in particular situations: scarcely available due to the cognitive constraints of the human intellect (e.g., working memory limitations, see Maxwell et al., 2003; Beilock & 2005), yet indefinitely accumulable and in principle universally transferable (via linguistic and figurative forms of notation), contentful knowledge about action is usable by any agent but does not belong to (let alone constitute) any of them. Contentful knowledge is never truly owned because access to it is necessarily mediated: know-that is always merely borrowed by the agent, who can use it only upon concession of an external source. Every situation in which an expert relies on a defined set of contentful rules or instructions is somehow similar to that of a chess player who needs to consult her move-by-move handbook between a move and another; football players who can play in formation only when their coach provides them with the relevant strategic instructions; a boxer who needs a transmitter concealed in his ear to receive precise instructions about where, when, and how he has to hit the

opponent's body; an Alzheimer patient who writes detailed instructions in a notebook to remind himself all his daily activities (Clark & Chalmers, 1998). The knowledge provided by these epistemic add-ons can be instrumentally exploited by their agents but it is never deeply ingrained in them: as convincingly argued by Facchin (forthcoming), phenomenal transparency is more of a complication than a requirement to cognitive extension, which is why many extended cognitive activities are not skilful. These supplements may work just fine and produce outcomes that are functionally equivalent to the outcomes they are meant to reproduce or replace (hence, the validity of parity principle is not at stake here, Wheeler, 2012), but (as I will systematically argue in Sect. 5.3.3) the bare ability to produce outcomes characteristically associated with a skilful performance does not guarantee that those outcomes are produced skilfully.

3.3.2 Context-sensitive or non-neutrally situated (specifically tailored to particular sets of unrepeatable circumstances)

Only the habitual/dispositional form of action control can be truly skilful, not the reflective form, because only contentless know-how (associated with the former) can be context-sensitive and contingently situated, not contentful know-that (associated with the latter). The key difference between the former and the latter is that contentful representations, unlike contentless dispositions, presuppose the very distinction between vehicles and contents of cognition, and such a distinction implies the functional neutrality of the instantiating medium and the multiple realizability of the content (Wheeler, 2010a).

This means that contentful knowledge about objective facts and verbal instructions can be *re-implemented* in a variety of conventional formats (oral, written, visual, etc.), *physically externalised* (e.g., stored in a tool, like a notebook, or assigned to another physical support of information, like an architecture) or *entirely outsourced* (when judgments are derogated to another cognitive system, whether natural or artificial): in all these cases, the vehicles of cognition remain neutral to its content, thus the way the agent learns some know-that does not modify their performance. The contentful know-that that guides reflective control (rules, schemas, and instructional maxims) can be identically replicated over time or relocated in different spatially situated vehicles while remaining functionally unchanged (Cappuccio & Wheeler, 2010a, 2010b; Wheeler, 2010b).

On the contrary, know-how does not allow for the content-vehicle distinction because the details of the means (how the performance is executed) and the details of the ends (the intended outcome of the performance) concretely map into one another and establish reciprocal constraints as they are two constitutive sides of one and the same performative experience: from the pre-reflective perspective of our tacit dispositions, recognizing the motivating goal of an action in a given context amounts to knowing how to execute that action in that context and, conversely, recognizing how a particular action is executed means immediately understanding why that action is executed in that particular way (Fusche Moe, 2016).

Now, skilful expertise always arises from a unique history of interactions characterised by the particular spatio-temporal situation of the learner. Skilful control gradually forms and develops through consecutive stages in accord to the specific

rhythm and style of each learning process (Francesconi & Gallagher, 2019). Due to its acute sensitivity to the holistic configuration of sensorimotor contingencies (Ilundáin-Agurruza, 2016) and ecological constraints (Gray, 2021), skilful control is always informed by the unrepeatable contextual circumstances in which the agent-environment coupling forms and by the specific situation of the agent (Bocanegra & Hommel, 2014). Such rich and adaptive sensorimotor responsiveness to granular, massively interconnected circumstances cannot be learned like a formula from a book or transferred unmodified from a cognitive system to another: it is effectively learnt only when it is practiced in ecologically relevant, and highly cohesive, value-laden conditions that instantiate a set of unique affordances, i.e. opportunities for action highly specific of a context (Gray, 2021; Segundo-Ortín & Heras-Escribano, 2021). That is why skilful control necessarily has the nature of contentless habitual dispositions, not of contentful representations: skilful know-how must be finely calibrated to match the tacit context in which embodied habitual dispositions form and to which they are coupled.

3.3.3 Trusted (reliably available, confidently usable, familiarly applicable)

To control their actions skilfully, experts need a reliable access to the relevant embodied expertise: this means not only that they must perfectly trust the information they rely on (e.g., procedural memory, Eichenbaum & Cohen, 2004), but also that the recollection and application of this information must occur through an absolutely simple and immediate process. No additional interpretative or inferential effort should be attached to this process, as this would hamper action control, making it slow, hesitant, and unnecessarily complicated. This kind of simple and immediate access is offered by the embodied know-how embedded in habitual dispositions: when the agent's competency is habitual, the relevant pre-comprehension of the context of action is immediately, implicitly available to them through the solicitations that they perceive in the environment, as these solicitations exert on the agent an inherently motivating force in the form of affordances (Heras-Escribano, 2021). For a skilled expert, relying on their embodied habitual dispositions, detecting affordances in the environment and comprehending the practical implications of one's actions in the given context of activity are simultaneous aspects of one and the same perceptive experience (Bruineberg et al., 2019).

On the contrary, this immediate access is not available to cognizers who must process contentful rules and instructions to control their actions: they are hampered and delayed because they must constantly make interpretative and inferential attempts to decode the information contained in their stored representations. The information conveyed by representations is less trustworthy because its utilization is *mediated, tentative and demanding*.

Mediated While habitual dispositions allow us to directly pick up the most relevant contentless information from the perceptual environment (i.e., salient covariational regularities), the process of accessing information through representation is mediated by our capability to retrieve and correctly decode (or infer) its content. That is why knowing pre-reflectively how to do X via Y and being told that Y is the best way

to do X are two very different kinds of epistemic access (Flegal & Anderson, 2008); in the first case, the required knowledge is transparently accessible as part of our phenomenal experience, in the latter it must be parted away from online experiential evidence and linked to non-evident mechanisms and processes located beyond the horizon of perceptually perspicuous events (Kiverstein & Rietveld, 2018).

Tentative Every time an agent relies on contentful know-that to guide their actions they must reconstruct all the relevant semantic and causal relations through hypotheses, guesses and, if the source is questionable, a leap of faith. The testimony of a witness is a less trustable source than direct evidence: analogously, the performers who rely on rational inferences or representations of objective facts inevitably perform less competently than experts because their inferences need to be rationally validated and facts need to be verified (Wheeler, 2010b; Miyahara, Ransom & Gallagher 2020). This uncertainty prevents the agent from promptly responding to circumstances and confidently investing their motivational and attentional resources in a consistent project of action. A tentative access to information is incompatible with skilful control because it is uncertain both *probabilistically* (in terms of prediction of the chances of success) and *psychologically* (in terms of emotional self-regulation). The resulting confidence erosion can dramatically curb the control of actions requiring fine-grained spatio-temporal regulation (Cappuccio et al., 2019b).

Demanding Demanding information processing (e.g., instructions encoded in a declarative format, see Maxwell et al., 2003; Jackson et al., 2006) is incompatible with skilful control because it is taxing in both *volitional* and *cognitive* terms. Action control cannot be skilful if it requires struggling efforts that compromises the agent's readiness and responsiveness to contextual contingencies. Effortful action control tends to be imprecise, cumbersome, and sluggish and hence potentially detrimental to performance outcomes, which presuppose the parsimonious use of the agent's physical and cognitive resources (Cappuccio & Ilundáin-Agurruza, 2020). Note that not all effort is detrimental to skilful activity (Brownstein, 2014a, 2014b; Jackson, 1996; Swann, 2016): for example, the moderate, continuous effort perceived during flow activities feels like a stimulating and invigorating challenge and is accompanied by an attenuated or peculiarly modified sense of agency (they *let their action happen* as a spontaneous response rather than *making it happen* as the effect of an effortful deliberation, see Swann et al., 2016).

On the contrary, the intellectual effort required to process recurrent contentful thoughts (which can be involuntary, intrusive, and obsessively reiterated) can severely disrupt performance, even when the cognitive load is manageable and commensurate to the agent's cognitive capabilities (i.e., the task's cognitive load is not onerous enough to cause distraction or overload). Contentful thoughts can cause performance disruption in at least two different ways (Cappuccio 2015; Cappuccio et al., 2019a, 2019b): first, contentful thoughts can solicit a compulsive self-monitoring stance, which interrupts the expert's continuous, highly optimized, flux of pre-reflective sensorimotor activity, fragmenting it into discrete components (performance disruption by execution/internal-focus, see Beilock et al., 2002a, 2002b, 2004, 2008; Beilock &

Carr, 2001, 2004; Beilock, 2011; Gray, 2004; Wulf, 2015; Wulf & Su, 2007); second, contentful thoughts risk causing an unwanted psychological regression in the expert performer, triggering in them an uncontrollable search for rules and instructional maxims, thus bringing the expert performer back to a novice-like condition (performance disruption by dispositional reinvestment, see Masters, 1992; Masters & Maxwell, 2004, 2008; Jackson et al., 2006).

3.4 Summary of the three arguments and provisional conclusions

These arguments defend the idea that performance is either skilful or controlled reflectively. It cannot be both. Consequently, the kind of action control that we legitimately call “skilful” must be grounded in contentless habitual dispositions and cannot be grounded in contentful representation. Habit is the only source of the pre-reflective pre-comprehension that experts need to adjust promptly and flexibly to complex, diverse contexts of practical activity. Reflective and pre-reflective control are distinct and incompatible as they have entirely different geneeses, obey different normative principles, and demand different epistemic resources. Thus, pace Conciliatory Habitualism, habitual control is *not only* a distinctive marker and an enabling condition of skilful expertise (i.e., one of the reasons sufficient to infer that an agent masters the relevant skills and one of the prerequisites necessary to recognize an action as the skilful performance of an expert); *it is also* the exclusive epistemic root and the only cognitive constituent of expertise. In contrast, contentful judgments deriving from representational cognitive processes, like decisions based on propositional knowledge and rules, do not play any necessary, sufficient, constitutive, or causal role in expertise because they do not meet the genetic, the normative, and the epistemic requirements of skilful control.

4 Intermezzo: an important clarification

4.1 Three prima facie exceptions

There are at least three prima facie exceptions to Radical Habitualism, which seem to challenge Dreyfus’ intransigent anti-representationalism. First, Dreyfus himself (2002b; Dreyfus & Dreyfus, 1999) recognizes that propositional knowledge plays a beneficial, often indispensable, role in the process of skill acquisition (learning) and maintenance (exercise). Second, he admits that moderately experienced performers (“competent” level and above) may appeal to representational resources when they face unfamiliar or problematic tasks that, while falling under their general domain of expertise, cannot be successfully dealt with relying exclusively on habitual dispositions (Dreyfus, 2007b, 2008; Wheeler, 2005, 2008). I will consider a third exception too: regardless of how difficult and unfamiliar the task at hand is, performers may find it convenient to rely on explicit instructions and other contentful representations, as the knowledge they provide can offer a competitive advantage and help producing superior outcomes (Sutton 2007; Cappuccio & Wheeler, 2012). While these three

cases represent instances of unhabitual, reflective control, none of them contradicts the DisC.

Radical Habitualism does not exclude that some representational processes may opportunistically serve the performer's best interests, operating (in a merely contingent and non-constitutive manner) as a temporary scaffold or external prosthesis of performance (Dreyfus, 2002b, 2008). Like in Dreyfus' famous metaphor, contentful representations are like *training wheels*, which undoubtedly constitute beneficial aids to any child who is still learning to ride their bicycle (this paper's opening quotation, taken from Dreyfus, 2005).

What Radical Habitualism contests is that the advantages provided by representational content could ever be a legitimate criterion to determine when agents perform as experts and which performances count as skilful: training wheels eventually must be set aside to permit the full maturation of our skill as cyclists. They don't magically turn into invisible, intangible mechanisms when their cognitive function becomes redundant: the training wheels entirely disappear from the horizon of an expert cyclist's performance.

Crucially, the valence of DCT is *also normative* and not only descriptive: it does not simply tell us what a successful *performance looks like* and which cognitive processes underpin their production (e.g., if, when, and how our training wheels improve our biking performances); it also gives us rigorous principles to appropriately rethink the notion of expertise and accordingly *decide which performances deserve to be considered skilful* (e.g., no cyclist relying on prosthetic wheels to keep their balance could ever be considered an expert). DCT does not ignore the consolidated scientific understanding of how well-trained agents do or do not access know-that to perform successfully: rather, it builds on scientific evidence to correctly situate the function of know-that within the processes of skill development and application.

4.1.1 First exceptional case: representations in skill development

Regarding skill training and re-training, agents obviously do not operate as experts when they rely on explicit rules and instructions for developing a new skill or when they deliberately practice some component movements to maintain or correct a related existing skill or to upgrade that skill to the next level. Therefore, the fact they focus on instructions and rules to develop new embodied habitual dispositions (as in learning), to fine-tune the existing ones (as in deliberate practice), or to regain full access to the damaged ones (as in rehabilitation, see Stanley & Krakauer, 2013) does not undermine Radical Habitualism. It is true, as emphasized by Ericsson (2008) and Montero (2016), that a serious practitioner (regardless of their expertise level), *never stops practicing* – in the sense that skill development is a life-long endeavour and in the sense that even competition in the agonistic context constitutes an opportunity to learn. But this does not change the well-demonstrated fact that exerting one skill with the explicit intent to improve the skill itself (deliberate practice) or rather to produce the best possible performance (peak performance) involve two entirely different modalities of control: the former aims to produce well-trained habits, the latter to effectively use those habits.

Montero (2016, p. 130) defends a different perspective on deliberate practice, as she claims that “experts, in at least many domains, deliberately try to get better during

performance” and that “deliberate improvement is typically an element of not only practice but also of a performance or tournament”. It seems to me that these claims are problematic in at least two ways. Firstly, they are *empirically problematic* because they are systematically contradicted by evidence: the competition-vs-practice distinction is universally valid across all domains of expertise, including the most intellectual ones, like mathematical problem solving and linguistic expression (Beilock, 2008; Beilock et al., 2002a, 2002b; Wulf, 2013, 2015, etc.) This cannot be accidental. Training practices benefit from attentive self-observation since the trainee needs to deliberately, temporarily, selectively use their focussed reflection to single out and correct this or that element of their execution; peak performance, in turn, demands an entirely different emotional, cognitive, and motivational profile, as the agent must leave in the background all the aspects of the execution that are already automatized in order to maximize their focus on the immediate outcomes of their action. Secondly, these claims are *conceptually problematic*. Not only because they presuppose that reflective processes are needed to deliberately improve expert performance (as if training wheels were needed by a professional cyclist), not only because the presence of these processes is hardly verifiable (like the material presence of invisible wheels), but also because, if confirmed, the presence of reflective processes would not constitute legitimate evidence of a superior skill (reliance on training wheels is not a sign of mature expertise).

4.1.2 Second exceptional case: representations in unfamiliar and troubled circumstances

Regarding troubled and challenging performances in unfamiliar scenarios, Dreyfus (2012) acknowledges that, in order to overcome unexpected obstacles, experts may rely on minimally or quasi-representational processes. Wheeler (2005, 2008) calls “action-oriented” these representations and, in contrast with fully-fledged representations, characterise them as *agent-centric*, *pragmatic*, and *context-sensitive* instances of “special purpose adaptive coupling”. Dreyfus (1990, 2008, 2012), using Heidegger’s (1927/1996) terminology, calls “unready-to-hand” the troubled modality of performance that requires an agent to produce on-the-fly and to immediately consume, without storing them, “minimal” or quasi representations of the actions necessary to surmount the difficulties. This modality is situated between the “ready-to-hand”, or familiar modality (pre-reflectively controlled) and the “present-at-hand”, or fully-fledged representational modality (reflectively controlled), because it combines the direct modality of access specifically granted by habitual experience via adaptive coupling with the detachedness of content-mediated understanding (see Wheeler, 2005, 2008 and Cappuccio & Wheeler, 2010a, 2010b, 2012).

By means of action-oriented representations, competent and proficient agents may provisionally adopt contentful rules, maxims, or models as prosthetic aids or scaffolds to compensate for the inadequateness or limitedness of their habitual dispositions, thus integrating the exploratory and tentative stance of novices and the improvisational capabilities acquired with experience. Importantly, DisC is not violated by the fact that, during troubled circumstances, agents may recall some contentful heuristics learnt during their training, for example repeating to themselves “instructional

nudges” as self-regulatory expedients (Sutton, 2007; Porter, Wu & Partridge 2010), or even inventing new rules-of-thumb (Cappuccio & Wheeler, 2012): in the first place, whether or not these contentful aids are useful to deal with the circumstances, their performance cannot be considered expert *if* the unfamiliar situation was not dealt with using the know-how embedded in habitual dispositions, because that implies the situation was dealt with (effectively or not) using some external, expediential knowledge that is instrumentally accessed by, but not constitutive of, the expert performance. Therefore, whether or not action-oriented representations are effectively used to overcome unexpected challenges, the phenomenology of the unready-to-hand does not justify the belief that mixing habits with content could make more skilful the control exerted by an experienced performer.

First, the issue has been raised that minimal representations might not be properly contentful, at least not in the sense of fully-fledged reflective control, as they do not display all the essential features of rule-based intelligence and inferential decision-making (Gallagher, 2008; Hutto, 2013). In particular, the instructional nudges evoked by performers through self-talk work more as “material symbols” (Gallagher & Francesconi 2018) serving emotional and perceptual self-stimulation and attentional self-regulation than fully-fledged contentful representations with a definite truth-value: in this case, the role they play in performance is merely exhortative or transformative, not isomorphic/projective, because they leverage the pathic, suggestive function of verbal expression (characterized by improvement, or far-from-balance conditions), not its denotative, analytic function (characterized by accuracy, or similarity-seeking conditions).

Second, in a Dreyfusian perspective, actions performed in an unready-to-hand mode are truly skilful only if the agent ultimately reverts to habitual control and limitedly to the extent that they rely on habitual dispositions (Dreyfus, 2008; Wheeler, 2008). Thus, the guiding principle is still compatible with DisC: in the event, and to the extent, that performers rely on representation when facing unpredictable circumstances, they do not deal with those circumstances like real experts, even if they are experienced and well-trained practitioners.

4.1.3 Third exceptional case: representations as aids to boost performance outcomes

Know-that may be involved in successful, outstanding performances, but this does not mean that those performances are executed skilfully or that the agent controls them like an expert. It would be hubristic, for a practitioner, to believe that every action they perform successfully is the result of their expertise and their expertise only. True experts know that even the most inspired and serendipitous performance cannot be flawless like a textbook description. Dreyfus’ approach (2014) is at once more true-to-life and intellectually humble: it recognizes that many performances, including the most efficacious and outstanding ones, are transient patchworks of skilled and unskilled behaviours and even the most seasoned practitioner should be willing to admit that their performance is punctuated by more or less frequent moments in which they do not act skilfully, as they intermittently disconnect and reconnect with the familiar background of their habitual dispositions.

4.2 Assessing expertise: radical habitualism in hybrid scenarios

In all these cases, when confidence mixes with tentativeness and familiar execution demands intense mental and physical effort, judging whether and to what extent a performance is truly skilled is a multifaceted, complex task. Not only because know-how, unlike know-that, is a gradable quality (a classical Rylean puzzle thoroughly discussed by Pavese, 2017 and Stanley & Williamson, 2017), but also because assessing one's skill level involves multiple dimensions and different (possibly reciprocally competing) variables. In comparison, judging whether the behavioural components of an action are controlled skilfully is simpler.

That is why we should assume, like Wu (2013) does, that each and every component of a performance is either habitual ("automatic") or unhabitual ("reflective"). Among the key components (i.e., the cognitive/functional building blocks) of a performance we should include the fine-grained movements that compose an action (shaped by "motor acuity" according to Stanley & Krakauer, 2013), the simple goal-oriented actions composed by such movements, complex motor chains composed by simple actions, and then the processes governing the agent's situational awareness, their tactical decision, their strategic planning, their self-regulation, their online coordination with other agents, etc. This list (which is not meant to be exhaustive) includes also the "higher-order mental processes that structure and coordinate multiple lower-order automatic processes toward the completion of a goal" (Bermúdez, 2017), because –pace Bermúdez– most higher-order mental processes (i.e., the processes that organize, initiate, or inhibit the lower-order ones) can be automatized too as there is no known upper limit to the kinds of mental faculties that can become habitual (Araújo et al., 2006; Bruineberg et al., 2019; Kiverstein & Rietveld, 2018, 2021; Legg, 2021).

It is well possible that some skills don't allow some of these components to be automatized (I will discuss this in Sect. 5.2). However, if all or most of the processes underpinning performance control can become habitual, then how do we determine whether the performance as a whole is controlled in a mode habitual enough to be called "skilled"? Let us consider two different approaches to this question: both approaches admit that a performance may result from any combination of habitual and representational components, but they ultimately disagree about *which kinds of performances are controlled skilfully*.

The first, call it the "Skill-No-Skill" approach to hybrid scenarios, is integral to Radical Habitualism. This approach assumes that a performance is perfectly/completely skilled *iff* its components are optimally arranged under the control of habitual dispositions and proportionally less skilled in reason of the number and type of components that are controlled by contentful know-that: in Dreyfus & Dreyfus model, the lower levels of skill (proficiency, competence, etc.) are more imperfect and immature exactly because, unlike levels 5 and 6 (expertise/mastery) they may still involve residual elements of contentful representation (like instructions).

In a way compatible with this framework, the three epistemic qualities of habitual control offer an additional heuristic tool to classify imperfect/incomplete forms of skill: the higher or lower magnitude of these qualities reflects the expertise level of the performance and depends on how many discrete components of the performance are

controlled habitually. From a “Skill-No-Skill perspective”, the fact that most actions involve mixes of habitual and unhabitual components does not imply that both the habitual and the unhabitual contribute to make that action skilled or (unskilled): only the habitual (automatized) components contribute to qualify a certain instance of action execution as expert; on the contrary, the representational and contentful components (reflective) not only do not make an action more skilled but often make it less skilled, either because they directly interfere with the pre-reflective component of control or because they water down the overall skilled quality of the performance by reducing the pre-reflective/reflective ratio.

As for the second approach to hybrid scenarios, call it “Knowledge-No-Knowledge”, it arguably represents the baseline position endorsed by both Intellectualism and Conciliatory Habitualism. Instead of construing performance as a mix of skilled and unskilled components, this approach construes skilful performance as a successful blend of representational (content) and habitual processes (no-content) and skill acquisition as a process based on the “continuing symbiosis or bootstrapping between knowledge and non-knowledge” (Stanley & Krakauer, 2013). This view is epitomized by the original “Mesh” approach spearheaded by Christensen, Sutton, McIlwain (2016), which systematically articulates the principle that “controlled and automatic processes are closely integrated in skilled action, and that cognitive control directly influences motor execution in many cases” (ivi, p. 43).

Cognitive and automatic processes both make a major contribution to skilled action. The basic idea on which this theory is based is that cognitive control is not eliminated in advanced skill, but is rather shifted primarily to higher-level action control. This idea is not new—it can be found in varied forms in prior work. Thus, although Schmidt and Wrisberg (2008) characterize skill learning as a progression to increasingly automatic action production, they also say that increased automaticity in motor production and sensory analysis ‘frees the best performers to engage in higher-order cognitive activities, such as split-second shifts in strategy during a basketball game or spontaneous adjustments in the form or style of a movement in dance or in figure skating’ (ivi, p. 38).

Mesh has the remarkable merit of pointing out that human performance depends on the deep integration of different types of informational processes underlying perceptual, memory, and control systems. This architecture was further complexified by Gallagher and Varga (2020), who have incorporated bottom-up regulation, affective, and external material factors within the framework originally proposed by Christensen and colleagues.

Radical Habitualism does not deny that several varieties of performances, including some skilled ones, may occasionally display a mix of reflective (“cognitive”) and pre-reflective (“automatic”) control or that action automatization may free-up the cognitive resources needed by reflective control processes, as observed by Schmidt and Wrisberg (2008). But Radical Habitualism does dispute two key premises of Mesh: first, the idea that a performer relies more and more on reflective control as their skill level increases (the fact that a greater amount of resources become available for reflection does not allow us to infer that reflection becomes necessary for or beneficial to performance, as discussed in Sect. 5.4.3); and, secondly, the idea that a greater reliance on reflective

control constitutes legitimate evidence that reflective control is constitutive of, skilful performance (because the performances producing better outcomes are not necessarily those executed more skilfully and also because the most experienced performers do not always perform more skilfully than the others, as argued in Sect. 5.3).

In other words, Radical Habitualism primarily questions Mesh because it is not persuaded about the normative criteria chosen by the Knowledge-No-Knowledge approach to evaluate whether a performance is skilled or not. From any Mesh-like perspective it does not matter whether a performance is controlled habitually or not: a performance is skilful, according to the Knowledge-No-Knowledge approach, insofar as the combination of habitual and representational components/aspects, whatever they are, ultimately produces successful outcomes. This definition of expertise is more permissive but requires the introduction of *extrinsic criteria* to define success, that is: criteria that don't track how the action is controlled because they consider only the action's outcomes, which are supposed to provide an exhaustive measure of the performance's competitive efficacy. As I will argue in the next section, this approach is problematic from Radical Habitualism's perspective, which rejects the idea that bare performance outcomes could be a valid and sufficient indicator of the performance true value.

5 Notable objections

5.1 Summary of the three objections

Appealing to extrinsic criteria that verify exclusively the production of desirable and competitive outcomes, the defenders of the Knowledge-No-Knowledge Approach are in a position to raise the following question against the Skill-No-Skill Approach: instead of trying to disentangle the skilful components of a performance from the unskilful ones, is it not more convenient to assume that a performance is overall skilful if it successfully produces the outcomes that the performer is wishing for? Both Intellectualists and Conciliatory Habitualists may be inclined to believe that Radical Habitualism is somehow compelled to answer positively. One reason to expect such a concession is that it would help Radical Habitualism avoid the theoretical complications highlighted by three objections. The "Scope" Objection (Sect. 5.2): Many expert activities are exquisitely intellectual. Their control does not depend on habit because they have no embodied basis and/or do not fully automatize (e.g.,). The "Efficacy" Objection (Sect. 5.3): Reliance on reflective control, sustained by self-monitoring and deliberate decisions, allows agents to control their actions more efficaciously, producing superior performance outcomes (e.g., Montero, 2016). Finally, the "Modality" Objection (Sect. 5.4): The execution of skilful performances is distinctively associated with representational practices and the manipulation of contentful knowledge (e.g., Fridland, 2019, 2021).

In truth, Radical Habitualism can confidently give a negative answer to the previous question, having robust reasons to oppose any appeal to extrinsic evaluative criteria: first, the spatial contiguity and temporal concomitancy of contentful instructions/judgments and effective control does not prove that the former is integral or necessary to the latter; second, any representational component can undermine the

efficacy of the performance as a whole (empirical claim); third, consistently with our definition (Sect. 2), a performance is not skilful simply because it produces outstanding results but also because the way each component individually and all of them as a whole are executed displays the epistemic qualities that characterize the performance as the doing of an expert (normative claim).

Based on these considerations, Radical Habitualism can deploy exhaustive counterarguments against the Three Objections. I will present these counterarguments in the following sections (Sect. 5.2, 5.3, 5.4), arguing that each objection presupposes a misguided phenomenology of skilful performance. They erroneously presuppose, respectively, that intellectual activities don't have a habitual basis, that performance outcomes are a sufficient indicator of expertise, and that some contentful practices may be constitutive of skilful execution.

5.2 The “scope” objection: when intellectual skills are controlled reflectively

The first objection questions the reach of Radical Habitualism more than its intent. This objection fails to acknowledge that, by definition, all skilful performances, whether eminently sensorimotor or not, are characterized not only by their tendency to produce desirable outcomes but also by their distinctive form of execution. In turn, the quality of execution is closely tied to the details of action control. By underestimating that expert action control necessarily emerges from a background of direct experiential familiarity, this objection fails to recognize that all skilful abilities, not only sporting skills, have an embodied and dispositional component as they develop thanks to the incremental sophistication of the agent-environment adaptive coupling. This coupling establishes the expert's inclination to produce richer, more accurate and faster responses to specific perceptual solicitations (i.e., environmental affordances, see Gibson, 1979; Rietveld & Kiverstein, 2014; Miyahara, Ransom & Gallagher 2020) and, correlatively, the expert's ability to predictively anticipate the sensory input following either self-generated (behaviour) or environment-generated (circumstances) events (Robertson & Kirchhoff 2020). As the consolidation of adaptive coupling primarily expresses itself through action automatization, all *expert activities must automatize to some degree* and a practical activity that did not automatize at all (if one ever existed) would be unable to fulfill the key requirements of skilful expertise. All authentically skilful activities are abilities of an embodied, embedded, and dispositional nature, as they emerge from the agent's coupling with the environment (Noë, 2005): this applies, to an extent, also to the activities that *do not entirely automatize*, for example calculus, hermeneutical exegesis, or improvisational abilities infused with cultural, symbolic, and aesthetic meanings, like doodling.

The presence of a robust agent-world coupling primarily grounded in perception–action feedback loop mechanisms is thus a universal criterion to evaluate whether performance execution is skilful, and it applies to all domains of expert activity. Embodied know-how always is habitual *insofar as* and *to the extent that* action control automatizes (Dreyfus, 1996, 2007b) and this principle holds even for those kinds of performances that cannot become entirely habitual: in fact, even an imperfect or incomplete know-how ultimately relies on the meta-stable attunement that the embodied agent establishes with the familiar contexts of their performance (Araújo &

Davids, 2011; Bruineberg et al., 2021), revealing the precarious functional autonomy and self-regulatory capabilities of the agent's body (Barandiaran, 2017; Di Paolo & Iizuka, 2008). Whether it is perfectly or imperfectly attuned to the worldly circumstances, the expert's bodily know-how transpires from their characteristic readiness to spontaneously anticipate and flexibly adjust to fluid contingencies. Pace intellectualists about know-how (Stanley, 2011; Stanley & Krakauer, 2013; Stanley & Williamson, 2017) and intellectualists about skill (Fridland, 2017; Montero, 2016; Toner, Montero & Moran 2015; Wu, 2013), habits are the fundamental constituents of this flexibility, not perils to its exertion.

Even the kinds of performances commonly deemed reflective and exquisitely intellectual (like chess playing, poems composition, and mathematical puzzle-solving), if truly skilful, must have a dispositional root, as they rest on core adaptive capabilities that clearly were acquired and refined via agent-environment coupling. This root is the reason why experts can read real-life contingencies as "rich landscapes of affordances" (Rietveld & Kiverstein, 2014), including linguistic and symbolic affordances (Bruineberg et al., 2019; Kiverstein & Rietveld, 2021), being able to confidently navigate such landscapes to deliver the most appropriate and timely embodied and affective responses. The adaptive coupling is manifested by how identifying the action opportunities relevant to a certain skill engages the expert perceptually and affectively, soliciting their online behaviour through multiple fields of attraction and repulsion that motivate or inhibit this or that pattern of action independently of any detached judgment (Dreyfus, 2008, 2012).

5.3 The "efficacy" objection: when contentful knowledge enhances performance outcomes

The second objection defends a belief that, while largely unsubstantiated, is nowadays singularly popular among academic philosophers.

The *Know-that Improves Performance* belief [KIP]: contentful knowledge and judgments (like those conveyed by verbal instructions, exhortations, self-talk, or well-designed action plans) can help agents make the most out of their skills, producing better performance outcomes.

My refutation of KIP is articulated in three steps, aiming to show –contra intellectualism about skill– that no competitive advantage provided by know-that could ever make a performance skilful or expert.

Step 1 (Sect. 5.3.1): the beneficial role played by contentful knowledge in the production of superior performance outcomes is overestimated. The empirical evidence supporting KIP is disputed and limited to particular domains: in some circumstances, the competitive efficacy of habitually controlled actions increases when automatized routines are accompanied by reflective judgments, e.g. strategic indications, execution-guiding instructions, and self-regulation principles, but in other circumstances these judgments decrease the efficacy of habitual actions and put performance at risk.

Step 2 (Sect. 5.3.2): even if performance scientists discovered a general correlation between contentful knowledge and superior performance outcomes, this would

not imply that contentful knowledge is the cause of those outcomes. In particular, even if it was proven that performances accompanied by contentful knowledge tend to produce comparatively superior outcomes, it would still be unclear how representational content could ever play a causally efficacious or semantically relevant role in the control process.

Step 3 (Sect. 5.3.3): even if contentful knowledge was the cause of superior outcomes, that would say nothing about skilful expertise. Even if KIP was validated empirically, it could not disprove Radical Habitualism because it would still be orthogonal to DisC. That is because, in order to assess whether a performance is skilful or not, we need not only criteria to evaluate the *action outcomes* (i.e., the aspects of the performance that make an athlete “score” during a competition) but also, and more importantly, criteria to evaluate whether the *action execution displayed* the distinctive epistemic qualities of skilful control.

5.3.1 Refutation of KIP—Step 1: contentful knowledge is less helpful and more detrimental than usually assumed

Is contentful knowledge beneficial or detrimental to competitive performance? Empirical data do not allow a simple answer. On the one hand, the forms of practical engagement categorized by Dreyfus as “present-at-hand”, which include content-manipulating practices of a representational and symbolic nature, are demanding in terms of consumption of attentional and cognitive resources (working memory, attention, processing power, inferential abilities, etc.). Clark & Toribio (1998) call these processes “representation hungry” because they require a greater cognitive effort. Today, embodied cognitive scientists hotly debate which activities are truly representation-hungry (if any), and what methodology should be used to identify them, as it is now becoming clear that many activities traditionally considered cognitively demanding (“imagination, memory, planning, and language processing”, *ibidem*) can in fact be explained without appealing to content-involving representational processes (Dreyfus, 2002a; Hutto & Myin, 2017; Wheeler, 2010b). Despite this controversy, the great majority of philosophers and psychologists agree that, whichever exactly they are, content-manipulating activities tend to be slower and less fluid when they compete for the available cognitive resources with other content-manipulating processes (Hutto & Sánchez García, 2015; Kiverstein & Rietveld, 2018; Maxwell et al., 2003).

On the other hand, knowledge of facts and maxims may facilitate performance outcomes when optimal control substantially depends on a relatively small number of kinematic and biomechanical constraints and, at the same time, explicit instructions allows the performer to regulate the parameters of their performance in accord with such constraints; for example explicit instructions telling a swimmer how to form non-habitual patterns of movements to increase the efficiency of their strokes (Cappuccio & Ilundáin-Agurruza, 2020); or indicating a batter how to maintain non-spontaneous bodily postures to minimize the response time and precision of their actions (Sutton 2007; Fridland, 2021). It is also often remarked that some transfer of knowledge in a declarative format, prevalently through speech, is involved during training, especially if the trainees need linguistic instructions during the process of skill acquisition (Porter,

Wu, Partridge 2010; Fridland, 2013; Komar et al., 2014; Francesconi & Gallagher, 2019).

While acknowledging all these platitudes, Radical Habitualism also stresses the merely derivative and ancillary function of explicit instructions in the activities of novices and experts alike. Two facts, concerning learning (Sect. 5.3.1.1) and peak performance (Sect. 5.3.1.2), respectively, encourage us to deflate the importance usually attributed to contentful judgments to skilful expertise.

Learning Coaching and preparation methods that emphasize explicit instructions and rules are not the most beneficial form of skill transfer. Research on dispositional reinvestment indicates that teaching models based on declarative communication engender long-term problems to which analogical (e.g., metaphoric and pantomimic) communication are less exposed (Cappuccio et al., 2019a; Hutto & Sánchez García, 2015; Masters, 1992; Masters & Maxwell, 2008; Maxwell et al., 2000, 2003): skill-learning methods building prevalently on explicit models and instructions offer an experientially impoverished type of training practices because reliance on contentful knowledge can delay the acquisition of the appropriate sensitivity to affordances (Abrahamson & Abdu, 2020; Gray, 2014, 2021); also, such training methods represent a common cause of underperformance in pressured-filled situations, because they predispose experts to compulsively recall principles and instructions, thus compelling them to revert to a novice condition that undermines their confidence in their habitual dispositions and, consequently, their fluid control (Masters, 1992; Masters & Maxwell, 2004; Maxwell et al., 2000).

Peak performance Contentful judgment is often detrimental to peak activity and performance optimization. Advices and instructions given by coaches prevalently promote an internal focus in the practitioner (Porter, Wu & Partridge 2010), soliciting them to voluntarily or involuntarily redirect their attention toward components and processes of their performance that have already been automatized (Baumeister, 1984; Cappuccio et al., 2019a, 2019b). This shift from the proximal goals of the performance to the procedural aspects of the execution disrupts the optimal allocation of attentional resources, as demonstrated by several experimental studies on attentional focus (Gray & Cañal-Bruland, 2015; Wulf, 2013, 2015; Wulf et al., 2001). Also, anxiogenic situations tend to trigger a counterproductive tendency to self-representation and introspective thoughts that are prevalently verbal (hence, contentful) in nature (Beilock & Gray, 2012; Castaneda & Gray, 2007): this tendency disrupts performance not only because it overloads the agent's mind with excessive informational content but also and primarily because the analysis of action promoted by reflective thoughts about the agent's own performance hinders the habitual control of the automatised routines that instantiate that very performance, discouraging the agent from confidently relying on habitual dispositions and thus making their action less fluid and adaptive (Beilock & Carr, 2005).

Now, not all contentful judgments and instructions incite self-awareness, and self-awareness may or may not have the assertive (thetic) force of a contentful judgment (Dow, 2017). Therefore, in principle, reflective body awareness and representational content are dissociable like two orthogonal kinds of processes, as correctly remarked

by Montero (2010, 2014), Toner, Montero & Moran (2016), and Bermúdez (2017), among others. However, in practice, these processes tend to occur together and recall each another, promoting reflective over pre-reflective control (Cappuccio et al., 2019b): contentful instructions that single-out specific body parts and component movements predispose the agent to prioritize self-awareness over environmental solicitations (Beilock, 2011; Beilock et al., 2004; Ilundáin-Agurruza, 2016); conversely, spontaneous introspective self-focus incentivizes the recollection and use of contentful heuristics over habitual dispositions (Gray, 2004; Komar et al., 2014). This is confirmed empirically by an “impressive” (Christensen, 2019) array of bodies of evidence: first, distractive interventions aiming at reducing self-directed attention are effective in reducing performance disruption, because they prevent the practitioner from compulsively recalling rules and other verbal thoughts due to anxiety (Beilock et al., 2008; Cappuccio et al., 2019a; Carr, 2015); second, only novices, not experts, are facilitated by explicit instructions about the posture and execution, while experts tend to perform better in conditions that facilitate their pre-reflective engagement (Beilock et al., 2002a, 2002b; Beilock et al., 2002a, 2002b); third, poor reflective judgments (e.g., when an expert is asked to explicitly assess the situation in which they have to perform) do not predict the impairment of habitual competences (experts who judge poorly can still perform familiar actions effectively), while the automatization of sensorimotor routines predict more accurate judgment in the relevant domains (the formation of habitual competences facilitates better judgments, see Beilock & Holt, 2007).

In summary, while it is well-established that contentful knowledge can interfere with expert control in several tasks and add a cognitive burden to control, it is still unclear whether a correlation exists between contentful knowledge and superior performance outcomes, and, if such a correlation exists, it is certainly not universal, but situational and limited to particular solutions adopted in specific domains of activity, mostly instrumentally to enhance performance outcomes.

5.3.2 Refutation of KIP—Step 2: contentful knowledge is impotent and irrelevant

Step 1 has questioned the possibility of a universal correlation between contentful knowledge and superior performance outcomes. Step 2 will now specify that even if such correlation existed, uncritically assuming that contentful knowledge is what produces these benefits would be a problematic simplification, for at least three reasons: first, without appropriate task manipulations and counterfactual evidence, it is impossible to ascertain what causal role content plays in the improvement of performance outcomes (if any), see Cappuccio et al. (2019a); second, there might be a trade-off between advantages and disadvantages engendered by contentful knowledge in specific activities that structurally rely on propositional content (e.g., the outcome of the same swimming performance may be improved by some explicit heuristics employed by the swimmer to optimize the position of their arms and, at once, damaged by the reflective self-observation triggered by the very same instructions, see Cappuccio, 2017 and Cappuccio & Ilundáin-Agurruza, 2020); and third, because the advantages and disadvantages brought by contentful knowledge can only be assessed comparing them with the advantages and disadvantages brought by training practices that use little or no contentful knowledge at all. Two facts suggests that, if know-that

has any beneficial role at all in a skilful performance, that role always presupposes, and actually is parasitic on, the concomitant role played by know-how in the same performance (Cappuccio & Wheeler, 2010a, 2010b, 2012).

Verbal communication is grounded in non-verbal practices The contentful dimension of speech is sustained by the rich background of bodily and affective interactions in which it is embedded (Kiverstein & Rietveld, 2021): illustrative actions, tool use demonstrations, iconic and indicative gestures, postures, gaze, facial expressions, and proxemics. The acquisition of know-how requires non-verbal information because skill transfer demands communication to be filled with a concrete, situated, operational meaning which derives from examples and exercises, not from the linguistic domain: even when a trainer can accurately describe the objectives that the trainees must achieve, those descriptions can operate causally (i.e., in a functionally efficacious manner) only when they are finally embedded in a background of bodily actions that not only *illustrate* but also concretely *instantiate* those objectives (Abrahamson & Abdu, 2020; Abrahamson et al., 2020). The comprehension of the content of rules and instructions always already presupposes the pre-comprehension of the contentless background of affordances (Bruineberg et al., 2019) to which such rules and instructions refer (Holt & Beilock, 2006) and it is impossible without it (Dreyfus, 1999, 2007b; McManus, 2008).

The severity of skill disruption grows with the amount of representational content leveraged by rules and instructions Although recalling the rules learned during training generally interferes with skilful control, some explicit instructions are less detrimental to performance than others. For example, external focus and speed instructions can often be objectively beneficial to performance, and anyway less detrimental than internal focus and execution focus instructions, precisely because, while being conveyed in a reflective (verbal) modality, they promote the pre-reflective regulation of action (i.e., they bring the agent's attention closer to its goal, away from its component processes, see Wulf et al., 2001, 2002; Wulf & Su, 2007; Komar et al., 2014). Thus, they ultimately facilitate absorption in the holistic flux of lived activity, instead of interrupting it with an effort to analyse action for singling out this or that mechanism (Ilundáin-Agurruza, 2016). The reason external focus (Wulf, 2015) and speed instructions (Beilock et al., 2004) are comparatively beneficial is that, while leveraging a minimally reflective process, they prevent action control from becoming fully reflective, thus reducing the overall dependency on explicit knowledge about single action components (Cappuccio et al., 2019a).

That is why we must differentiate two possible functions for contentful representations in performance, considering whether rules and instructions are meant to leverage and support pre-existing habits or rather offer a systematic procedure to act and decide that is completely abstracted from and substitutive of the agent-environment adaptive coupling: in the first case, contentful know-that essentially works as a *trigger* (enabler), *catalyzer* (facilitator), or *enhancer* (multiplier) for the agent's embodied habitual dispositions; in the second, it has instead the function to replace those very dispositions, substituting them with a protocol that is externally-sourced because provided to, rather than developed by, the performer.

Either way, the contentful knowledge conveyed by representations cannot surrogate the know-how embedded in the underlying background of pre-reflective dispositions or operate without such a background, by reason of the *causal impotency* and *semantic irrelevance* of representations, that I will discuss in Sect. 5.3.2.3 and Sect. 5.3.2.4, respectively.

Causal impotency problem Representational content cannot exert any causal power over action, as such power can be exerted only by psychophysical processes, not purely formal (i.e., logical, semantic, or syntactic) ones. Explaining how the content of mental representations could engender physical action represents a challenge to philosophy of mind (Adams & Aizawa, 2017). Postulating intermediate cognitive layers between content and action indefinitely delays and pointlessly complicates the treatment of this problem without offering a solution to it. Several philosophers assert the causal impotency of representations because semantic relations ultimately derive from causal efficacy, not viceversa (Chemero, 2009; Hutto, 2013; Kiverstein & Rietveld, 2015). What concretely determines behaviour is the sensorimotor and affective responsiveness embedded in the underlying dispositional mechanisms (muscular, nervous, hormonal systems, etc.). As skilful control is a natural variety of causal power, contentful instructions and rules cannot be its legitimate source. Talks of contentful instructions soliciting and shaping skilful control are essentially metaphorical, as the causal efficacy they portray is merely epiphenomenal or ultimately reducible to dispositions (Hutto & Myin, 2017). In short, know-that merely stands in for the real efficacy of embodied know-how and the advantage allegedly brought by it is either vicarious in nature or entirely apparent. Either way, know-that cannot make an agent perform like an expert because content can only *inform about* the natural world (describing it), not *inform it* (by constituting or changing it).

Certainly, contentful knowledge can boost performance outcomes by offering cognitive scaffolds or epistemic prostheses that ultimately are either parasitic on or reducible to embodied habitual dispositions (Bäckström Gustafsson 2017). However, like training wheels attached to your child's bicycle, these representational scaffolds cannot be considered a legitimate cause of skilful performance until their function is effectively incorporated by the performer as a set of habits... at which point they are not contentful representations anymore.

Semantic irrelevance problem An agent can attribute an operational meaning to representational content only if they already possess a performative familiarity with the concrete details of the contexts of action described by such representations (Dreyfus, 2000). Contentful knowledge may happen to be relevant to the execution of this or that action, but it cannot specify all the circumstances in which it is or it is not relevant (Dreyfus, 1999). In this sense, it is fundamentally different from the contentless know-how that we embody when we unconsciously produce fast ocular responses or habitual gestures, a knowledge that is entirely self-sufficient because immediately tailored to the context. In turn, contentful rules and instructions can help us navigate a context of action only insofar as their practical relevance is specified by underlying habitual dispositions (Cappuccio & Kiverstein & Rietveld, 2018; Wheeler, 2010a,

2010b, 2012). When an agent becomes able to perform skilfully without being explicitly informed about the contextual conditions (e.g., descriptions of when and how to act) it is because their know-that has already been replaced by implicit know-how (sensitivity to background conditions). So, even when effectively applied to improve or correct one's *know-that does never become one with that ability*: content is not embedded into the performer's skillset as such.

Some authors have argued that practical knowledge encoded in "executable action concepts" is simultaneously or diachronically accessible by multiple cognitive systems to produce either propositional representation or action, based on the needs of the moment, hence its function is to mediate between contentless and contentful levels of cognitive processing (Ferretti & Zipoli Caiani, 2021). Ironically, this hypothesis undermines representationalism, instead of corroborating it, because it requires that some implausibly rich and detailed concepts are incessantly formed, stored, and manipulated to discretely represent the incalculable varieties of situated contingencies in which the agent is immersed. Our experience of skilful control reveals, in contrast, the simple and direct nature of our response to these contingencies, a response that does not display any sign of the cognitive effort that would be necessary to process them as contentful representations. Even if the format of the relevant concepts were actional (as suggested by the proponents), this would not make their content less discrete and finite, hence it would not make them more adequate to the infinite complexity of real contexts. The content of rules and instructions is never concretely meaningful per se but can operate efficaciously as a proxy of practical dispositions only if the agent is already experientially acquainted with the pre-reflective background of performative know-how.

These considerations recall and expand Ryle's (1949/2009) Crucial Objection against Intellectualism about Know-How [CROKH] (see Pavese, 2021). Stanley (2011) summarizes CROKH in the following way: if know-how were entirely reducible to propositional know-that, as intellectualism maintains, then Premise 1 and Premise 2 should be simultaneously true:

Premise 1: "The Intellectualist view entails that 'for any operation to be intelligently executed', there must be a prior consideration of a proposition".

Premise 2: "The consideration of propositions is itself an operation the execution of which can be more or less intelligent, less or more stupid".

But this would imply that "acting intelligently requires the performance of an infinite number of prior actions, which is a vicious regress" (Stanley, 2011, p. 214).¹³ A proposition could be a valid guide to expert performance only if it exhaustively specified how the relevant action is to be correctly executed, but applying the proposition to the context is a skilful action itself, an action that can be executed correctly

¹³ "The Intellectualist is committed to the view that each intelligent action is preceded by a prior action of considering a proposition. Since considering a proposition is something that can be done intelligently or stupidly, it is an intelligent action. So, acting intelligently requires a prior action of considering a proposition, and considering a proposition intelligently requires a prior action of considering a proposition intelligently. Presumably, if any of these prior actions is performed stupidly, then the original action will not be performed intelligently. But then acting intelligently requires the performance of an infinite number of prior actions, which is a vicious regress" (Stanley 2011, p. 215).

only if another proposition exhaustively specifies how this is to be done. To skilfully implement the instructions provided by the first proposition, some supplementary propositional knowledge is necessary, which in turn requires the correct implementation of the instructions provided by other propositions, and so on ad libitum. No finite amount of know-that per se could exhaustively specify the correct modality of its own application. CROKH implies that intellectualism about know-how is false because the skilful application of any bit of know-how requires the prior manipulation of an infinite amount of know-that.

Intellectualists about know-how have proposed a number of theoretical solutions to break the regress highlighted by CROKH (Stanley & Williamson 2017; Pavese, 2021) but they have yet to address the puzzle in its most general and fundamental form. We find this general formulation in Dreyfus, where the infinite regress, instead of being caused by the underspecified nature of linguistic denotation, arises from the inherent finitude and hermeneutical situatedness of representational knowledge itself. In fact, while formulated in a different domain, Ryle's diagnosis closely resembles Dreyfus' phenomenological and existential analyses of the *Frame Problem*, which point to an equally debilitating form of infinite regress (Dreyfus & Dreyfus, 1988; Dreyfus, 2008, 2012, see also Dennett, 1984).

For Dreyfus the problem is that the finite informational content carried by a finite set of representations would never be sufficient to correctly and exhaustively denote a horizon of concrete real-world circumstances, because any attempt to represent such a horizon that is not infinitely complex like the horizon itself is inevitably incomplete and thus meaningless (Andler, 2000; McManus, 2008; Fusche Moe, 2004; Wheeler, 2008; Cappuccio & Wheeler, 2010a, 2010b, 2012). The information conveyed by representations (of which propositions are just a subset) could help an agent meaningfully navigate a practical context only if such representations (A) were connected to a non-representational background of pre-reflective familiarity that enabled their hermeneutical pre-comprehension or (B) exhaustively portrayed all the defining aspects of all the relevant practical contexts in which that information is applicable. A is exactly the option that intellectualism excludes. The problem with the remaining option (B) is that distinguishing the aspects characterizing the relevant contexts from the aspects characterizing the irrelevant ones requires meta-contextual information (i.e. more contentful representations), which again cannot specify all the preconditions of its contextual application. Thus, the correct interpretation of any representation requires the prior interpretation of infinitely nested levels of meta-representations.

This is particularly evident when representations are meant to convey rules and instructional maxims as guides to skilful performance.¹⁴ Unlike artificial toy-worlds (simulations based on representations characterised by a finite number of parameters),

¹⁴ "Naturally, we can exchange the skills of language with the skills of soccer and reach the same end result as Dreyfus has done. The argument will then go like this: If we are to establish a full-blown theory of the skills of a soccer player, we not only must have rules for how the player shall kick and run, but we also need a specification of the rules that makes the player able to recognize the contexts of where the rules are applied. And with the constant changes of the game, we are talking about many rules. The cognitivist must therefore be able to specify an enormous number of rules that make sure that the soccer player is capable of recognizing the different situations and so on. But if the theory then requires even more rules to be able to explain how the former rules are applied, as the pure cognitivist must do according to Dreyfus, then we are in an infinite regress of rules required for applying rules." (Fusche Moe 2005, p. 169).

real-life scenarios are characterised by an incalculable number of finely grained practical contingencies, so massively interconnected that even a minor variation in their global configuration may generate systemic effects capable to impact the whole performance (Dreyfus, 2008; Wheeler, 2008). Rules and instructions cannot be a valid guide to skilful control because the informational content they carry is inherently underdetermined to specify how the agent must negotiate all the possible contextual contingencies: it would never be sufficiently comprehensive and granular to inform the kind of on-line agile and flexible decisions required by the circumstances. Subsequently, the practical indications generated by explicit commands and rules cannot be fluid and adaptive like those spontaneously arising from embodied, habitual dispositions. Without the adaptiveness distinctively granted by habitual dealings with the action context, an agent cannot effectively implement general rules to particular actions, which remain rigid, mechanical, and “out-of-context” (Ramírez-Vizcaya & Froese, 2020).

Both Dreyfus and Ryle point to the epistemic constraints of representational cognition. Causal impotency and semantic irrelevance are consequences of the fact that a representation, regardless of how rich, cannot provide real-world grounding to the information it carries, because it remains causally decoupled from the dynamical context that makes them relevant. If contentful representations are unable to specify in which contexts they are relevant/useful it is because, without a background of pre-comprehension, they are just static, discrete, self-contained units of information that do not co-vary with the context to which they are supposed to be relevant or the cognitive systems that makes sense of this information, which in turn are dynamic, continuous, and expansive like other natural processes (Hutto, 2012). That is why the epistemic quality of representations is incompatible with the skilful processes they are supposed to guide. Habitual dispositions, on the contrary, materially originate from and are spatio-temporally grounded in the infinitely complex details of the very adaptive coupling that makes expert behavioural responses causally efficient and, at once, inherently relevant to the varying contexts of activity.

5.3.3 Refutation of KIP—Step 3: better outcomes do not imply greater expertise

So far, we have questioned the capability of contentful knowledge to enhance performance outcomes. In truth, even if contentful knowledge played a relevant causal role in making an agent more efficacious in a competitive context, that competitive advantage would still tell nothing about whether their performance was executed skilfully or not.

Assessing a performance based on its outcome serves to produce a comparative judgment about its efficacy. An assessment exclusively based on outcomes reflects a cognitive bias similar to *selective abstraction* (Maric et al., 2011): a particular aspect or feature of a performance (e.g., the number of goals scored during a match, the height of a jump) is singled out to become “the outcome” of the performance, i.e. the result of an action conceptually separated from the quality and form of its execution. In many domains of expertise (e.g., who is the most skilled painter? The most skilled coder?), this process of abstraction is largely arbitrary in the sense that it depends on conventions and decisions established by communities of humans within their

shared cultural practices (Cappuccio & Ilundáin-Agurruza, 2020; Ilundáin-Agurruza, 2016). An abstractive process of this kind may be to an extent unavoidable, but its arbitrariness often risks confounding the recognition of authentic expertise: even when empirically accurate, a value judgment on skill *exclusively* based on outcomes de facto reduces a performance to an instrumental action, contradicting the very definition of performance and thus preventing the broader appreciation of the particular contextual contingencies in response of which performances are produced.

Although sport business puts brute results under the spotlight and idolizes competition winners, it is clear to experienced coaches and club managers that the most skilled player in their team is not necessarily the one who has scored more points in the season, as there are myriads of other qualities to be considered in a comparative evaluation of their overall performance. Often, in the professional and artistic domains, the value of outcomes is not clearly defined by objective measurements systems, thus assigning an objective “score” to a performer proves even more difficult, as the correct evaluation of their performance involves situated normative principles that are qualitative and value-laden in character (Rietveld, 2008). Unless these principles are deeply incorporated in our immediate ways of feeling and perceiving, performance comparative assessment is an inherently complex, possibly endless, endeavour. Even in contexts characterised by objectively measurable and perfectly quantifiable outcomes (i.e., standardized, controlled environments such as the sporting field or the experimental laboratory), performance assessment is never exhausted by the measurement of this or that variable alone, as it also depends on the interpretation of their particular contextual significance.

The inadequacy of outcome-centric assessment methods is illustrated by the possibility to artificially enhance or reduce performance outcomes without affecting the agent’s skill level. Let us consider three categories of *expedients used to enhance performance outcomes*. Under the first category we find expedients like muscle development routines or meditation and relaxation practices for self-regulation and emotional control. These expedients, if proven safe for the athletes, are accepted by the sport community because they enhance the same kinds of abilities that normally underpin skill. By enhancing the abilities that univocally and specifically constitute the foundation of a certain skill, these expedients are considered legitimate aids to support that skill’s development.

The second category includes expedients such as chemical boosts or cybernetic augmentation. They are forbidden in agonistic activities not only because they can be harmful and violate the athlete’s psychophysical integrity but also because they violate the principles of sportive competition, as they artificially enhance performance outcomes without promoting legitimate skill improvement.

Our evaluation becomes more difficult with the third category of expedients, which includes *auxiliary skillsets* that are acquired with the intent to enhance the outcomes of the skillset immediately relevant to competition. An auxiliary skillset ψ is an expedient to support their *performance skillset proper* φ : for example, a professional golfer may decide to learn meditation techniques to improve his golf putting performance. The fact that ψ can be instrumentally exploited to optimise/enhance the outcome of their φ -performance indicates their expertise in ψ , not φ . In turn, expertise in φ per se does not entail skill ψ , unless ψ clearly is a structural component of φ : this would clearly be

the case only if ψ -expertise were (1) systematically required in all φ -performances and (2) specifically tailored to enhance φ -performances. These two circumstances rarely occur and anyway almost never together: golfers can perfectly put without meditation techniques, which, in turn, are not specifically designed to improve golf putting. This means that the efficacy of these auxiliary skillsets is not constitutive of the performer's expertise: in fact, the magnitude of the competitive advantage that they provide is pretty much orthogonal to the experiential history of the agent. Thus, a φ -performance is not more skilful because its outcome was enhanced by an expedient ψ if the nature of ψ is merely instrumental to increase the outcomes of φ or if ψ was acquired independently of φ , as either circumstance contradicts the definition of skilful performance.

Only the expedients in the first category contribute to increase one's expertise, because –unlike the other two– they are legitimate constituents of the process of skill development. The expedients listed in the third category, which use auxiliary knowledge to improve one's performance, are more similar to the expedients of the second category in the sense that they enhance performance outcomes without increasing the know-how embedded in one's inherent control capability. This confirms my point that success is not a valid indicator of one's skilful expertise.

The same reasoning applies, in reverse, to the disruptive factors that typically cause a performance to fail. Some cases of underperformance are engendered by conditions (e.g., physical debilitation or contingent distractors) that undermine the agent's efficacy in a competition without reducing their habitual control. A poor performance outcome engendered by one of these objective extrinsic factors does not give reasons to doubt that the agent performed like an expert. Remarkably, this reasoning does not apply to other debilitating factors, like performance anxiety, which affect experts, and experts only, by interfering with their habitual control. In these cases, underperformance is caused exactly by the agent's incapability to perform "like an expert".

Choke, a psychological effect particularly feared by expert athletes and performers, is effectively explained as the subconscious tendency to reinvest attention into the explicit monitoring of action execution (Baumeister, 1984; Beilock, 2008, 2011; Hill et al., 2010). Skilful expertise deteriorates when habitual control is disrupted by self-focus. That is why leading sport psychologists construe choking as a *debilitating paralysis* of the automatized behaviours caused by compulsive self-monitoring, either spontaneously produced or promoted by explicit instructions (Beilock & Carr, 2005; Castaneda & Gray 2007; Mesagno & Hill, 2013). The sufferers of choking are diagnosed with a *temporarily loss of their normal ability to perform as experts* exactly because they involuntarily revert to a novice condition as soon as they start to compulsively monitor their own actions (Cappuccio et al., 2019a, 2019b; Jackson et al., 2006; Masters & Maxwell, 2008).

Of course, experts tend to be more successful than novices: skilful expertise (S) goes hand in hand with outstanding outcomes (O) and we expect to find them together. Radical Habitualism does not deny that an inductive correlation between S and O exists but reminds us that it goes only in one direction (the implication $S \rightarrow O$ does not justify the converse implication $O \rightarrow S$); moreover, correlation does not mean causation; causation is only one aspect of constitution; and constitutive accounts (i.e., essentialist accounts that characterise x in terms of x 's ontological components) are not more informative than explanatory ones (i.e., processual accounts that describe x

in terms of how x functions and what it concretely does). Even the fact that a non-lucky agent can compete with experts, reliably producing performances comparable to those generally expected from skilled performers, is insufficient to prove that they truly have the expertise: a junior boxer with limited experience on the ring and poor technique is likely to win several matches if he is naturally gifted with an extraordinarily robust built and lightning-fast fists. The evidence that experts frequently produce outstanding outcomes is per se insufficient to infer that any agent producing outstanding outcomes truly is an expert. To ensure that O is a reliable and accurate indicator of S we need an account that satisfactorily explains their causal relation.

Performance measurements collected in highly controlled environments can, in principle, justify an inference to the best explanation from O to S , which is why such measurements are conducted in laboratories and athletic competitions alike: by filtering out all accidental factors (including random circumstances and the agent's natural "gifts"), they artificially construct a link between outcomes and putative skill level. Such a selective construction, however, offers only a narrow picture of skilful expertise because the information it provides is less ecologically valid than the information provided by a more comprehensive assessment (e.g., it ignores that exploiting accidental circumstances and abilities is a prerogative of experts, see Davids et al., 2012).

"Skill is relative rather than absolute" warned us Logan (1985), referring to the fact that a performance can be considered more or less skilful only in relation to a particular context, or world-environment, to which their skill is originally attuned. "We may talk about skilled performers versus unskilled performers, but we immediately recognize that this is only a convenient fiction. There is no sharp demarcation between the skilled and the unskilled" at least not if we assume that a skill is a simple, objective feature of an agent or if we evaluate a skilful performance simply looking at its outcomes abstracted from the modality of its execution.

However, this does not mean that expertise level is an illusion or that it is an indiscernible property. We can and we normally do recognize an expert boxer because, in addition to scoring many victories, he demonstrates an elegantly balanced stance, an impenetrable defence, perfectly timed moves, and a confident use of fancy footwork (inter alia). *While being situated and relational, these features tell us exactly why* he is very often the one who prevails on the ring, thus – by providing an accurate description of how the performance was executed, these features offer also an explanation of his outstanding outcomes.

Obviously, the boxer who can count on such a bodily know-how tends to be also the one who wins his matches more frequently. After all, their know-how was expressly developed through practice for winning matches. But the point is that even an unimpressive or common outcome on the ring may be perfectly skilful, if the boxer performs a complex sensorimotor task with flawless perceptual, motoric, and emotional readiness, dealing effectively with an adversary that most boxers would consider dangerous. The skilful quality of an action shines through its form of execution, revealing how confidently the boxer's actions are regulated and how harmoniously they are organized. Conversely, a performance cannot be considered skilful, despite its impressive outcome, if its execution is poor or does not entirely originate from the agent (i.e., when it depends on supplementary aids).

This three-step refutation of KIP definitively dismisses the Efficacy Objection: the competitive advantage brought by contentful knowledge is situational, modest, uncertain, or merely apparent, and anyway orthogonal to skill level because insufficient to explain successful performances.

5.4 The “modality” objection: when execution is informed by representation

The line of argument used to block the Efficacy Objection cannot work against the third objection, which asserts that mastering certain content-involving practices distinctively characterizes the expert’s performance’s mode of execution, independently of its outcomes.

The Modality Objection correctly acknowledges that skill transpires from the form of execution, but wrongly assumes that contentful knowledge contributes to make the execution of a performance more skilled. This objection misinterprets some empirical facts broadly confirmed by experimentation and the practitioners’ reports: experts, compared to novices, are capable to give more accurate judgments and richer verbal accounts of their performance (Sect. 5.4.1); experts, more frequently than novices, are able to deal creatively and improvisationally with difficult or unprecedented problems (Sect. 5.4.2); experts, more competently than novices, strategize and make complex decisions during performance (Sect. 5.4.3).

The objection assumes that, because verbal reporting, creative improvisation, and strategic decision-making predict higher skill levels, they must essentially define the distinctive way experts perform (Fridland, 2021). So, Objection #3 argues that these cases undermine Radical Habitualism because they (a) indicate the agent performed like an expert and (b) their performance involved reflective, contentful judgements and deliberative processes that cannot be automatized. I shall question these two points, arguing that verbal reporting, strategic decision-making, and creative improvisation either (a) are insufficient to prove that the agent performed like an expert or (b) can be automatized, or both.

5.4.1 Regarding rich accounts of ongoing or past performances

One of the main arguments against intellectualism about know-how is that the ability to execute certain actions and the ability to report on them seem totally unrelated: experts often struggle to describe “how they did it”, while many excellent coaches are not able to perform the actions they effectively describe to their trainees. Nonetheless, intellectualists find it remarkable that some expert performers tend to indulge in copious efforts of recollection and verbalisation of their prior actions. Moreover, these performers’ potential to generate rich verbal accounts after performance seems to increase with the agent’s expertise level. This is plausible. However, a greater potential of this kind does not predict more skilful performances. The loquacity of experts is largely explained by the well-known fact that people tend to have a broad (but not always deep or reliable) factual knowledge in their areas of personal or professional interest, as they hold stronger and more articulated (but not always accurate or insightful) opinions about it (Logan, 1985).

Obviously, retrospectively analysing one's own performances during the debriefing phase can help a professional performer identify mistakes and areas of improvement, thus increasing the effectiveness of their successive training. However, there is no evidence that performances accompanied by accurate verbal accounts are executed more skilfully. On the contrary, experiments show that, when they perform skilfully, experts, compared to novices, tend to generate fewer explicit judgments about their ongoing performance (Beilock et al., 2002b). Also, offline verbalization about a new motor experience tends to damage subsequent performances in novices (Chauvel et al., 2013). Last but not least, experts, compared to novices, recollect specific episodes or aspects of their performance with lower accuracy and granularity. This well-documented phenomenon is called "expertise-induced amnesia" (Beilock & Carr, 2004, 2005): skilled performers can produce rich accounts concerning the extra-personal dynamics underpinning execution (e.g., the scientific knowledge of the metabolic processes relevant to energy-use maximization) or the personal dynamics that oversight execution (e.g., the strategic knowledge relevant to effective action planning and rational decision making), but their verbal recollection of the details of specific instances of performance execution (episodic memory) tends to be, compared to novices, vague and impressionistic (Beilock & Carr, 2004; Beilock et al., 2002a, 2002b).

This tendency is explained by studies on attentional focus (Wulf, 2015) and embodied perception (Gray & Cañal-Bruland, 2015): the attention of experts, during peak performance, spontaneously tends to target their immediate intended effects, not the components of their movement: their habitual engagement disinclines (but does not mechanically prevent) them from closely monitoring their ongoing actions, thus reducing the amount of contentful information they can extract and encode from their ongoing execution (Cappuccio et al., 2019a, 2019b; Gray, 2004).

That is because reflective self-monitoring tends to disrupt expert performance, as testified by the fact that catastrophic performance deterioration correlates to both compulsive self-monitoring and pressure-filled environments. Gray (2004) indicates that, in situations of psychological distress, the attention of experts tends to target the procedural components of their movements, deteriorating the execution of their habitual actions while improving the accuracy of their judgments about those very actions. The cognitive effort required to report on one's own actions, making concurrent judgments, is part of the process that deteriorates their execution (Flegat & Anderson, 2008). This is enough to reject the claim that verbal processes are a defining, positive feature of the skilful control of experts.

5.4.2 Regarding the ability to act creatively and improvisationally

The fact that experts are more confident and effective than novices in exerting their creativity during peak performance could undermine DisC only if creatively improvised action was obviously unhabitual in nature. However, the idea that creativity and imagination necessarily involve contentful knowledge is disputed (Hutto & Myin, 2017), while there are many reasons to understand the imaginative and visualisation processes

behind improvised performance as non-representational in nature (Rucińska & Aggerholm, 2019) and fundamentally based on embodied habitual dispositions (Beilock & Gonso, 2008).

Besides, our previous considerations on unready-to-hand forms of interaction indicate that the DisC would still be valid even if the creative and improvised control of an expert proved completely unhabitual: the transient formation and situational application of action-oriented representations as temporary scaffolds to deal with unfamiliar circumstances can be facilitated by richer and opportunely diversified constellations of habitual dispositions. This explains why expertise level positively contributes to the process that brings an expert to envision and apply minimally representational forms of control in troubled scenarios (Baggio, 2021; Cappuccio & Wheeler, 2012; Wheeler, 2005). Therefore, whether or not action-oriented representations constitute truly contentful cognitive devices, their performative efficacy is ultimately to be credited to pre-reflective processes (Hutto, 2013). This means that even unready-to-hand action is truly skilful only to the extent that it is controlled pre-reflectively, i.e. when habitual dispositions played an active role in improvising a coping strategy, and it is unskilled (or progressively less skilled) commensurately to how such strategy was not supported by any habitual dispositions.

That is because, when imaginative abilities are sustained by decoupled, contentful processes, they must rely on auxiliary and expediential sources of knowledge that are essentially different from the expertise deriving from historical agent-environment coupling, thus such abilities cannot have the characteristic epistemic qualities (and constraints) of skilful control (Rucińska & Gallagher, 2021).

5.4.3 Regarding decision-making and strategic reflection

Dreyfus' account of skill levels emphasizes that experts tend to have a superior awareness of the bigger picture, which helps them plan their actions with a strategic perspective, possibly a long-term one. Papineau (2015) remarks that experts prefer planning their purposeful activity in accord with holistic styles and distal goals. Strategic reasoning is often constructed as the high-level decisional process that organizes motor intentions subordinating them to distal and more general intentions (Fridland, 2021). The objection emphasizes not only that representations can be used to strategize, but also that they are used more frequently, effectively, and confidently by experts than novices (Christensen et al., 2016): this is used as evidence that contentful knowledge can play a constitutive role in the control of expert performance (Gallagher, 2018; Gallagher & Varga, 2020).

This inference is, however, flawed: as in the case of imagination used improvisationally, we must again remark that, while rational $\times$ aiming to build a hierarchy of means and ends is often sustained by reflective awareness and explicit representation (Mylopoulos & Pacherie, 2019), this is neither necessary nor constitutive in many kinds of skilful activities; moreover, strategic reasoning, like any other component of a performance, can be truly reflective without being simultaneously skilful. Once again, we need to distinguish the instances of strategic reasoning that are grounded in pre-reflective control from the reflective ones, while assessing which of them deserve to be called "skilful".

When strategic reasoning skills are dispositional and incorporated in the same habits that control the performance This happens, paradigmatically, when the practical knowledge relevant to strategic decision-making is contextually embedded: know-how solutions are created “on the fly” from patterns of thought and volition spontaneously solicited as part of one’s automatic dispositions (Araújo et al., 2006). This is illustrated by fast chess playing in which complex anticipations of the adversary’s behaviour are generated in a purely habitual manner and the depth of the player’s strategic vision is strictly contingent upon their embodied responsiveness to familiar patterns of spatial organization and correlated activity (Dreyfus, 2002b). Some strategically significant preferences and inclinations have, as a matter of fact, an embodied and affective root determined by sensorimotor familiarity (Beilock & Holt, 2007). In these paradigmatic cases, decision is implicit and its burden is taken on by embodied habitual dispositions, not representation, thus control is pre-reflective (Brownstein).

Pre-reflective decision-making and automatized problem-solving are normally available at the highest levels of expertise, starting from “Competence” level. Dreyfus and Dreyfus (1980, 1988) portray this kind of strategic intuition as an intelligent propension that, in many way, is more similar to a gut reaction or an acquired instinct than to detached and unemotional calculus. While largely pre-reflective, this “deliberative rationality” is often supported by planning practices that can be lengthy and demanding. One way to construe complex rational planning as an embodied, dispositional process is using the notion of embedded or *nested affordances* (Rucińska, 2021), i.e. articulated chains of action possibilities that are perceived in the environment as indicating secondary action possibilities, and secondary action possibilities that indicate tertiary ones, and so on – with the number of embedded levels of action possibilities being limited uniquely by the agent’s perceptual specifications, nor their cognitive resources. Elaborating on similar premises, Cappuccio, Sánchez García & Ilundáin-Agurreza (forthcoming) have developed a general pre-reflective non-representational account of strategic decision and skilful planning inspired by the ancient concept of “Metis”, described by the Greek as a cunning, crafty, resourceful, and often deceptive form of embodied intelligence and epitomized, for example, by Ulysses in the Homeric poems and, in more recent times, by Captain Kirk in Star Trek. Metis is a dispositional quality deeply related to the classical virtue of *practical wisdom*, that Dreyfus conceives as the equivalent of skilful coping in the moral domain (Gehrman, 2016).

When strategic reasoning skills are representational and rely on contentful knowledge Know-that ψ is used by the agent expedientially, to enhance their performance ϕ . As such, ψ is decoupled from the context of performance ϕ itself, which is why it often is processed asynchronously from online performance control. In this case, the expediential knowledge is based on contentful rules, models or instructions, and its contribution to the control of performance ϕ is reflective. However, as discussed earlier, the knowledge informing the control of ϕ cannot be considered part of ϕ if it does not belong to the agent in an intimately familiar manner (Sect. 5.3.3) and does not display the three epistemic qualities of skilful control (Sect. 3.3). A sport performance cannot be controlled in a familiar manner if the agent relies on know-that about principles and facts established by academic disciplines such as performance science,

kinesiology, logistics, cognitive neurosciences, metabolic medicine and biochemistry, or any other discipline that studies the skills of experts without presupposing direct practice or exercise. By extension, if the strategic reasoning skill ψ is based on contentful knowledge acquired independently of skillset φ then, at best, ψ proves the performer's ability to effectively extract some instructional maxims or notions that are relevant to performance φ . However, this only proves the performer's erudition and their ability to apply such erudition to a practical domain φ like an advanced beginner, not the performer's competency in the φ skillset.

To summarize, also control based on strategic decision is unable to undermine Radical Habitualism because (consistently with DisC) such control is either: (i) habitual and deeply rooted in the same skillset used to perform; (ii) reflective because based on contentful knowledge and, thus, unskilled because irrelevant to determine the agent's expertise level. Consequently, it is not true that performers, even the most strategic ones, rely more and more on reflective control as their skill level increase. Moreover, a greater reliance on reflective control does not constitute legitimate evidence that reflection plays a greater control function in skilful performance.

5.5 Provisional conclusions: in favour of the Skill-No-Skill approach

Our analysis of the three objections has proven that even intellectual activities have a habitual basis, that performance outcomes are not reliable evidence of expertise, and that reflective practices cannot be constitutive elements of skilful execution. These analyses definitively undermine the Knowledge-No-Knowledge approach to hybrid scenarios because they prove that the performer's ability to effectively combine some auxiliary know-that ψ with the habitual dispositions underpinning performance φ is not a valid indicator of the performer's expertise in φ . The fact that the ability to use ψ increases with the performer's skill level in φ does not prove that representations have a greater role in their skilled performance: on the contrary, it suggests that the habitual dispositions previously acquired by the performer scaffold his abilities to act pre-reflectively and reflectively alike.

So, the Knowledge-No-Knowledge approach is wrong because the agent's ability to combine elements of reflective and unreflective control is ultimately irrelevant to assess the agent's expertise in φ . Conversely, a robust pre-reflective control capability emerging from well-trained habitual dispositions in φ is a key precondition to adequately and adaptively apply the reflective resources provided by ψ . Thus, whether know-that is involved instrumentally in a performance or not, only the agent's habitual dispositions reliably indicate the agent's authentic potential to execute a performance skilfully and moreover they represent a crucial factor for the agent's odds to perform successfully (producing desirable outcomes); conversely, if such know-that is not supported by an adequate background of habitual dispositions, it improves performance only situationally, conditionally, and partially, and certainly cannot make their performance more skilful.

This gives us some strong reasons to prefer the Skill-No-Skill approach over the Knowledge-No-Knowledge approach, rejecting the supposition that a skilful performance is essentially the one that produces the outcomes wished for by the

performer. By systematically neutralizing the three objections against the Skill-No-Skill approach, this section has eventually corroborated the DisC: if the performer's actions are ultimately governed by contentful knowledge, then they are not controlled skilfully. It has also helped us articulate the idea that, when a performer appeals to contentful knowledge to enhance or modulate behaviours that are already under the control of robust habitual dispositions, that performer can act like an expert only to the extent that a pre-reflective attunement with the performance environment sustains their engagement with the task's context.

This result finally confirms that the contribution offered to skilful control by rules and instructional maxims is either epiphenomenal or entirely reducible to the habitual dispositions in which the comprehension of rules and instructional maxims is embedded, because the ability to control a performance *skilfully* (regardless of its outcomes) derives solely from the history of their online embodied engagement. In short, what we have learned is: first, the fact that actions involve mixes of habitual and unhabitual components does not mean that the habitual and the unhabitual contribute to make that action skilled (or unskilled): only the habitual components contribute to qualify an instance of action execution as expert; on the contrary, the non-representational or contentful components (reflective) not only do not always make an action more skilled but are also likely to make it less skilled.

6 An illustration: the BBQ scenario

Let us conclude with a thematic analysis meant to exemplify the concepts reviewed so far. Perfectly grilling steaks on a barbeque is a skilled activity that demands the competent use of several tools and techniques. Contentful indications learned from written or verbal accounts are insufficient to master the tricks of the trade, which must be painstakingly practiced in real-life scenarios until one is entirely confident about their implementation in a range of quite different situations.

This does not mean that the performance of a BBQ master can never include any contentful knowledge at all. Experts continuously recall old information or search for new one, for example when they consult a recipe book to confirm the cooking time of a certain kind of meat. The point is that, while this contentful knowledge may or may not help them complete their task, it certainly does not help them do it *as experts*. The knowledge of recipes and procedures becomes an effective guide to action only once it is incorporated as a set of habitual dispositions, which prepare the cook to recognize and appropriately respond to the relevant environmental solicitations, making grilling their second nature. Every bit of contentful knowledge initially learned from books or verbal instructions must become contentless if it has to be applied skilfully. The cook must incorporate this know-that as a set of pre-reflective dispositions.

For, example both unexperienced and experienced grillers may find it useful to rely on a digital BBQ thermometer to check that a steak has reached the right internal temperature. Determining when the steak is ready is a complex decision that must take into account the thickness of the steak, the meat type, its cut, the fire temperature, the grilling technique, the BBQ specifications, and several other variables. Without the BBQ thermometer's aid and without cutting the meat to check its internal colour,

no novice and only few amateurs can track these variables, which is why non-experts are typically advised to use a thermometer to avoid disappointing outcomes. However, nobody would say that their cooking performance became more skilful because they started relying on such an external aid. The novice's dependency on this tool is exactly what reveals that their control of the grilling process is still uncertain: such dependency betrays their inexperience despite their steaks being perfectly done. What betrays their unpreparedness is not the artificial nature of the technological tool they rely on, but the fact that their decision is entirely delegated to an extrinsic judgment of a contentful nature. We would have equal reasons to consider them inexperienced if the same contentful information originated from a different source (e.g., the advice of another cook in the kitchen, a manual of BBQ techniques, a microchip implanted in their pre-frontal cortex, etc.): in all these cases, control is entrusted to some discrete, self-contained operation entirely independent of the agent's performative history and decoupled from the context in which performance is executed.

Some BBQ experts pride themselves to avoid thermometers because –they say– a true master checks the doneness of the meat simply using their tactile and visual abilities. Other professional chefs systematically use thermometers to minimize the risk of disappointing their diners, especially when they must supervise numerous grills at once: they deem thermometers objectively useful and see no shame in using them. These are just subjective preferences. What crucially distinguishes experts from novices is that, when they are not allowed to use a thermometer, experts have significantly better chances to guess the right cooking time. Also, out of two perfectly done steaks, respectively cooked by two different cooks with and without the help of a thermometer, the steak cooked without the thermometer certainly is –all the other things being equal– the one cooked by the most experienced cook between the two. Also, the cook's expertise is indicated by the cooking style (elegance, parsimony, rapidity, confidence, etc.) no less than its final outcome, although the former might be less apparent than the latter: true skill shows first of all in the fine control exerted by the master, when the distinctive epistemic qualities of expert know-how make a crucial difference.

There is more. Experts and novices don't apply the thermometer with the same effectiveness: using the thermometer appropriately is challenging for a beginner, even after they duly memorized all the applicable instructions, as these instructions (regardless of how detailed they are) are never sufficient to make appropriate decisions in all possible contexts. What is the best way to position the thermometer? What is the best course of action when the thermometer reaches the critical temperature? Abruptly removing the steak from the fire, flipping it to the other side, moving it to a less central area of the grill, or diminishing the heat? These are only some of the many plausible options and each of them, to be implemented correctly, requires dozens of other context-sensitive micro-decisions.

Reading the thermometer's value and applying some generic instruction associated with it is not even the beginning of what is required to appropriately deal with a steak that quickly approaches its ideal doneness: when the meat has reached the recommended internal temperature, it takes an expert to complete all the right operations in the right order with the right care. Novices risk to overcook or undercook their steak even if they blindly follow all the instructions and tips associated with the thermometer,

as they lack the constellations of habits necessary to skilfully apply those instructions and tips while effectively navigating the infinite contingencies of the grilling process. The correct application of the thermometer “by the book” requires only the narrowest skillset, which makes the agent proficient *exclusively in* the “thermometer use” skill (if anything at all), *not in* the “cooking with a thermometer” skill. An inexperienced cook turned into an expert thermometer user would still have only the bare competence strictly sufficient to cook mediocre steaks and avoid the most common mistakes. To be truly skilful, the application of the thermometer must become an integral part of, and immediately cohere with, a galaxy of context-specific micro-practices that is as broad and deep as the art of grilling itself, which is why either the know-that invoked to instruct the use of a thermometer represents the entire skillset of a BBQ master or it is practically useless and possibly even deleterious.

Now, the greatest advantage of thermometers is that they are designed to be a universal source of user-friendly contentful knowledge. They can be used both by experts and novices and, in a sense, they establish a standard that is valid in any circumstances. They certainly tend to reduce the chances to ruin a steak, which represents the best possible outcome for the novice and a desirable minimum for the expert. But the fact thermometers provide a serviceable, advantageous, skill-level-neutral bit of information is exactly what makes the cook’s skill level less relevant: in no way that information, per se, contributes to increase the agent’s know-how; conversely, that information is practically useful and relevant *only insofar as some know-how is already pre-reflectively available* to the cook in the form of a set of embodied habits. Paradoxically, the thermometer becomes a legitimate element of the cook’s skilful activity only once its function becomes redundant: not an informative instrument providing a necessary bit of content but just another element in the system of environmental solicitations to which the dispositional know-how of the agent is coupled; not a critical, discrete decision-point but a perceptual cue within a continuous horizon of possible affordances.

Considering this constraint, should we consider the use of the thermometer an integral component of the grilling skillset or not? Again, the answer has nothing to do with the extrabodily and instrumental nature of tools (which, under the right circumstances, can be considered legitimate extensions of the expert cook’s body schema, i.e. active functional components of an extended cognitive system), but on the modality of control brought about by the thermometer, which can be either unhabitual or habitual. On the one hand, we must note that, unlike the cook’s well-trained perceptual and motoric dispositions, the thermometer is neither necessary, nor sufficient, nor constitutive of the skilful control of grilling:

- First, unlike the novice, the expert can obtain the same result with or without thermometer, as their control of the grilling process is grounded only in their grilling habits;
- Second, unhabitual control might compete with or even confound habitual control, in the short term, and produce deskilling, in the long term, progressively weakening the cook’s ability to perceptually check the doneness.
- Third, a jury tasked to evaluate the cook’s skill level would not be surprised to learn that they rely on a digital thermometer but would not be impressed either. On

the contrary, they would positively evaluate the fact that the cook did not need the thermometer or the fact that the thermometer's use was harmonized with a multitude of relevant habits confidently and effortlessly exerted by the cook.

On the other hand, we must acknowledge that a cook who treats the information provided by the thermometer as a narrow prescriptive indication, rigidly executing the pre-defined script associated with it, can only be a novice as they cannot do much more than blindly follow a set of received rules and instructions; conversely, a cook who, having fully automatized the use of the thermometer, experiences it as a transparent element of their grilling practices and pre-reflectively responds to it as just another environmental solicitation (not different from the colour or the consistency of the meat) must be an expert because they have fully incorporated the tool among the affordances to which their embodied expertise is responsive.

Therefore, the performance with the thermometer can be considered skilful *iff* the cook has already made it habitual, systematically relying on their relevant embodied dispositions to effectively control the grilling process via the thermometer. The thermometer does not play an effective causal role and does not have a concrete practical relevance per se, but only to the extent it is incorporated in the background of habitual practices that pre-reflectively control the cooking process. The representational content attached to the thermometer (the mere logical and informational value of the number displayed by the tool) has no causal efficacy or semantic relevance per se because it is decoupled from its user and the context of its own use, thus either it does not contribute to the grilling performance or its contribution is not of a skilful (positive) nature.

To sum up, the thermometer does not help a novice grill steaks like an expert but can help them prevent many common yet avoidable amateurish disasters: thus, the more the thermometer proves useful the more its user has already acquired some relevant grilling know-how. Using the thermometer effectively does not provide the cook with a supplement of skilful control but, on the contrary, presupposes that very skilful control. Whether the use of the thermometer is skilful or not is not attested by how advantageous that usage is; rather, it is attested by how the grilling performance displayed the characteristic qualities of expert execution. In a way or another, performance is never made skilful by the contentful information provided by the thermometer. As per Radical Habitualism, performance is either skilful or controlled by contentful knowledge, because know-that is never an authentic source of skilful control.

7 Conclusion: the broader implications of Radical Habitualism

Experienced performers obviously *can*, typically *do* and, moreover, often *must* use reflective control informed by contentful knowledge during peak performance. Without denying this, the previous analyses have shown that an agent performs like an expert, and their performance is truly skilful, only if their actions are confidently controlled by well-trained habitual dispositions. We must then conclude not only that the intellectualist maxim “all knowing how is knowing that” (Stanley & Williamson, 2017) is false but also that the opposite maxim is true: “no knowing how is knowing

that.” This conclusion can now be appreciated in all its compelling force as our analyses have systematically shown that the account of skilful expertise offered by Radical Habitualism is *universal, positive, innovative, phenomenologically sound, consistent with naturalistic explanation, and based on empirical evidence*.

First, our investigation releases Dreyfus from the allegation of arbitrarily narrowing-down the standard definition of skilful performance to exclude the cases that contradict his phenomenology. Some authors (e.g., Montero, 2016, p. 53; Christensen et al., 2016, p. 59) allege that the habitualist perspective (like those defended by Dreyfus, in philosophy, and Beilock and Carr, in sport psychology) conveniently relies on canonical descriptions of agents performing routine actions (e.g., car driving in “automatic pilot” mode) or producing serendipitous performances in rare “flow” conditions, thus limiting his analyses to a specific set of well-chosen cases in which explicit knowledge does not play a major role. Against these allegations, my full-spectrum analyses confirm the universality of Radical Habitualism insofar as all known forms and levels of skilful expertise have a fundamentally embodied and dispositional basis. In fact, Dreyfus’ approach promotes an expanded and more comprehensive picture of skilful expertise, not a reduced one, because it invites us to appropriately appreciate the qualities of the execution of a performance in addition to its competitive outcomes and its propositional description.

Second, Dreyfus has been unjustly accused of offering a merely negative account of skill, i.e. limited to describing what expertise *is not*: un-reflective, non-representational, etc. (Stanley & Williamson, 2017). In truth, Radical Habitualism offers a positive account that is both viable and sound: the original attunement of intentional agents with their world-environment constitutes the authentic source of skilful control and the true mark of expertise, which is why the skilful valence of a performance is to be found in the network of dynamic, interactive relations that reciprocally link an agent and their environment (Freeman, 2000). Appreciating the ecological qualities of this network of relations means valuing how a performance is executed confidently, swiftly, adaptively, creatively, parsimoniously and whether the agent demonstrated the characteristic readiness to cope with the situation that distinguishes experts (Davids et al., 2012). Skill is a form of familiarity or propensity arising from concrete dealings. Also, expertise is an organisational, dispositional quality that expresses the agent’s personal history and reflects their robust metastable attunement with the environment (Heras-Escribano, 2021; Miyahara, Ransom & Gallagher 2020; Bruineberg et al., 2021).

Third, while being old at least like Merleau-Ponty’s and Heidegger’s phenomenologies, the general account offered by Radical Habitualism is truly innovative and likely to produce a concrete transformative impact on the practices of skilled experts and the normative criteria used to evaluate their performances: Radical Habitualism invites us to rethink the notion of expertise beyond stereotypes that are still too common in both the agonistic and the academic worlds, where the value of competitive efficacy and the importance of formal titles tend to be overemphasized, in a way that promotes the superficial identification of the skilful expert with the competition winner or the degree holder, respectively. Skilful expertise is too often construed as the conformity of a subject to some essentialist standard:

- (A) as some sort of epistemic authority or curricular status (professional credentials or similar entitlement to have an opinion);
- (B) as an indication of the length of the practice that the agent has done (for example 10,000 h of practice or ten years of experience);
- (C) as the amount of knowledge or information that the agent carries in their mind like a stored commodity or usable resource (facts, data, rules, etc.);
- (D) as a power or ability that the agent can exploit instrumentally to achieve whatever goals.

Radical Habitualism consists in a form of *relationalism about skill* (Coeckelbergh, 2012, 2015) because it helps us realize that expertise does not reduce to any of these essentialist definitions (i.e., A–D), although each of them may be heuristically useful, in specific settings, to monitor specific aspects of specific performances: for example, when we need to explain the reliability of a professional or predict their probabilities of success. But neither one's epistemic entitlement (A), nor the length of their training (B), nor the depth and breadth of their scholarship (C), nor the capability to serve instrumentally one's goals (D) are sufficient to capture the experiential quality (the phenomenological relation) and the cognitive form (the informational relation) that distinctively characterise the expert's skilful engagement.

Radical Habitualism consists also in a form of *particularism about skill* because it emphasizes that know-how is a *situation-specific* and *applied* kind of knowledge that arises dynamically from a meta-stable (thus always precarious) agent-environment coupling process. Even the most experienced agent does not perform like an expert all the time: their modality of attunement continuously fluctuates, co-varying with a myriad of relational aspects that the four stereotypes (A–D) cannot capture. For example: how resolutely the agent approaches a challenge and whether they are truly ready to perform, sustained by a robust familiarity with similar tasks, confident self-regulation routines, and a harmonious attunement with the context. That is why Radical Habitualism advocates appropriate ecological consideration for the galaxies of interconnected practices in which embodied competences became relevant and useful.

Importantly, Radical Habitualism offers a robust and flexible framework to appreciate the relational and situational qualities of skilful control that the rationalist approaches systematically neglect: skill level increases when the agent-environment coupling becomes rich, robust and mature enough to allow the agent to navigate both routine and unexpected challenges in a *spontaneous, confident, self-reliant, harmonious, parsimonious, and effortless* manner. These qualities can consolidate only moving through an experiential itinerary that makes familiar and predictable the kinds of practical challenges that novices and unprepared experts can only encounter as discomfortingly unfamiliar and uncontrollable. If one reliably performs like an expert, that is normally because they eventually stopped using rules and instructions as *expedients to produce the outcomes of an expert*. To become an expert, the content of their know-that must finally dissolve to give life to habitual dispositions: crucially, this does not only mean that skilful performance cannot be based on know-that; it also means that the prolonged persistence of know-that is an inherent limit, a burden, and possibly a threat to one's capability to perform like an expert.

Also, Radical Habitualism is superior to the other accounts because it has a truly *explanatory* valence and not a merely *constitutive* one (refer to Sect. 5.3.3). With regard to the essential constitution of skill, Dreyfus' account is useful to recognize both the necessary and the sufficient conditions of expert performance: *necessary* is the progressive automatization of the action components, while *sufficient* is the presence of certain patterns of epistemic qualities that distinctively characterize skilful action control. This means that Radical Habitualism gives us universal and unambiguous principles, concerning the general constitution of skilled actions, to determine whether an action is performed skilfully or not. The first criterion is whether action is controlled habitually (necessary condition). Notably, complex actions are composed of different sub-actions and components, which is why, before giving a simple judgement about the action as a whole, one has to consider whether each component is controlled habitually (automatized) or not. The universal principle to be adopted here is: a greater incidence of habitually controlled action components indicates a higher level of skilful expertise. The second criterion (sufficient condition) is whether the action's form of execution displays the typical epistemic properties of expert performances (private, context-sensitive, trusted) and the associated virtues of expert execution (self-reliant, harmonious, parsimonious, etc.), to be considered jointly as a coherent configuration of markers of skilful expertise. The principle to be adopted here is that of family resemblances: the more closely the form of execution approximates the ideal normative profile of an expert action, through any one of its distinctive patterns of qualities and virtues, the higher is the expertise level of the performer.

Most important, Dreyfus' account also offers a theoretical platform to articulate concrete, accurate, and consistent *explanations* of the systems and processes that underpin the genesis and manifestation of skilful control in real life scenarios. Unlike the arm-chair approach favoured by analytic epistemologists and metaphysicians, who rely prevalently, if not exclusively, on "linguistic analyses and cultivated intuitions" (Noë, 2005, p. 288), the ambitious synthesis offered by Radical Habitualism is backed by the experimental data collected by three generations of sport & performance psychologists, the experiential accounts of professional practitioners, the methodological indications of naturalistic research, and the dynamicist and probabilistic models of machine learning.

Author contributions Single author submission.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest I have no conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative

Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abrahamson, D., & Abdu, R. (2020). Towards an ecological–dynamics design framework for embodied–interaction conceptual learning: The case of dynamic mathematics environments. In T. J. Kopcha, K. D. Valentine, & C. Ocak (Eds.), *Embodied cognition and technology for learning*. Educational Technology Research and Development.
- Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education*. <https://doi.org/10.3389/feduc.2020.00147>
- Adams, F., & Aizawa, K. (2017). *Causal Theories of Mental Content*. The Stanford Encyclopedia of Philosophy.
- Andler, D. (2000). Context and background. Dreyfus and cognitive science. *Heidegger, Coping and Cognitive Science* (pp. 137–159). MIT Press.
- Araújo, D., & Davids, K. (2011). What exactly is acquired during skill acquisition? *Journal of Consciousness Studies*, 18(3–4), pp. 3–4.
- Araújo, D., Davids, K., & Hristovski, R. (2006). The ecological dynamics of decision making in sport. *Psychology of Sport and Exercise*, 7, pp. 653–676.
- Baggio, G. (2021). Imagery in action: G. H. Mead's contribution to sensorimotor enactivism. *Phenomenology and the Cognitive Sciences*, 20(5), pp. 935–955.
- Barandiaran, X. E. (2017). Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, 36, pp. 409–430.
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3), pp. 610–620.
- Beilock, S. L. (2008). Beyond the playing field: Sport psychology meets embodied cognition. *International Review of Sport and Exercise Psychology*, 1(1), pp. 19–30.
- Beilock, S. L. (2011). *Choke*. Little, Brown Book Group.
- Beilock, S. L. (2015). *How the Body Knows Its Mind: The Surprising Power of the Physical Environment to Influence How You Think and Feel*. Atria Books.
- Beilock, S. L., Bertenthal, B. I., Hoeger, M., & Carr, T. H. (2008). When does haste make waste? *Speed–Accuracy Tradeoff, Skill Level, and the Tools of the Trade*, *Journal of Experimental Psychology*, 14(4), pp. 340–352.
- Beilock, S. L., Bertenthal, B. I., McCoy, A. M., & Carr, T. H. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review*, 11(2), pp. 373–379.
- Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology*, 130(4), pp. 701–725.
- Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and 'choking under pressure' in math. *Psychological Science*, 16(2), pp. 101–105.
- Beilock, S. L., Carr, T. H., MacMahon, C., & Starkes, J. L. (2002a). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology Applied*, 8(1), pp. 6–16.
- Beilock, S. L., & Gonso, S. (2008). Putting in the mind versus putting on the green: Expertise, performance time, and the linking of imagery and action. *The Quarterly Journal of Experimental Psychology*, 61(6), pp. 920–932.
- Beilock, S. L., & Gray, R. (2012). From attentional control to attentional spillover: A skill-level investigation of attention, movement, and performance outcomes. *Human Movement Science*, 31(6), pp. 1473–1499.
- Beilock, S. L., & Holt, L. E. (2007). Embodied preference judgments: Can likeability be driven by the motor system? *Psychological Science*, 18(1), pp. 51–57.

- Beilock, S. L., Wierenga, S. A., & Carr, T. H. (2002b). Expertise, attention, and memory in sensorimotor skill execution: Impact of novel task constraints on dual-task performance and episodic memory. *The Quarterly Journal of Experimental Psychology*, 55A(4), pp. 1211–1240.
- Beilock, S. L. & Carr, T. H. (2004) From novice to expert performance: Memory, attention and the control of complex sensorimotor skills, in Williams, A. M., Hodges,
- Bermúdez, J. P. (2017). Do we reflect while performing skillful actions? Automaticity, control, and the perils of distraction. *Philosophical Psychology*, 30(7), pp. 896–924.
- Bermúdez, J. P. (2021). The skill of self-control. *Synthese*, 199(3–4), pp. 6251–6273.
- Birch, J. E., Breivik, G., & Fusche Moe, V. (2019). Knowledge, consciousness and sporting skills. In M. Cappuccio (Ed.), *Handbook of Embodied Cognition and Sport Psychology*. MIT Press.
- Bocanegra, B. R., & Hommel, B. (2014). When cognitive control is not adaptive. *Psychological Science*, 25(6), pp. 1249–1255.
- Breivik, G. (2007). Skillful Coping in everyday life and in sport: A critical examination of the views of heidegger and dreyfus. *Journal of the Philosophy of Sport*, 34(2), pp. 116–134.
- Breivik, G. (2013). Zombie-like or superconscious?: A phenomenological and conceptual analysis of consciousness in elite sport. *Journal of the Philosophy of Sport*, 40, pp. 85–106.
- Brownstein, M. (2014a). Rationalizing flow: Agency in skilled unreflective action. *Philosophical Studies*, 168(2), pp. 545–568.
- Brownstein, M. (2014b). *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford University Press.
- Brozzo, C. (2017). Motor intentions: How intentions and motor representations come together. *Mind & Language*, 32(2), pp. 231–256.
- Bruineberg, J., Chemero, A., & Rietveld, E. (2019). General ecological information supports engagement with affordances for “higher” cognition. *Synthese*, 196(12), pp. 5231–5251.
- Bruineberg, J., Seifert, L., Rietveld, E., & Kiverstein, J. (2021). Metastable attunement and real life skilled behaviour. *Synthese*. <https://doi.org/10.1007/s11229-021-03355-6>
- Bryan, W. L., & Harter, N. (1899). Studies of the telegraphic language. The acquisition of a hierarchy of habits. *Psychological Review*, 6, pp. 345–375.
- Butterfill, S. A., & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88(1), pp. 119–145.
- Button, C., Seifert, L., Chow, J. Y., Araujo, D., & Davids, K. (2020). *Dynamics of Skill Acquisition: An Ecological Dynamics Approach*. Human Kinetics.
- Bäckström, S., & Gustafsson, M. (2017). Skill, drill, and intelligent performance: Ryle and intellectualism. *Journal for the History of Analytical Philosophy*. <https://doi.org/10.15173/jhap.v5i5.3205>
- Cappuccio, M. L. (Ed.). (2019). *Handbook of Embodied Cognition and Sport Psychology*. MIT Press.
- Cappuccio, M. L., Gray, R., Hill, D., Mesagno, C., & Carr, T. (2019a). The Many Threats of Self-Consciousness: Embodied Approaches to Choking under Pressure in Sensorimotor Skills. In M. L. Cappuccio (Ed.), *Handbook of Embodied Cognition and Sport Psychology* (pp. 101–156). The MIT Press.
- Cappuccio, M. L., & Ilundáin-Agurruza, J. (2020). Swim or Sink: Habit and Skillful Control in Sport Performance. In F. Carauna & I. Testa (Eds.), *Habits: Pragmatist Approaches from Cognitive Neuroscience to Social Science*. Cambridge University Press.
- Cappuccio, M. L., Kirchhoff, M. D., Alnajjar, F., & Tani, J. (2019b). Unfulfilled prophecies in sport performance: active inference and the choking effect. *Journal of Consciousness Studies*, 27(3–4), pp. 152–184.
- Cappuccio, M., Miyahara, K., & Ilundáin-Agurruza, J. (2020). Wax on wax off! habits, sports skills, and motor intentionality. *Topoi*, 40(3), 609–622.
- Cappuccio, M. L. (2017). Flow, choke, skill: The role of the non-conscious in sport performance. In Z. Radman (Ed.), *Before consciousness: In search of the of the Fundamentals of Mind*. Imprint Academic.
- Cappuccio, M., & Wheeler, M. (2012). Ground-level intelligence: action-oriented representation and the dynamics of the background. In Z. Radman (Ed.), *Knowing without Thinking*. Palgrave-MacMillan.
- Cappuccio, M., Wheeler, M. (2010). ‘When the twain meet: Could the study of mind be a meeting of minds?’ In: J. Reynolds, J. Chase, J. Williams, and E. Mares (eds) *Postanalytic and Metacontinental: Crossing Philosophical Divides Continuum*
- Cappuccio, M. L. (ed.) (2015). Unreflective action and the choking effect, *Phenomenology and the Cognitive Sciences*, 14 (2).
- Cappuccio (forthcoming). Dreyfus is Right: Knowledge-That Constrains Your Skill, *Synthese*.

- Carr, D. (1979). The logic of knowing how and ability. *Mind*, 88, pp. 394–409.
- Carr, T. (2015). Strengths and weaknesses of reflection as a guide to action: Pressure assails performance in multiple ways. *Phenomenology and the Cognitive Sciences*, 14, pp. 227–252.
- Caruana, F., & Testa, I. (2020). *Habit: Pragmatist approaches from cognitive neurosciences to social sciences*. Cambridge University Press.
- Castaneda, B., & Gray, R. (2007). Effects of focus of attention on baseball batting performance in players of differing skill levels. *Journal of Sport & Exercise Psychology*, 29(1), pp. 60–77.
- Chauvel, G., Maquestiaux, F., Ruthruff, E., Didierjean, A., & Hartley, A. A. (2013). Novice motor performance: Better not to verbalize. *Psychonomic Bulletin & Review*, 20(1), pp. 177–183.
- Chemero, T. (2009). *Radically enactive cognitive science*. MIT Press.
- Christensen, W. (2019). Skilled action. *Philosophy Compass*, 14(11), p. e12631.
- Christensen, W., & Bicknell, K. (2022). Cognitive control, intentions, and problem solving in skill learning. *Synthese*, 200(6), p. 460.
- Christensen, W., & Bicknell, K. (2019). Affordances and the Anticipatory Control of Action. In M. L. Cappuccio (Ed.), *Handbook of Embodied Cognition and Sport Psychology*. The MIT Press.
- Christensen, W., Sutton, J., & Bicknell, K. (2019a). Memory systems and the control of skilled action. *Philosophical Psychology*, 32(5), pp. 692–718.
- Christensen, W., & Sutton, J. (2019). Mesh: Cogssnition, body, and environment in skilled action—a new introduction to “cognition in skilled action.” In M. L. Cappuccio (Ed.), *Handbook of Embodied Cognition and Sport Psychology*. The MIT Press.
- Christensen, W., Sutton, J., & McIlwain, D. J. F. (2014). Putting pressure on theories of choking: Towards an expanded perspective on breakdown in skilled performance. *Phenomenology and the Cognitive Sciences*, 14(2), pp. 253–293.
- Christensen, W., Sutton, J., & McIlwain, D. J. F. (2016). Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind and Language*, 31(1), pp. 37–66.
- Christensen, W. D., Sutton, J., & Bicknell, K. (2019b). Meshed control in skilled action.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), pp. 7–19.
- Coeckelbergh, M. (2012). Technology as skill and activity: revisiting the problem of alienation. *Techné: Research in Philosophy and Technology*, 16(3), pp. 208–230.
- Coeckelbergh, M. (2015). *Environmental Skill: Motivation, Knowledge, and the Possibility of a Non-Romantic Environmental Ethics*. Routledge.
- Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: A motor-visual attentional effect. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), pp. 1673–1692.
- Davids, K., Araújo, D., Hristovski, R., Passos, P., & Chow, J. Y. (2012). Ecological dynamics and motor learning design in sport. In M. Williams & N. Hodges (Eds.), *Skill Acquisition in Sport: Research, Theory & Practice* (2nd ed., pp. 112–130). Routledge.
- Dennett, D. (1984). Cognitive wheels: The frame problem of AI. *Mind, Machines, and Evolution*. University Press.
- Dewey, J. (1983). Human nature and conduct. In *The middle works of John Dewey, 1925–1953. Volume 14: 1922, Human nature and conduct*, edited by Jo Ann Boydston, 1–437. Southern Illinois University Press, Carbondale.
- Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. E. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.
- Di Paolo, E. A., & Iizuka, H. (2008). How (not) to model autonomous behaviour. *Bio Systems*, 91(2), pp. 409–423.
- Dow, J. M. (2017). Just doing what I do: On the awareness of fluent agency. *Phenomenology and the Cognitive Sciences*, 16(1), pp. 155–177.
- Dreyfus, H. L. (1990). *Being-in-the-World: A Commentary on Heidegger's Being in Time*. MIT Press.
- Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.
- Dreyfus, H. L. (1993). Review of “The Embodied Mind.” *Mind*, 102, pp. 542–546.
- Dreyfus, H. L. (1996). The current relevance of Merleau-Ponty's theory of embodiment. *The Electronic Journal of Analytic Philosophy*, 4(4), pp. 1–16.
- Dreyfus, H. L. (1999). The primacy of phenomenology over logical analysis. *Philosophical Topics*, 27(2), pp. 3–24.
- Dreyfus, H. L. (2000). A Merleau-Pontyan Critique of Husserl's and Searle's Representationalist accounts of action. *Proceedings of the Aristotelian Society*, 100, pp. 287–330.

- Dreyfus, H. L. (2002a). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, 1, pp. 367–383.
- Dreyfus, H. L. (2002b). Refocusing the question: Can there be skillful coping without propositional representations or brain representations? *Phenomenology and the Cognitive Sciences*, 1, pp. 413–425.
- Dreyfus, H. L. (2005). Overcoming the myth of the mental: How philosophers can profit from the phenomenology of everyday expertise. *Proceedings and Addresses of the American Philosophical Association*, 79, pp. 47–63.
- Dreyfus, H. L. (2007a). The return of the myth of the mental. *Inquiry*, 50, pp. 352–365.
- Dreyfus, H. L. (2007b). Response to McDowell. *Inquiry*, 50, pp. 371–377.
- Dreyfus, H. L. (2013). The Myth of the Pervasiveness of the Mental. *Mind, reason, and being-in-the-world*. Routledge.
- Dreyfus, S. E. (2015). System 0: the overlooked explanation of expert intuition. *Handbook of Research Methods on Intuition*. Edward Elgar Publishing.
- Dreyfus, S. E., & Dreyfus, H. L. (1980). *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. Storming Media.
- Dreyfus, H. L., & Dreyfus, S. E. (1988). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. The Free Press.
- Dreyfus, H. L., & Dreyfus, S. E. (2014). The Perceptual noema. In H. L. Dreyfus (Ed.), *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*. Oxford University Press.
- Dreyfus, H. L., & Dreyfus, S. E. (1999). The Challenge of Merleau-Ponty's Phenomenology of Embodiment for Cognitive Science. In H. Haber & G. Weiss (Eds.), *Perspectives on Embodiment: The Intersections of Nature and Culture*. Routledge.
- Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland, & M. Wheeler (Eds.), *The Mechanical Mind in History* (pp. 331–371). MIT Press.
- Dreyfus, H. L. (2012). The mystery of the background qua background. In Z. Radman (Ed.), *Knowing without Thinking: The Background in Philosophy of Mind* (pp. 1–10). Palgrave Macmillan.
- Dreyfus, H. L. (2014). Skillful coping: essays on the phenomenology of everyday perception and action. In M. Wrathall (Ed.), *Oxford University Press*.
- Dreyfus, H. (1997). Intuitive, Deliberative, and Calculative Models of Expert Performance. In C. Zsombok & G. Klein (Eds.), *Naturalistic Decision Making*. Psychology Press.
- Egbert, M. D., & Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience*, 8, p. 590.
- Eichenbaum, H., & Cohen, N. J. (2004). Habits, Skills, and Procedural Memory. In H. Eichenbaum & N. J. Cohen (Eds.), *From Conditioning to Conscious Recollection*. Oxford University Press.
- Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 15(11), pp. 988–994.
- Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*. Houghton Mifflin Harcourt.
- Eriksen, J. W. (2010). Mindless coping in competitive sport: Some implications and consequences. *Sport, Ethics, & Philosophy*, 4, pp. 66–86.
- Evans, J. S. B. T., St. B. T., Evans, J., & Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspectives on Psychological Science*, 8(3), pp. 223–241.
- Facchin, M. (forthcoming). Phenomenal transparency, cognitive extension, and predictive processing. *Phenomenology and the Cognitive Sciences*:1–23.
- Ferretti, G., & Zipoli Caiani, S. (2021). How knowing-that and knowing-how interface in action: The intelligence of motor representations. *Erkenntnis: an International Journal of Analytic Philosophy*, 88, pp. 1103–1133.
- Fitts, P. M., & Posner, M. I. (1967). *Human Performance*. Brooks/Cole.
- Flegal, K. E., & Anderson, M. C. (2008). Overthinking skilled motor performance: Or why those who teach can't do. *Psychonomic Bulletin & Review*, 15, pp. 927–932.
- Fodor, J. (1981). *Representations*. MIT Press.
- Fodor, J. (1983). *The Modularity of Mind*. MIT Press.
- Foerster, A., Gray, R., & Cañal-Bruland, R. (2015). Size estimates remain stable in the face of differences in performance outcome variability in an aiming task. *Consciousness and Cognition*, 33, pp. 47–52.
- Francesconi, D., & Gallagher, S. (2019). Embodied cognition and sport pedagogy. In M. L. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology* (pp. 249–272). The MIT Press.
- Freeman, W. (2000). *How Brains Make Up Their Minds*. Columbia University Press.

- Fridland, E. (2014). They've lost control: Reflections on skill. *Synthese*, 191(12), pp. 2729–2750.
- Fridland, E. (2017). Skill and motor control: Intelligence all the way down. *Philosophical Studies*, 174(6), pp. 1539–1560.
- Fridland, E. (2019). Longer, smaller, faster, stronger: On skills and intelligence. *Philosophical Psychology*, 32(5), 759–783.
- Fridland, E. (2021). Skill and strategic control. *Synthese*, 199, pp. 5937–5964.
- Fridland, E. (2013). Skill learning and conceptual thought: Making our way through the wilderness. In B. Bashour & H. Muller (Eds.), *Philosophical naturalism and its implications* (pp. 77–100). Routledge.
- Fridland, E. (2020). The nature of skill: functions and control structures. In E. Fridland, C. Pavese, E. Fridland, & C. Pavese (Eds.), *The Routledge handbook of philosophy of skill and expertise*. Routledge.
- Fridland, E., & Pavese, C. (2020). Introduction to the Routledge handbook of philosophy of skill and expertise. In E. Fridland, C. Pavese, E. Fridland, & C. Pavese (Eds.), *the Routledge handbook of philosophy of skill and expertise*. Routledge.
- Fusche Moe, V. (2004). How to understand skill acquisition in sport. *Bulletin of Science, Technology & Society*, 24(3), pp. 213–224.
- Fusche Moe, V. (2007). Understanding the background conditions of skilled movement in sport: A study of Searle's "background capacities. *Sports Ethics and Philosophy*, 1(3), pp. 299–324.
- Fusche Moe, V. (2016). On phenomenological and logical characteristics of skilled behaviour in sport: cognitive and motor intentionality. *Skills, Knowledge and Expertise in Sport*, 10(3), pp. 35–52.
- Gallagher, S. (2004). Hermeneutics and the cognitive sciences. *Journal of Consciousness Studies*, 11(10–11), pp. 162–174.
- Gallagher, S. (2007). Moral agency, self-consciousness, and practical wisdom. *Journal of Consciousness Studies*, 14(5–6), pp. 199–223.
- Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies*, 16(3), pp. 351–369.
- Gallagher, S. (2009). Book review of reading Merleau-Ponty: On phenomenology of perception. *Mind*, 118(472), pp. 1105–1111.
- Gallagher, S. (2014). Pragmatic interventions into enactive and extended conceptions of cognition. *Philosophical Issues*, 24(1), pp. 110–126.
- Gallagher, S. (2017). *Enactivist Interventions*. MIT Press.
- Gallagher, S. (2018). Mindfulness and mindlessness in performance. *Reti, Saperi, Linguaggi, Italian Journal of Cognitive Sciences.*, 1(2018), pp. 5–18.
- Gallagher, S. (2016). The practice of thinking: Between Dreyfus and McDowell. In T. Breyer (Ed.), *The Phenomenology of Thinking* (pp. 134–146). Routledge.
- Gallagher, S., Sparaci, L., & Varga, S. (2022). Disruptions of the meshed architecture in autism spectrum disorder. *Psychoanalytic Inquiry*, 42(1), pp. 76–95.
- Gallagher, S., & Varga, S. (2020). Meshed architecture of performance as a model of situated cognition. *Frontier Psychology*, 11, p. 2140.
- Gehrman, K. (2016). Absorbed coping and practical wisdom. *The Journal of Value Inquiry*, 50(3), pp. 593–612.
- Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.
- Gottlieb, G. (2015). Know-how, procedural knowledge, and choking under pressure. *Phenomenology and the Cognitive Sciences*, 14(2), pp. 361–378.
- Gray, R. (2014). Embodied perception in sport. *International Review of Sport and Exercise Psychology*, 7(1), pp. 72–86.
- Gray, R., & Cañal-Bruland, R. (2015). Attentional focus, perceived target size, and movement kinematics under performance pressure. *Psychonomic Bulletin & Review*, 22(6), pp. 1692–1700.
- Gray, R. (2004). Attending to the execution of a complex sensorimotor skill: Expertise, differences, choking, and slumps. *Journal of Experimental Psychology: Applied*, 10(1), pp. 42–54.
- Gray, R. (2021). *How We Learn to Move: A Revolution in the Way We Coach & Practice Sports Skills*, Perception Action Consulting & Education LLC
- Heidegger, M. (1927/1996) Being and Time, John Stambaugh (trans.) (Albany: State University of New York Press). Originally published as. (1927). *Sein und Zeit*. Max Niemeyer Verlag.
- Heras-Escribano, M. (2021). Pragmatism, enactivism, and ecological psychology: Towards a unified approach to post-cognitivism. *Synthese*, 198, pp. 337–363.
- Van den Herik, J. C., & Rietveld, E. (2021). Reflective situated normativity. *Philosophical Studies*, 178(10), pp. 3371–3389.

- Hill, D. M., Hanton, S., Matthews, N., & Fleming, S. (2010). Choking in sport: A review. *International Review of Sport and Exercise Psychology*, 3(1), pp. 24–39.
- Hipólito, I., Baltieri, M., Friston, K., & Ramstead, M. J. D. (2021). Embodied skillful performance: Where the action is. *Synthese*, 199, pp. 4457–4481.
- Holt, L. E., & Beilock, S. L. (2006). Expertise and its embodiment: Examining the impact of sensorimotor skill expertise on the representation of action-related text. *Psychonomic Bulletin & Review*, 13(4), pp. 694–701.
- Hutto, D. (2013). Exorcising action-oriented representations: ridding cognitive science of its Nazgûl. *Adaptive Behaviour*, 21(3), pp. 142–215.
- Hutto, D. D., & Myin, E. (2013). *Radicalising enactivism*. MIT Press.
- Hutto, D. D., & Myin, E. (2017). *Evolving enactivism*. MIT Press.
- Hutto, D. (2012). Exposing the background: Deep and local. In Z. Radman (Ed.), *Knowing Without Thinking: Mind, Action, Cognition and the Phenomenon of the Background* (pp. 37–56). Palgrave Macmillan.
- Hutto, D., & Robertson, I. (2020). Clarifying the Character of Habits: Understanding what and how they explain. In C. Fausto & I. Testa (Eds.), *Habit: Pragmatist Approaches from Cognitive Neurosciences to Social Sciences*. Cambridge University Press.
- Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia*, 43(3), 521–536.
- Hutto, D. D., & Sánchez García, R. (2015). Choking REctified: Embodied expertise beyond Dreyfus. *Phenomenology and the Cognitive Sciences*, 14(2), pp. 309–331.
- Ilundáin-Agurruza, J. (2016). *Holism and the Cultivation of Excellence in Sports and Performance – Skillful Striving*. Routledge.
- Jackson, S. A. (1996). Toward a conceptual understanding of the flow experience in elite athletes. *Research Quarterly for Exercise and Sport*, 67(1), 76–90.
- Jackson, G. B. (2018). Maurice Merleau-Ponty's concept of motor intentionality: Unifying two kinds of bodily agency. *European Journal of Philosophy*, 26(2), pp. 763–779.
- Jackson, R. C., Ashford, K. J., & Norsworthy, G. (2006). Attention focus, dispositional reinvestment, and skillful motor performance under pressure. *Journal of Sport and Exercise Psychology*, 28, pp. 49–68.
- Jackson, R. C., & Beilock, S. (2008). Attention and performance. In D. Farrow, J. Baker, & C. MacMahon (Eds.), *Developing Elite Sports Performers: Lessons from Theory and Practice*. Routledge.
- James, W. (1890). *Principles of psychology*. Holt.
- Kelly, S. D. (2000). *Grasping at straws: Motor intentionality and the cognitive science of skillful action, in Essays in Honor of Hubert Dreyfus* (Vol. II). MIT Press.
- Kelly, S. D. (2002). Merleau-Ponty on the body. *Ratio*, 15(4), pp. 376–391.
- Kirchhoff, M. D., & Kiverstein, J. (2020). Attuning to the world: The diachronic constitution of the extended conscious mind. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2020.01966>
- Kiverstein, J., & Rietveld, E. (2015). (2015), The primacy of skilled intentionality: On Hutto & Satne's the natural origins of content. *Philosophia*, 43, pp. 701–721.
- Kiverstein, J., & Rietveld, E. (2018). Adaptive Behavior Reconceiving representation-hungry cognition: an ecological-enactive proposal. *Adaptive Behavior*, 26(4), pp. 147–163.
- Kiverstein, J., & Rietveld, E. (2021). Scaling-up skilled intentionality to linguistic thought. *Synthese*, 198, pp. 175–194.
- Komar, J., Chow, J.-Y., Chollet, D., & Seifert, L. (2014). Effect of analogy instructions with an internal focus on learning a complex motor skill. *Journal of Applied Sport Psychology*, 26(1), pp. 17–32.
- Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 15(5), pp. 1256–1271.
- Legg, C. (2021). Discursive habits: A representationalist re-reading of teleosemiotics. *Synthese*, 5–6, pp. 14751–14768.
- Legg, C., & Reynolds, J. (2022). Habits of mind: new insights for embodied cognition from classical pragmatism and phenomenology. *European Journal of Pragmatism and American Philosophy*. <https://doi.org/10.4000/ejap.2994>
- Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Experimental Psychology/revue Canadienne De Psychologie Experimentale*, 39(2), pp. 367–386.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95(4), pp. 492–527.
- Magri, E. (2019). Situating attention and habit in the landscape of affordances. *Rivista Internazionale Di Filosofia E Psicologia*, 10(2), pp. 120–136.

- Maric, M., Heyne, D. A., van Widenfelt, B., & M., & Westenberg, P. M. (2011). Distorted cognitive processing in youth: The structure of negative cognitive errors and their associations with anxiety. *Cognitive Therapy and Research*, 35(1), pp. 11–20.
- Martens, J. H. (2021). Habit and skill in the domain of joint action. *Topoi*, 40, pp. 663–675.
- Masters, R. S. W. (1992). Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure. *British Journal of Psychology*, 83(3), pp. 343–358.
- Masters, R., & Maxwell, J. (2008). The theory of reinvestment. *International Review of Sport and Exercise Psychology*, 1(2), 160–183.
- Masters, R. S. W., & Maxwell, J. P. (2004). Implicit motor learning, reinvestment and movement disruption: What you don't know won't hurt you? In A. M. Williams & N. J. Hodges (Eds.), *Skill Acquisition in Sport: Research, Theory and Practice*. Routledge.
- Maxwell, J. P., Masters, R. S. W., & Eves, F. F. (2000). From novice to know-how: A longitudinal study of implicit motor learning. *Journal of Sports Sciences*, 18(11), p. 120.
- Maxwell, J. P., Masters, R. S. W., & Eves, F. F. (2003). The role of working memory in motor learning and performance. *Consciousness and Cognition*, 12(376), p. 402.
- McDowell, J. (2007). Response to Dreyfus. *Inquiry*, 50(4), pp. 366–370.
- McDowell, J. (1981). Non-cognitivism and rule-following. In S. Holtzman & C. M. Leich (Eds.), *Wittgenstein: To Follow A Rule*. Routledge.
- McManus, D. (2008). Rules, regression and the “background”: Dreyfus, Heidegger and McDowell. *European Journal for Philosophy of Science*, 16(3), pp. 432–458.
- Menary, R. (2010). *The Extended Mind*. MIT Press.
- Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Routledge.
- Mesagno, C., & Hill, D. M. (2013). Definition of choking in sport: Reconceptualization and debate. *International Journal of Sport Psychology*, 44(4), pp. 267–277.
- Miyahara, K., Ransom, T. G., & Gallagher, S. (2020). What the situation affords: Habit and heedful interrelations in skilled performance. In C. Fausto & I. Testa (Eds.), *Habit: Pragmatist Approaches from Cognitive Neurosciences to Social Sciences*. Cambridge University Press.
- Miyahara, K., & Robertson, I. (2021). The pragmatic intelligence of habits. *Topoi*, 40(3), pp. 597–608.
- Montero, B. (2010). Does bodily awareness interfere with highly skilled movement? *Inquiry a Journal of Medical Care Organization, Provision and Financing*, 53(2), pp. 105–122.
- Montero, B. G. (2014). Is monitoring one's actions causally relevant to choking under pressure? *Phenomenology and the Cognitive Sciences*, 14(2), pp. 379–395.
- Montero, B. G. (2016). *Thought in Action: Expertise and the Conscious Mind*. Oxford University Press.
- Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), pp. 297–326.
- Moya, P. (2014). Habit and embodiment in Merleau-Ponty. *Frontiers in Human Neuroscience*, 8, p. 542.
- Mylopoulos, M., & Pacherie, E. (2019). Intentions: The dynamic hierarchical model revisited. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(2), p. e1481.
- Mylopoulos, M., & Pacherie, E. (2021). Skilled action control. *Review of Philosophy and Psychology*, 12(3), pp. 469–480.
- Hodges, N. J., Scott, M. A. & Court, M. L. J. *Skill Acquisition in Sport: Research, Theory and Practice*, pp. 309–328, Routledge
- Neemeh, Z. A. (2022). Smooth coping: an embodied, Heideggerian approach to dual-process theory. *Adaptive Behavior*, 30(4), pp. 329–344.
- Newen, A., De Bruin, L., & Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford University Press.
- Noë, A. (2005). Against intellectualism. *Analysis*, 65(4), pp. 278–290.
- Pacherie, E., & Mylopoulos, M. (2021). Beyond automaticity: The psychological complexity of skill. *Topoi: an International Review of Philosophy*, 40(3), pp. 649–662.
- Paolucci, C. (2021). Cognitive Semiotics: Radical Enactivism, Pragmatism and Material Engagement. *Cognitive Semiotics: Perspectives in Pragmatics, Philosophy & Psychology*. Springer.
- Papineau, D. (2015). Choking and the yips. *Phenomenology and the Cognitive Sciences*, 14(2), pp. 295–308.
- Pavese, C. (2017). Know-how and gradability. *Philosophical Review*, 126(3), pp. 345–383.
- Pavese, C. (2019). The psychological reality of practical representation. *Philosophical Psychology*, 32(5), pp. 784–821.

- Pavese, C. (2021). Knowledge-how, *Stanford Encyclopedia of Philosophy*.
- Porter, Wu., & Partridge,. (2010). Focus of attention and verbal instructions: strategies of elite track and field coaches and athletes. *Sport Science Review*, 19(3–4), pp. 199–211.
- Ramírez-Vizcaya, S., & Froese, T. (2019). The enactive approach to habits: New concepts for the cognitive science of bad habits and addiction. *Frontiers in Psychology*, 10, p. 301.
- Ramírez-Vizcaya, S., & Froese, T. (2020). Agents of habit: Refining the artificial life route to artificial intelligence. In J. Bongard, J. Lovato, L. Hebert-Dufrésne, R. Dasari, & L. Soros (Eds.), *Artificial Life Conference Proceedings* (pp. 78–86). MIT Press.
- Reynolds, J. (2006). Dreyfus and Deleuze on L'habitude, coping, and trauma in skill acquisition. *International Journal of Philosophical Studies*, 14(4), pp. 539–559.
- Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind*, 117(468), pp. 973–1001.
- Rietveld, E., & Kiverstein, J. (2014). A Rich landscape of affordances. *Ecological Psychology*, 26(4), pp. 325–352.
- Rietveld, E. (2012). Context-switching and responsiveness to real relevance. In J. Kiverstein & M. Wheeler (Eds.), *Heidegger and Cognitive Science: New Directions in Cognitive Science and Philosophy* (pp. 105–135). Palgrave Macmillan.
- Rizzolatti, G., & Craighero, L. (2010). Premotor theory of attention. *Scholarpedia Journal*, 5(1), pp. 6311.
- Robertson, I., & Kirchhoff, M. (2020). Anticipatory action active inference in embodied cognitive activity. *Journal of Consciousness Studies*, 27(3–4), pp. 38–68.
- Robertson, I., Hutto, D. (forthcoming). Against Intellectualism about Skill.
- Rousse, B. S. (2019). Self-awareness and self-understanding. *European Journal for Philosophy of Science*, 27(1), pp. 162–186.
- Rousse, B., & Dreyfus, S. (2021). Revisiting the six stages of skill acquisition. *Teaching and Learning for Adult Skill Acquisition: Applying the Dreyfus & Dreyfus Model in Different Fields*. Charlotte, NC, USA
- Rucińska, Z. (2021). Enactive planning in rock climbing: Recalibration, visualization and nested affordances. *Synthese*. <https://doi.org/10.1007/s11229-021-03025-7>
- Rucińska, Z., & Aggerholm, K. (2019). Embodied and Enacted Creativity in Sports. In M. Cappuccio (Ed.), *Handbook of Embodied Cognition and Sport Psychology* (pp. 669–694). MIT Press.
- Rucińska, Z., & Gallagher, S. (2021). Making imagination even more embodied: Imagination, constraint and epistemic relevance. *Synthese*, 199(3–4), pp. 8143–8170.
- Ryle, G. (2009). *The concept of mind*. Routledge.
- Schmidt, R. A., & Wrisberg, C. A. (2008). *Motor learning and performance: a situation-based learning approach*. Human Kinetics.
- Segundo-Ortín, M., & Heras-Escribano, M. (2021). Neither mindful nor mindless, but minded: Habits, ecological psychology, and skilled performance. *Synthese*, 199(3), pp. 10109–10133.
- Springle, A. A., & Humphreys, J. (2021). Anti-intellectualism, instructive representations, and the intentional action argument. *Synthese*, 199, pp. 7919–7955.
- Stanley, J. (2011). *Know how*. Oxford University Press.
- Stanley, J., & Krakauer, J. (2013). Motor skill depends on knowledge of facts. *Frontiers of Human Neuroscience*. <https://doi.org/10.3389/fnhum.2013.00503>
- Stanley, J., & Williamson, T. (2017). *Skill. Noûs*, 51(4), pp. 713–726.
- Stephens, A., & Felix, C. V. (2020). A cognitive perspective on knowledge how: Why intellectualism is neuro-psychologically implausible. *Philosophies*, 5(3), p. 21.
- Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9(4), pp. 465–481.
- Sutton, J. (2007). Batting, habit, and memory: The embodied mind and the nature of skill. *Sport in Society*, 10(5), pp. 763–786.
- Sutton, J., & Bicknell, K. (2020). Embodied Experience in the Cognitive Ecologies of Skilled Performance. In E. Fridland, C. Pavese, E. Fridland, & C. Pavese (Eds.), *The Routledge Handbook of Philosophy of Skill And Expertise*. Routledge.
- Sutton, J., & McIlwain, D. J. F. (2015). Breadth and depth of knowledge in expert versus novice athletes. *Routledge Handbook of Sport Expertise*. Routledge.
- Sutton, J., McIlwain, D. J. F., Christensen, W., & Geeves, A. (2011). Applying Intelligence to the Reflexes: embodied skills and habits between Dreyfus and Descartes. *JBSP: Journal of the British Society for Phenomenology*, 42(1), pp. 78–103.

- Swann, C. (2016). Flow in sport. In L. Harmat, F. Ø. Andersen, F. Ullén, J. Wright, & G. Sadlo (Eds.), *Flow Experience: Empirical Research and Applications* (pp. 51–64). Springer.
- Swann, C., Keegan, R., Crust, L., & Piggott, D. (2016). Psychological states underlying excellent performance in professional golfers: ‘Letting it happen’ vs. ‘making it happen.’ *Psychology of Sport and Exercise*, 23, pp. 101–113.
- Sánchez García, R., Cappuccio, M. L., and Ilundáin -Agurruza, J. (forthcoming). Metis as a skill and an embodied virtue in sport.
- Taylor, C. (1993). Engaged agency and background in Heidegger. In C. B. Guignon (Ed.), *The Cambridge Companion to Heidegger* (pp. 317–336). Cambridge University Press.
- Thompson, E. (2014). *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. Columbia University Press.
- Thompson, E. (2018). Review of Daniel D. Hutto and Erik Myin, evolving enactivism: Basic minds meet content. *Notre Dame Philosophical Reviews*.
- Toner, J., Montero, B. G., & Moran, A. (2014). Considering the role of cognitive control in expert performance. *Phenomenology and the Cognitive Sciences*, 14(4), pp. 1127–1144.
- Toner, J., Montero, B. G., & Moran, A. (2015). The perils of automaticity. *Review of General Psychology*, 19(4), pp. 431–442.
- Toner, J., Montero, B. G., & Moran, A. (2016). Reflective and prereflective bodily awareness in skilled action. *Psychology of Consciousness: Theory, Research, and Practice*, 3(4), pp. 303–315.
- Toner, J., Montero, B., & Moran, A. (2021). *Continuous Improvement: Intertwining Mind and Body in Athletic Expertise*. Oxford University Press.
- Toner, J., & Moran, A. (2021). *Exploring the orthogonal relationship between controlled and automated processes in skilled action*. Springer.
- Ungureanu, C., & Rotaru, I. (2014). Philosophy of skillful coping. motor intentionality vs. representations for action. *Procedia—Social and Behavioral Sciences*, 163, pp. 220–229.
- Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.
- Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press.
- Wheeler, M. (2008). Cognition in context: Phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies*, 16(3), pp. 323–349.
- Wheeler, M. (2010b). The Problem of Representation. In S. Gallagher & D. Schmicking (Eds.), *Handbook of Phenomenology and Cognitive Science*. Springer.
- Wheeler, M. (2010a). In Defense of Extended Functionalism. In R. Menary (Ed.), *The Extended Mind*. The MIT Press.
- Wheeler, M. (2012). Minds, Things, and Materiality. In J. Schulkin (Ed.), *Action, Perception and the Brain*. Palgrave Macmillan UK.
- Witt, J. K., Linkenauger, S., Bakdash, J., & Proffitt, D. (2007). Golf performance makes the hole look as big as a bucket or as small as a dime. *Journal of Vision*, 7(9), p. 291.
- Wrathall, M. A. (2014). Introduction: Hubert Dreyfus and the Phenomenology of Human Intelligence. In M. A. Wrathall (Ed.), *Skillful Coping: Essays on the phenomenology of everyday perception and action*. Oxford Academic.
- Wu, W. (2013). Mental Action and the Threat of Automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the Will*. Oxford University Press.
- Wulf, G. (2013). Attentional focus and motor learning: A review of 15 years. *International Review of Sport and Exercise Psychology*, 6(1), pp. 77–104.
- Wulf, G. (2015). An external focus of attention is a conditio sine qua non for athletes: A response to Carson and Collins (2015). *Journal of Sports Sciences*, 34(13), pp. 1293–1295.
- Wulf, G., McConnel, N., Gärtner, M., & Schwarz, A. (2002). Enhancing the learning of sport skills through external-focus feedback. *Journal of Motor Behavior*, 34(2), pp. 171–182.
- Wulf, G., Shea, C. H., & Park, J.-H. (2001). Attention in motor learning: Preferences for and advantages of an external focus. *Research Quarterly for Exercise and Sport*, 72, pp. 335–344.
- Wulf, G., & Su, J. (2007). An external focus of attention enhances golf shot accuracy in beginners and experts. *Research Quarterly for Exercise and Sport*, 78, pp. 384–389.

Zahavi, D. (2012). Mindedness, mindlessness and first-person authority. In J. Shear (Ed.), *Mind, Reason and Being-in-the-World: The McDowell-Dreyfus Debate*. Routledge.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Support Publication #8 [SP8]

Sandoval E.B., Sosa R., Cappuccio M.L., Bednarz T (2022)

“Human–robot creative interactions:

Exploring creativity in artificial agents using a storytelling game”

Frontiers Robotics and AI - Section Human-Robot Interaction 24 August

2022 <https://doi.org/10.3389/frobt.2022.695162>

Reproduced with permission from Frontiers Media SA. All rights reserved.

Reproduced with permission from Frontiers Media SA. All rights reserved.



OPEN ACCESS

EDITED BY
Vasanth Sarathy,
Tufts University, United States

REVIEWED BY
Jill Dosso,
University of British Columbia, Canada
Patricia Alves-Oliveira,
University of Washington, United States

*CORRESPONDENCE
Eduardo Benítez Sandoval,
e.sandoval@unsw.edu.au

SPECIALTY SECTION
This article was submitted to Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

RECEIVED 14 April 2021
ACCEPTED 28 June 2022
PUBLISHED 24 August 2022

CITATION
Sandoval EB, Sosa R, Cappuccio M and
Bednarz T (2022), Human-robot
creative interactions: Exploring
creativity in artificial agents using a
storytelling game.
Front. Robot. AI 9:695162.
doi: 10.3389/frobt.2022.695162

COPYRIGHT
© 2022 Sandoval, Sosa, Cappuccio and
Bednarz. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](#). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Human-robot creative interactions: Exploring creativity in artificial agents using a storytelling game

Eduardo Benítez Sandoval^{1*}, Ricardo Sosa^{2,3},
Massimiliano Cappuccio⁴ and Tomasz Bednarz⁵

¹Creative Robotics Lab, School of Arts and Design, Faculty of Arts, Design and Architecture, University of New South Wales, Sydney, Australia, ²Faculty of Design and Creative Technologies, Auckland University of Technology, Auckland, New Zealand, ³Monash University, Design, Melbourne, Australia, ⁴University of New South Wales, Canberra, Australia, ⁵EPICentre and CSIRO's Data61, University of New South Wales, Sydney, Australia

Creativity in social robots requires further attention in the interdisciplinary field of human-robot interaction (HRI). This study investigates the hypothesized connection between the perceived creative agency and the animacy of social robots. The goal of this work is to assess the relevance of robot movements in the attribution of creativity to robots. The results of this work inform the design of future human-robot creative interactions (HRCI). The study uses a storytelling game based on visual imagery inspired by the game "Story Cubes" to explore the perceived creative agency of social robots. This game is used to tell a classic story for children with an alternative ending. A 2 x 2 experiment was designed to compare two conditions: the robot telling the original version of the story and the robot plot twisting the end of the story. A Robotis Mini humanoid robot was used for the experiment, and we adapted the Short Scale of Creative Self (SSCS) to measure perceived creative agency in robots. We also used the Godspeed scale to explore different attributes of social robots in this setting. We did not obtain significant main effects of the robot movements or the story in the participants' scores. However, we identified significant main effects of the robot movements in features of animacy, likeability, and perceived safety. This initial work encourages further studies experimenting with different robot embodiment and movements to evaluate the perceived creative agency in robots and inform the design of future robots that participate in creative interactions.

KEYWORDS

creative robots, robot games, story cubes, creative interactions, short-scale creative self scale, storytelling, the ugly duckling

1 Introduction

An important dimension in social interaction is how agents perceive the intelligence and creativity of other agents. However, creativity is an under-explored area in the study of human–robot interaction (HRI) (Saunders et al., 2013). Creativity can be defined as the capacity to imagine alternative futures. Despite its relevance in shaping everyday interactions, we have a limited knowledge of how creativity is perceived and attributed by humans to self and others. Furthermore, the display of creativity is a subjective phenomenon that is challenging to study using experimental methods of inquiry (Svanaes, 2013). This study is drawn from work in the area of self-assessment of creativity to evaluate the possible connections between the perceived creative agency and the animacy and kinesthetics of social robots.

In artificial agents such as social robots or screen-based avatars, the attribution of creativity has remained largely unaddressed in the field of HRI until recently (Alves-Oliveira et al., 2017, 2020). While artificial intelligence has been investigated extensively, and artificial creativity defined as the creativity attributed to artificial agents remains to be addressed particularly in experimental studies. Some creative behaviors have been simulated using language, pattern recognition, and evolutionary generative systems as shown (Pham et al., 2017; Gizzi et al., 2019; Myoo, 2019; Uneeq Ltd, Digital Humans, 2021; OpenAI, 2021). Similarly, some work has been performed on users' non-verbal behavior, self-presence, social presence, and interpersonal attraction in one-to-one human–agent interaction on collaborative virtual environments using realistic humanoid avatars (Herrera et al., 2020). However, in artificial agents such as social robots that rely on physical embodiment to interact with the users and the real-world, artificial creative behavior will need to be communicated to users verbally and kinesthetically, namely, using movement as part of their communicative means. To our knowledge, the display of creative behavior *via* physical movement (animacy) in robots is an area requiring further investigation by human–robot interaction researchers.

In human–human interaction (HHI), the study of non-verbal behavior in creative and collaborative tasks has been extensively studied. For instance, Won et al. (2014) suggest that there is a significant correlation between synchronized movements of a pair of humans and creative outcomes. Hence, the display of creative behavior *via* movement is likely to be of high relevance to determine the ultimate value and usefulness of social robots interacting with humans in everyday settings.

The impact of robots' physical presence and their movements has been studied previously across contexts (Vignolo et al., 2017). However, more research is needed to better understand how social robots can effectively use movement in everyday interactions, especially in light of screen-based smart applications and disembodied products that use voice interfaces. What may physical robots offer in terms of

functionality and usability that screen and voice agents cannot? What are the design affordances enabled by their physicality and movement possibilities that are not available on screen or voice interaction? The work presented in this article seeks to contribute to the future design of social robots by analyzing how humans perceive the robots' movement when these perform a task that requires creative behavior, such as play.

We propose an adapted scale to measure the perceived creativity in social robots in the context of games and playful activities such as creative storytelling. This scale is inspired in the work of Karwowski et al. (2018) and Karwowski (2014) and is modified here to capture how participants rate the creative skills of robots. We are particularly focused on the domain of creative collaborative interactions that could be eventually implemented using social robots.

The work with social agents as robots is relevant because excessive screen time is associated with ergonomic, visual, and behavioral issues. Furthermore, excessive use of screens for entertainment negatively impacts people at the level of being recently classified as “gaming addiction” (WHO, 2018) and influences people's overall mood, among other effects. At the moment, users intensively use screen-based devices for both work and entertainment services, resulting in extended periods of eye strain and sedentary lifestyle (Aboujaoude and Starcevic, 2015; Alter, 2018; Desmurget, 2020).

This work is an early exploration of HRI that does not rely on a screen to interact with users. We aim to study alternatives to screen and audio interactions using physical robots for interactive creative activities, relying on more natural interactions with artificial agents. We consider that creativity is a central part of HRI due to its importance in the cognitive and social processes involved in playful interactions. Finally, we expect to contribute in the near future in the design of robots encouraging natural, long-term interactions with cognitive and social gains.

2 Background

The design of social robots has shown initial evidence of their potential value for usability and functionality to assist users in everyday life. So far, the main applications of notable market success have been to carpet and floor robot cleaners and toys including robotic pets for therapeutic purposes. For the last 2 decades, researchers and companies have searched for the “killer app” that builds on the affordances of physical robots to transform the lives of users around the world. In that time, screen-based devices and home assistants that use audio interfaces have made substantial gains in market penetration. Currently, there is a need to understand if and how physical robots will be of value for users in their everyday tasks as suggested by the fiction (Miller, 2021). However, to our knowledge, there are no sufficient studies using social

robots aiming to study creative interactions between humans and robots, creative performance by the robot, or human's creativity enhancement. In contrast, there are numerous apps, software, and websites for this purpose of screen-based devices.

2.1 Robot's embodiment and movement

Arguably, the most salient affordance that gives social robots an edge over screen and voice assistants is their physical presence. The importance of physicality and movement in communication is evident from "body language" to 4E cognition, that is, the principle that all human cognition is embodied, embedded, enactive, and extended (Lindblom and Alenljung, 2015). In other words, people are not "brains in jars" but rely heavily on their bodies, the physical world around them, other humans, and their contexts to be able to think, communicate, and be fully human.

There have been studies on issues relevant to the design of robots with movement in mind like the study developed by Hoffman and Ju (2014). In the early stages of research in human-robot interaction researchers as Van Breemen (2004) understood that body gestures are a natural channel to communicate robot's agency and social behaviors. Furthermore, the motion of robot agents (mechanic and organic), as one of the main features differentiating robots from AI or computers, has been explored to highlight its relevance from robot esthetic and functional context by Harris and Sharlin (2011). Similarly, Bainbridge et al. (2011) explored how the physical presence of a robot affects human judgments of the robot as a social partner. Looking for the effects of form and motion in robotic agents, Castro-González et al. (2016) studied the attributions of animacy and investigated how the combination of robot's bodily appearance and movement can alter attributions of animacy, likability, trustworthiness, and unpleasantness in the users. Apparently, a Baxter robot executing mechanistic movement was perceived as inanimate. However, the same robot performing naturalistic movements was unpleasant. However, all these previous works have not been placed in the context of creative expressiveness of the robot agents.

Our work aims to explore and understand how animacy plays a role in the perception of social robots in the future design of interactive tasks related to creativity. According to the Oxford English Dictionary, animacy is "...the quality or condition of being alive or animate" (Animacy, 2021) while kinesthetics refers to "...the effort that accompanies a voluntary motion of the body" (Kinaesthesia, 2021). In the context of this research, kinesthetics and animacy refer to the study and perception of body motion, and the kinetic design refers to the use of movement as a design material (Sosa et al., 2015).

2.2 Games as a setup in HRI

We are interested to examine the physicality and moving affordances of social robots (their animacy qualities), and their potential advantages over traditional board games or interactive games such as mobile apps and voice assistants. It may be possible to combine the best of digital and physical affordances to design social robots that can meaningfully augment social games. To this end, studies are needed to assess the impact of the physical presence and movement of robots in these contexts of use.

Social and creative games represent interesting settings to study the interaction with social robots as they create a space for playful semi-structured interactions. In many social games, clear rules exist but significant open-endedness is supported to exercise and enjoy the creativity of oneself and others. Creative games with open rules as Story Cubes have not been used often in HRI. However, games are a popular setup in HRI. For instance, Leite et al. (2009) used chess (to some extent a creative game) as a setup to understand how social presence of robots is perceived. Similarly, interactive storytelling in HRI has been reported as a promising scenario for children's social skills development (Leite et al., 2015, 2017). Storytelling games have the potential to support creative social interactions. We thus select social games, and particularly storytelling games such as the popular "Story Cubes" (Ros and Demiris, 2013; Eladhari et al., 2014; Bae et al., 2016; Gordon and Spierling, 2018) as the site of research for this study.

2.3 Displaying robot creativity

Creativity is an important component of social interaction (Rogers, 1954). Movement has communicative properties that make it an essential part of social interaction between humans (Goldman, 2004), and it is widely regarded as central to embodied experiences including everyday creativity (Svanæs, 2013). Social robots assisting develop creative capacities that have been proposed before demonstrating the importance of physical movement in this type of applications (Hoffman and Ju, 2014). Similarly, storytelling is a creative and social activity that has an entertainment value but is also used to support learning (Sadik, 2008) and health (Plaisant et al., 2000).

Kinesthetic creativity has been studied mostly in artistic performance by humans (Ros and Demiris, 2013; Tan et al., 2018). To our knowledge, this work represents an early approach to the study of kinesthetic creativity in social robots. The research methodology for this study is an experimental design based on previous studies as the ones performed by Salem et al. (2011); Hoffman and Ju (2014); Hoffman et al. (2015); Tung (2016). The effects of robot movement are thus evidenced by their perception of how movement is perceived as a cue, indicating creative agency in social robots.

2.4 Motivation for this exploratory study

In this study, we chose Story Cubes as an experimental setup. This is a game where players create or re-create stories to share with others, thus requiring to some extent creative skills found universally (Brent, 2014). Our study asks whether social robots will have an edge based on their physicality and their basic animacy features to engage in creative activities such as storytelling games. More specifically, it seeks to assess to what extent movement may make a difference in such settings. If the answer to this question is positive, then further work will be needed to identify and evaluate the kinetic principles for the design of playful and creative social robots more specifically. If the answer is negative, then it is more likely that screens and voice interaction devices will be more adequate and arguably even easier to develop, deploy, maintain, and operate in this type of activities rather than robots that use physical bodies to move and reinforce non-verbal interactions. To this end, our study addresses the research gap in how robot movement is perceived when they perform a creative storytelling activity.

In sum, with studies like this, we aim to contribute to the emergent exploration of human–robot creative interaction (HRCI). Thus, here we set to identify the hypothesized connection between the perceived creative agency and the animacy of social robots. Our goal is to evaluate the relevance of robot movements to attribute creativity to social robots. At this stage, we aim to provide a benchmark for future experiments using non-choreographed robot movements. Similarly, we use storytelling games supported by visual cues to study how movement shapes human's perception of creative agency in robots. We need to highlight that most social robots present limitations in terms of dexterity when manipulating objects on a human scale.

3 Methods

3.1 Research goals and questions

This study aims to assess the connection between perceived creative agency and the animacy of social robots. We evaluate the relevance of robot movements to how observers attribute kinesthetic creativity to social robots and the verbal delivery of a story as a creative act. The aim of this experiment is to explore the extent to which robots' movement supports the display of a creative act.

With this exploratory study, we aim to respond the following research questions without proposing any hypotheses due to the complex nature of the interaction and the exploratory nature of the study:

- (1) To what extent do humans perceive creative agency in robots when these tell a story as part of a game (display of a creative act)?
- (2) To what extent do humans perceive creative agency in robots when these display movements accompanying the delivery of the story?
- (3) How are robots perceived as creative agents compared with humans?

3.2 Materials and implementation

3.2.1 The robot

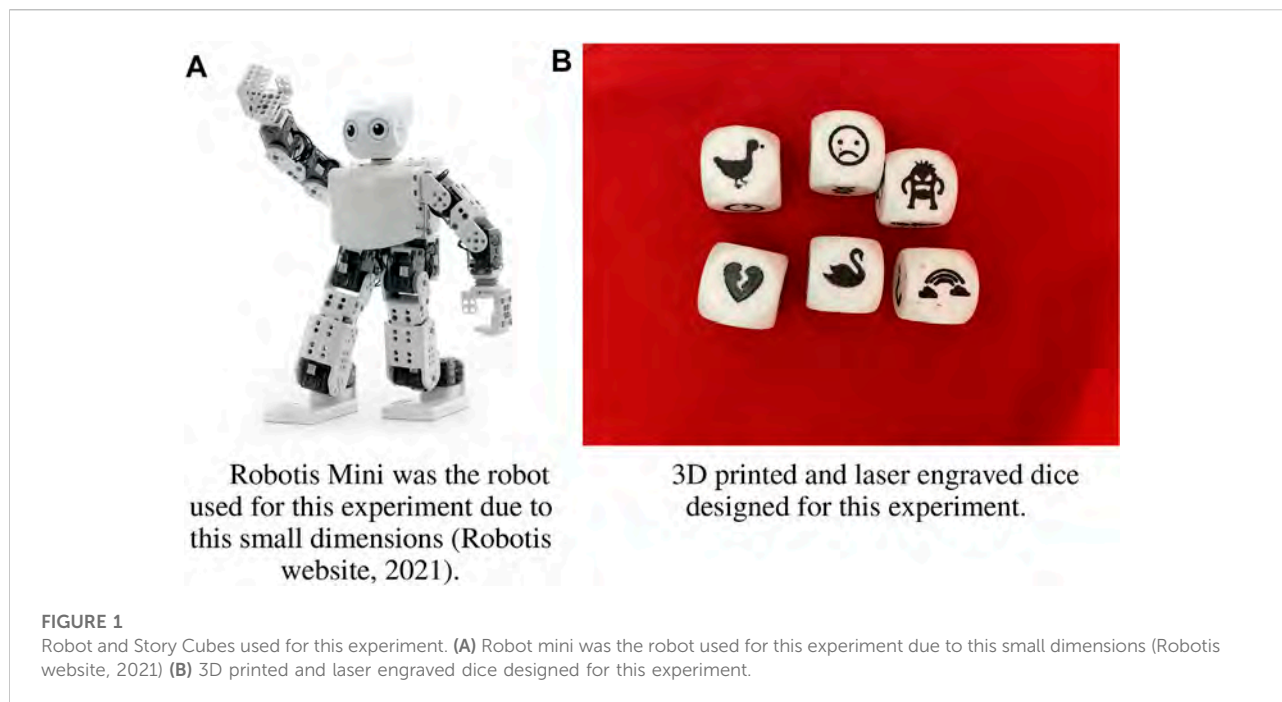
To evaluate the research questions, we programmed a Robotis Mini humanoid robot (See Figure 1A) to play our own version of Story Cubes (storytelling game) in a Wizard of Oz setup. The robot used a female voice in English language (Karen) generated in Mac OS 10.15.7 and played at slow speed (25%) using a bluetooth speaker next to the robot. We chose this robot due its small dimensions for transport and suitability to be used for future experiments using board games. The robot was animated using the Robotis Mini app in iOS 14.4. According to the work of Bernotat et al. (2021) and Kuchenbrandt et al. (2014), the robot design can lead to a male or female perception of the social robot. Hence, we decide to use a female voice to neutralize the possible male perception of the robot considering the very sharp and angular design of the mini humanoid. The robot movements were presented but not choreographed as they were just a stimulus to indicate robot animacy rather than supporting the delivery of the story.

3.2.2 The game

The Story Cubes game is a collaborative board game using six or more dice with adaptable rules and an indeterminate number of players. The goal of the game is to create a story with the contribution of all participants. We designed our Story Cubes for this experiment with a set of six oversized dice that could be visible in the video recordings. After five design iterations using different materials, we used a 30mm, solid white, PDA, 3D-printed dice, which was laser engraved and hand painted. We modified publicly available icons under Creative Commons license for our experiment. Thirty six icons were engraved and six icons were used to reveal “The Ugly Duckling” story for the two different versions used in the experiment.

3.2.3 The story

We use the well-known literary tale titled “The Ugly Duckling” by the Danish author Hans Christian Andersen (Andersen et al., 1995) and a modified version of that story with an alternative ending. Stories are an effective way to connect with other humans and social agents. The creative act of telling a story associated with the movement could possibly lead to significant different perceptions by users of the creative agency of robots. Similarly, a universal story as The Ugly Duckling helps us to reach a significant and diverse pool of participants for this experiment. Finally, the twist plot leading to



a different end of the story is a creative act performed by a robot, which could be or not perceived by the user.

3.3 Setup

The standard version of The Ugly Duckling was split in six short sections to match with the six icons shown in Figure 1B and two additional sections for Introduction and Wrap-up. The creative story was deeply discussed by the authors and validated by four experts in creative writing, English literature, and film scripts in its early and latest versions. The latest version uses dinosaur references, which were considered unexpected and perceived as creative without the controversy and the risk of unintentionally offend or hurt feelings of a particular community as the ethics committee suggested (Application HC200985). This last version was validated by one of the specialists in English literature. Both versions, namely, the standard and the creative stories can be read as follows:

Original Story:

-Ok, let us start. Once upon a time [Introduction].
 -There was a mama duck who was very surprised with one of her ducklings [1].
 -Everyone thought it was very ugly and rejected the poor little duckling [2–3].
 -She could not understand why everyone was so cruel only because she was different [3–4].
 But a year later, the “ugly duckling” grew into a beautiful swan [5]!

-Then she flew away with a flock of swans and every year returned to say hello to her foster mum [6].
 -She, her mum, and siblings celebrate with a happy party around the lake [7].
 -The moral of the story is that some people take longer to develop and find their true beauty [Wrap-up and Robot Reflection].

Creative Story (The bold font indicates the creative twist vs. the original story):

-Ok, let us start. Once upon a time [Introduction].
 -There was a mama duck who was very surprised with one of her ducklings [1].
 -Everyone thought it was very ugly and rejected the poor little duckling [2–3].
 -She could not understand why everyone was so cruel only because she was different [3–4].
-But a year later, the ‘ugly duckling’ grew into a beautiful flying dinosaur [5.1]!
-A pterosaurs called Quetzalcoatlus [5.2].
-She discovered that she was different, not ugly [5.3].
-Then she flew away with a flock of happy pterosaurs and every year returned to say hello to her foster mum [5.4].
 -Then she flew away with a flock of swans and every year returned to say hello to her foster mum [6].
 -She, her mum, and siblings celebrate with a happy party around the lake [7].
 -The moral of the story is that some people take longer to develop and find their true beauty [Wrap-up and Robot Reflection].

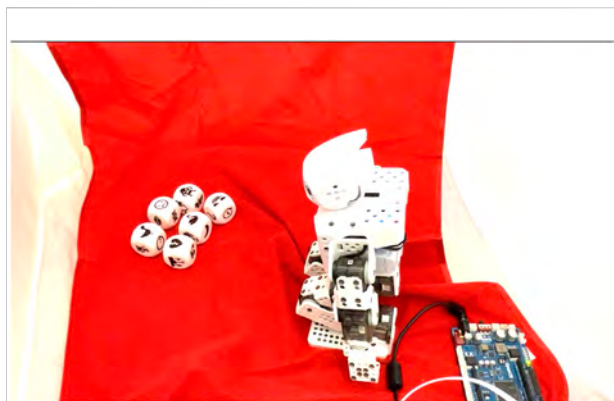


FIGURE 2

Snapshot of the video showed to the participants. The setup used under the four different conditions of the experiment is the same (cubes, robot, and microcontroller in the shot).

The robot, the dice, and the speaker (not visible) were allocated in a photo tent in order to isolate them from the external stimuli, avoid distractions for the users and make possible the replication of the experiment. A high contrast, red carpet was used to highlight the dice and the robot. After several tests, we decided to use the video, which was recorded in from the top-left corner of the photo tent using an iPhone 12 mini with 0.8x digital zoom. This setup allows to the viewers to have a full view of the robot and dice. We consider that using this point of view allows the user to understand the creation of the story by the robot. See [Figure 2](#). The length of the videos was shorter when the standard story was displayed due to the three extra sentences used in the creative story. These three extra sentences in the creative story aim to reinforce the novel character of the story and assure that the participant noticed the plot twist. This setup was implemented by the suggestions of the experts in writing and literature.

Four videos were displayed to the participants showing two different robots performing a creative task. The creative task consists of storytelling performed by the robot. The method to tell the story is supported by visual cues in the form of icons on dice and the robot manipulating them. Once the bowl covering the dice was removed, the robot started one of the proposed stories with the movements depending on the conditions.

In sum, we programmed a humanoid robot to tell the original version of the “The Ugly Duckling” and a modified version of the same story (creative story) under the experimental condition using cubes with visual imagery such as those used in the “Story Cubes” creative game. Under the control condition, the robot remains static while telling the story, while under the experimental condition the robot performs movements to accompany telling the story kinesthetically. These interactions were video recorded, and participants were requested to fill a survey evaluating the creative agency of the robot.

4 Experimental design

We designed a 2×2 between-subject online study. The factors are the story and robot movements. The robot can tell the standard story or the creative story (dinosaur plot twist), and the robot can display movements or not (still) during the storytelling. Hence, we test four conditions displayed by the robot: still robot and standard story (SS), still robot and creative story (SC), moving robot and standard story (MS), and moving robot and creative story (MC). As a between-subject study, the participants of this online study were exposed to one of the conditions mentioned earlier, that is, a participant under the condition MC would see a robot gesticulating and telling The Ugly Duckling story with the plot twist of the dinosaur. The dice and videos of this experiment can be requested contacting the corresponding author. The videos are unlisted in YouTube but can be reviewed using the next links: Video MS condition, Video MC condition, Video SC condition, and Video SS condition. See [Table 1](#).

4.1 Online survey and measurements

This study was approved by the human ethics committee of the University of New South Wales, application HC200985 reviewed by the HREAP Executive. Similarly, this study was funded using the Scientia Fellowship (PS-46183) development package provided by the UNSW. We implemented the survey on Qualtrics licensed for the UNSW. A survey flow was created to assign the participants under the four conditions randomly, see [Figure 3](#). The survey was distributed using prolific.co. The survey was designed as follows: First, the participant information and then the consent form. Once the participant agreed to participate in the studio, he/she was directed to the next page, and demographic information and confirmation of the Prolific ID was collected. The identity of the participants is anonymous, and demographic information was collected such as age, gender, occupation, location, and level of education.

Once the demographic information was collected, the Short Scale of Creative Self ([Karwowski et al., 2018](#)) questionnaire was applied. The questionnaire is confirmed by eleven questions measuring creative self-efficacy (CSE) and creative personal identity (CPI). The questions are as follow:

- (1) I think I am a creative person (CPI).
- (2) My creativity is important to who I am (CPI).
- (3) I know I can efficiently solve even complicated problems (CSE).
- (4) I trust my creative abilities (CSE).
- (5) Compared to my friends, I am distinguished by my imagination and ingenuity (CSE).

TABLE 1 Four experimental conditions. Each condition shows the factorial combination displaying a particular performance of the robot.

Experimental design	Standard story (control condition)	Creative story (experimental condition)
Still robot	Still standard condition (SS)	Still creative condition (SC)
Moving robot	Moving standard control (MS)	Moving creative condition (MC)

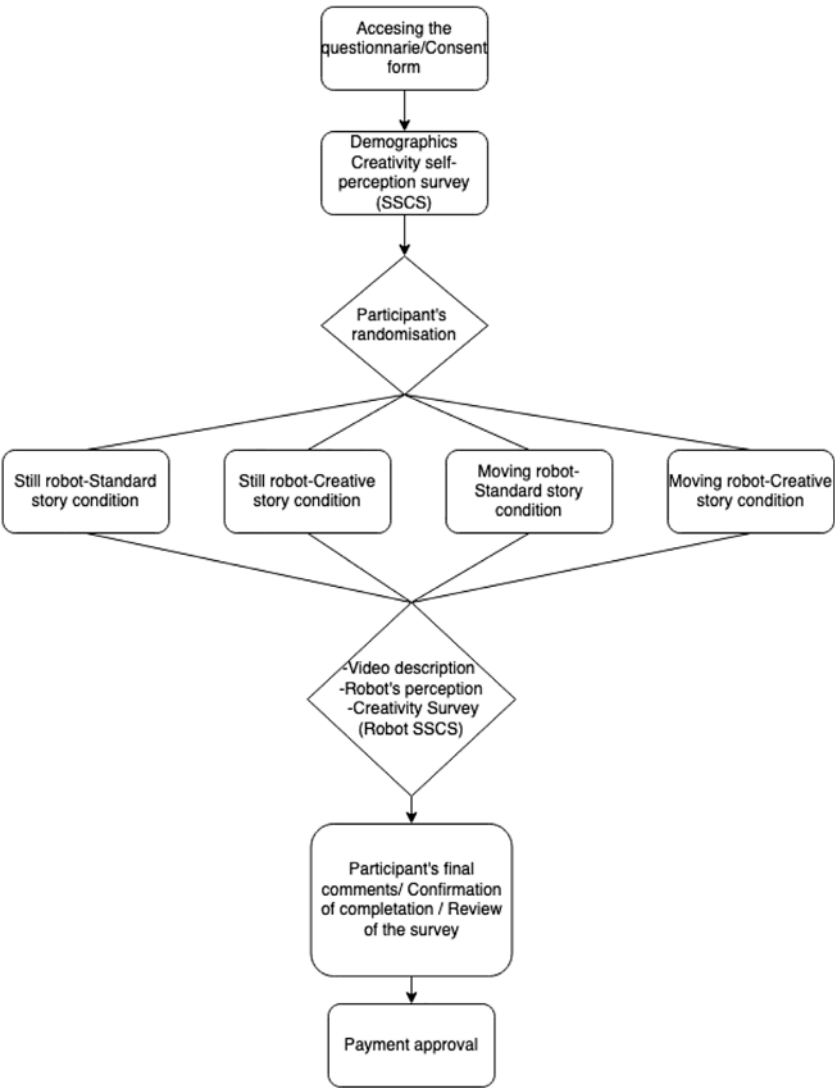


FIGURE 3 Experimental procedure. We applied the questionnaire using Qualtrics. We distributed it using Prolific.co.

- (6) Many times I have proven that I can cope with difficult situations (CSE).

(7) Being a creative person is important to me (CPI).

(8) I am sure I can deal with problems requiring creative thinking (CSE).
- (9) I am good at proposing original solutions to problems (CSE).

(10) Creativity is an important part of me (CPI).

(11) Ingenuity is a characteristic which is important to me (CPI).

Questions one, two, seven, ten, and eleven gauge CPI and questions three, four, five, six, eight, and nine are assigned to the CSE. Following the questions, the participants watched one of the four videos. We confirmed that the video was watched *via* a check button and requesting a brief description of the video by the participant. Next, the two first authors proposed a modified version of the SSCS to be applied to robots playing creative task. In this case, the task is framed in the playing of our version of Story Cubes. As the original SSCS, our Likert scale goes from one to five. The anchors are definitely not 1) and definitely yes (5). We will validate this adapted scale in a following article. The questions proposed are listed as follow.

- (1) I think that the way the robot played Story Cubes shows it is a creative robot.
- (2) The creativity of the robot playing Story Cubes is important to how it behaves.
- (3) The robot efficiently plays Story Cubes.
- (4) I trust the robot's creative abilities to play Story Cubes.
- (5) Compared to other players of Story Cubes, the robot is distinguished by its imagination and ingenuity.
- (6) I think that the robot can consistently create good stories when playing Story Cubes.
- (7) Being creative is important for a player of Story Cubes.
- (8) The robot can deal with problems requiring creative thinking.
- (9) The robot is good at proposing original stories in Story Cubes.
- (10) Creativity is an important part of how the robot plays Story Cubes.
- (11) Ingenuity is a characteristic which is important to how the robot plays Story Cubes.
- (12) I think that the robot can consistently create good stories when playing Story Cubes. Additional question:

In addition, we added a 12th question. The aim of this question is to summarize the general impression of the participant in just one score. The Godspeed questionnaire was applied after the SSCS and general impressions over the study were requested to the participant. The survey concludes confirming the end and providing a code that will allow the compensation for the participant by Prolific. The content of the survey is available by request to the corresponding author.

4.2 Participants

A total of 297 participants were recruited using the platform Prolific. All participants were over 18 years, and there were no restrictions on gender, formal education, income, or other demographics. Due to technical issues, just 242 participants were exposed to one of the four conditions and compensated by completing the

questionnaire. We used the data of 239 participants as the SSCS score as three of their data were not recorded. Participants were paid the equivalent of 1.27 British pounds (or 2.2 Australian dollars) for their participation. The first half of the study was run in February 2021, and the second part was 1 week later. One forty five participants were male, 89 female, three non-binary, and two did not specify. The average age was 26.9 years ($SD = 8.58$). Participants came from a range of locations; 38.5% from North America (Canada, United States, and Mexico), 34.7% from Europe, 13% from the United Kingdom, 7.5% from South America, 4.2% from Oceania, and 2.1% from Africa. The education levels are distributed as follows: 42.7% university degree, 38.5% high school, 9.6% masters, 1.7% PhD, 5.4% vocational education, 1.7% primary education, and 0.4% other. In terms of occupation, 42.3% were students, 38.9% employed (6.3% IT and software, 3.3% artist, 2.5% freelance, 1.3% researcher, and 25.5% other), 12.1% were unemployed, 2.5% homemakers, 0.4% retired, and 1.3% did not specify.

The average time to fill the survey was 9 min 20 s. We suggested to the participants to use a device with a large screen to have a better user experience and show the video and survey consistently; 72.8% used Windows 10, 8.4% used Macintosh, and 18.8% used other platform (Android 10 5.4%, iPhone 4.2%, Windows 6.1 3.8%, Windows 6.3 1.7%, Android 11 1.3%, Android 9 1.3%, Linux x86 0.8%, and Ubuntu 0.4%).

We aimed to allocate at least 60 participants per condition. We did not fully record data as three subjects lost in the SSCS score. Participants were randomly allocated as follows: (SS) = 62 participants (minus two missed SSCS scores), (SC) = 61 participants, (MS) = 61 participants (minus one missed SSCS score), and (MC) = 58 participants. The average human SSCS score was 3.89 ($SD = 0.66$), no significant differences were found among the different experimental conditions (ANOVA). Participant's results are stored in a standard online spreadsheet, and the statistical analysis was made using IBM SPSS.

5 Results

In order to address the exploratory questions of this study, we performed multiple 2×2 factorial analyses of variance (ANOVA); the factors were the story (standard vs. creative) and movements (still vs. moving). We defined seven dependent variables: the SSCS, the question twelve, and the five Godspeed items (anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety). In addition, we performed a Pearson's correlation present among the humans and robots CPI, CSE, and SSCS scores to check the internal consistency and possible human-robot significant correlations among the scores.

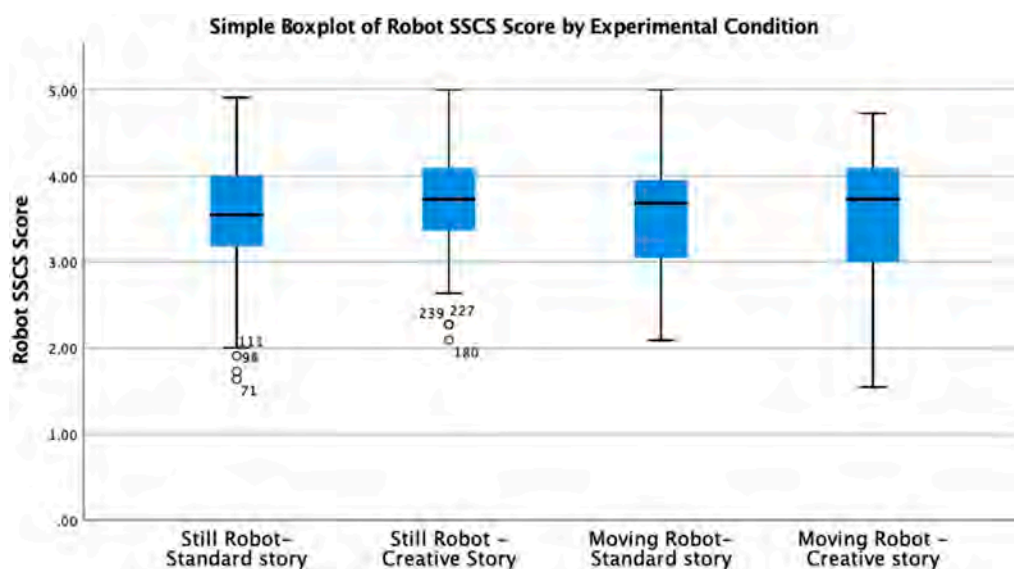


FIGURE 4

Boxplot per group of the SSCS applied to the robot with outliers. No significant differences were found among the means of the groups.

5.1 Short Scale of Creative Self (SSCS) score applied to the robots

This variable is our main indicator to assess how participants perceive robots as creative agents. The average SSCS score of the robots under the four different conditions are as follows: SS = 3.48 (SD = 0.74), SC = 3.66 (SD = 0.61), MS = 3.58 (SD = 0.68), and MC = 3.53 (SD = 0.80). As the original SSCS, our Likert scale goes from one to five. The anchors are definitely not 1) to definitely yes (5).

We ran a two-way ANOVA with the SSCS score as a dependent variable and the story and movements as factors. Residual analysis was performed to test for the assumptions of the two-way ANOVA. The assumption of homogeneity of variances was not violated, as assessed by Levene's test for equality of variances, $p = 0.122$. Data were normally distributed as assessed by the Kolmogorov-Smirnov test ($p = 0.200$). There were six outliers as assessed as being located less than 3 box-lengths from the edge of the box in a boxplot. They were not removed from the following ANOVA and showed that they did not affect the results even without them. Neither significant main effects nor interaction effects were found. See Figure 4. Similarly, pairwise comparisons were run aiming to find simple main effects. However, non-significant effects were found again.

Hence, as suggested in Laerd (2021), a further robust ANOVA was run. We used SigmaPlot to run a Kruskal-Wallis non-parametric ANOVA. Once again, neither main effect nor interaction effect was found. KW = 2.11, $df = 3$, and $p = 0.550$.

5.2 Question twelve and Godspeed items

A similar procedure was followed for the analysis of the rest of the dependent variables. The effects were found only for the items of animacy, likeability, and perceived safety. Data were normally distributed, as assessed by the Kolmogorov-Smirnov test for animacy ($p = 0.200$) and likeability ($p = 0.200$) but not for perceived safety ($p < 0.001$) and Q12 ($p < 0.001$). The assumption of homogeneity of variances was not violated for animacy (0.989), perceived safety ($p = 0.173$), and Q12 ($p = 0.909$) as assessed by Levene's test for equality of variances. However, it failed for likeability ($p = 0.006$).

We decided to maintain the outliers assessed as those being located less than 3 box-lengths from the edge of the box in a boxplot. The results indicate that movement has a main effect on how participants perceive the robots under the different conditions. Animacy $F(3,235) = 39.777$, $p < 0.001$, likeability $F(3,235) = 12.824$, $p < 0.001$, and perceived safety of the robot $F(3,235) = 19.127$, $p < 0.001$. Due to the violation of the assumption of homogeneity for likeability, a robust non-parametric Kruskal-Wallis ANOVA was run for this variable. Movement has a significant main effect. KW = 10.205, $df = 3$, and $p = 0.017$. See Table 2.

5.3 Person's correlations among CPI, CSE, and SSCS in humans and robots

A Pearson's 1-tailed correlation was carried out to assess the relationship between the human CPI, CSE, and SSCS scores and

TABLE 2 Bold font means this item had a significant main effect on the movement condition.

Means and SD for Question 12 and Godspeed items under the movement main effects condition	Still robot standard story	Still robot creative story	Moving robot standard story	Moving robot creative story
Q12	3.27 (1.06)	3.25 (1.06)	3.57 (0.98)	3.45 (1.08)
Anthropomorphism	2.16 (0.73)	1.99 (0.77)	2.27 (0.76)	2.15 (0.76)
Animacy	2.36 (0.76)	2.33(0.74)	3(0.68)	2.84(0.66)
Likeability	3.49(0.96)	3.61(0.96)	3.96(0.62)	3.89(0.73)
Perceived intelligence	3.48 (0.87)	3.47 (0.75)	3.67 (0.70)	3.45 (0.72)
Perceived safety	2.88(0.61)	2.74(0.37)	3.12(0.54)	3.10(0.56)

TABLE 3 Pearson's correlations among the CPI, CSE, and SSCS scores. The scores are internally consistent.

Pearson's Correlations among the CPI, CSE, and SSCS scores for humans and robots ($p < 0.001$)	Human CSE	Human CPI	Human SSCS	Robot CSE	Robot CPI	Robot SSCS
Human CSE	1					
Human CPI	0.573	1				
Human SSCS	0.883	0.891	1			
Robot CSE				1		
Robot CPI				0.673	1	
Robot SSCS				0.931	0.896	1

similar scores granted to the robot. We aimed to find if correlations between people perceiving themselves as highly creative project this on the robot's creative act. The scores of the 239 participants were analyzed. There were statistically significant, moderate, and strong positive correlations between CPI–CSE, SSCS–CSE, and SSCS–CPI in humans, and similar pattern of correlations among the robot scores suggesting consistency in our proposed scale and its internal scores. However, no significant correlation appeared between human scores and robot scores. We should highlight that overall the means of the human scores were higher than that of the robot scores under all the experimental conditions and all the scores. See [Table 3](#).

6 Discussion

This study aims to explore the design of future social robots using a quantitative approach and qualitative insights from the

users. With this study, we want to encourage a discussion in the domain of robot's perceived creativity and explore robot movement supporting the delivery of a creative act. However, studying creativity using quantitative approaches presents a number of challenges. We frame our findings in the context of games as a mean to sustain long-term, meaningful human–robot creative interactions and several considerations should be taken.

To answer the first research question: to what extent do humans perceive creative agency in robots when they tell a story as part of a game (display of a creative act)? We assured that participants watched the video requesting a checkbox validation and a brief description of what they saw. Participants frequently used the word “story” or even specifically “ugly duckling story” in these descriptions; 92.5% of the participants mentioned the word “story” or similar (tale, history, story line, and fable) when describing the video. In few cases typos were present and derivations such as history, story, or study were used but the intention was taken as the same. Furthermore, 100% of the

participants referred directly or indirectly to the act of telling a story even when they did not use the word in their descriptions. For instance, “it is about a robot that tells ugly duck that turns into swan,” “the process to find the real beauty,” or “It is about a robot that tells ugly duck that turns into swan.”

We highlight the fact that the SSCS scores of the robots, even when they are not significantly different, were above the mean (2.5) of the 1–5 scale of the SSCS score as indicated in section 5.1. Even though the robot scores were lower than the human scores, the minimal score was above 3.4 for the still robot standard story condition. The scores per condition were: SS = 3.48 (SD = 0.73), SC = 3.66 (SD = 0.61), MS = 3.58 (SD = 0.68), and MC = 3.53 (SD = 0.80). Hence, we can claim that participants were aware that the robot was performing a creative act delivering a story. Future work can involve further statistical analysis comparing with a specific benchmark, indicating a minimal score that a social agent considered for a creative agent. Similarly, a face-to-face setup could be more appropriate to perform an experiment of this nature since it is likely that the robot embodiment has a significant impact in the participant’s perceptions compared to virtual agents.

For the second question: to what extent do humans perceive creative agency in robots when these display movements accompanying the delivery of the story? We considered that robot movement would be a variable moderating the effect of the story on how people perceive robots. The marginal means graph can wrongly lead to conclude that movements moderate the SSCS score. However, when inspecting the boxplot, it is clear that means among all the experimental conditions are not significant. See 4. Although we did not notice main or interaction effects in the SSCS score, we did notice significant effects in three of the items of the Godspeed scale. These items are animacy, likeability, and perceived safety.

In the case of animacy, we observed that participants in this study could notice the movements of the robot as they score significantly higher in animacy to the moving robots independently of which story the robot is telling. This shows that participants are aware of the movement and how the movement impacts participant’s perceptions in terms of likeability as they rank moving robots MS = 3 (0.68) and MC = 2.84 (0.66) significantly higher than still robots SS = 2.36 (0.76) and SC = 2.33 (0.74). See Table 2.

The robot movements were not mentioned frequently in the description of the video as the story. However, some participants used anthropomorphic terms, that is, “a creepy robot gives its version of The Ugly Duckling by Hans Christian Andersen. It also does a decent MC hammer impression.”... The movements of the robot were however a bit erratic and did not match that well with the story it was telling. “In a way I feel like I was arranged to tell that story but I liked the movements and the appearance of the robot.”

In terms of likeability, as shown in Table 2, participants scored significantly higher to the moving robots in terms of

likeability. MS = 3.96 (0.62) and MC = 3.096 (0.62) significantly higher than still robots SS = 3.49 (0.96) and SC = 3.61 (0.96). The significant main effect of the movements in likeability is aligned with previous studies using games (Sandoval et al., 2016b; a, 2020). Apparently, humanoid robots tend to be likeable when they perform unexpected tasks that can be interpreted as social or creative. Future studies could test other robot embodiment perceived as less humanoid. An illustrative comment in how some participants perceive the robot was: “how this robot looks like and how it feels like compared to a human being. By his movements, he looked really happy, but by talking only sometimes we cannot understand how smart a robot can be. He was very smart, and we can definitely see that his voice was not recorded at some point.”

At the beginning of this experiment we considered that the factor of movement would be a moderating variable, supporting the delivery of the story by the robot. In other words, robot movement would lead to higher scores for the robots in both kinds of stories or at least in one of the stories. However, evidence for this were not registered, a possible reason for that is the type of the robot’s movement. For this experiment, we intentionally designed a set of robot movements that are part of the standard programming of the robot but are not synchronized with the delivery of the story. The reason for this is that in future applications using robots for social games, it is unlikely that robot movements will always be customized to their dialogs. Better synchronized and choreographed movements would be an obvious stimulus that could cause significant main effects in the SSCS score and how people perceive robots as creative agents. This type of movements can be tested in future studies. In terms of perceived safety, moving robots were perceived as safer than still robots. This is a result worth considering further, especially taking into account that this study is an online experiment and not a face-to-face setup.

Question twelve in the survey: “I think that the robot can consistently create good stories when playing Story Cubes” was added to allow the participants to summarize their impressions from the previous SSCS questions. The movement main effect was not significant ($p = 0.64$) when the participants answered this question. However, this provides an intriguing result to be explored in future studies as the comments of the participants suggest. Certainly, the participants perceived the movement of the robot but not necessarily the novelty of the story. Independently of the story, participants rated slightly higher for the moving robots than still robots as agents, which can create good stories when playing Story Cubes. Even as a marginal result, this finding points to the importance to study a range of robot movement approaches in future work. Further qualitative analysis is required to explore.

For question three in the survey: how are robots perceived as creative agents compared with humans? In all the experimental conditions robots scored lower than humans for the SSCS scores. Robot scores can be seen in Section 5.1 and human SSCS scores

are as follow $SS = 3.88$ ($SD = 0.68$), $SC = 3.85$ ($SD = 0.83$), $MS = 3.88$ ($SD = 0.70$), and $MC = 3.95$ ($SD = 0.66$). Pearson correlation does not suggest any significant correlation among the human and robot scores. However, there are significant correlations among the sub-scores CPI and SCE in both human and robot scores, which could suggest that the internal consistency of the original scale and our adaptation for assessing robot creativity have potential as measurement tools for future studies in human–robot creative interactions. See [Table 3](#).

Some of the participants' comments suggest that a further exploration of this question could be relevant as they were enthusiastic about the capabilities of the robots comparing with previous human creative experiences. To illustrate, first of all, the participants' adored the robot. they thought it was cute plus they were impressed with its abilities, too! they meant they can probably kind of guess how it works, but still exclaimed that it was just mind-blowing! The participants' loved that the message of the story told by the robot was so wholesome! As a musician and somewhat a songwriter, they find it astonishing that how it can come up with a good story in such a quick amount of time, and they wished keep it up. As for the study itself, they liked the way the text lighted up when the mouse hovered over it, they have not seen it a lot, if anytime. In addition, they had to check up on two of the English words used in the study, which they highly value as an educational feature. They thoroughly enjoyed the experience. "This was an interesting concept to consider, and the participants' would honestly like to see more content involving AI and Story Cubes." They are impressed how robot can tell people a story based on random images.

7 Conclusion

Creativity can be considered an aspect of autonomy and agency in social agents that is different from intelligence, logic, and strategy. The current understanding on how creativity is displayed by robots is still limited. This study aimed to inform the design possibilities of human–robot creative interaction (HRCI) and provides a reference for future studies exploring the main factors involved in the creative interaction between humans and robots. Our findings show that the setup used in this study does not trigger higher scores in the SSCS, differentiating the robots as creative agents. However, movement does show main effects in the scores of animacy, likeability, and perceived safety of the Godspeed scale. Furthermore, the scores of the moving robots were above the media in all the cases (although they were lower than the SSCS scores in humans). In terms of how robots are perceived as good storytellers (question 12 in the robot SSCS), even when the scores are not significantly different, the results provide an important insight in how to continue the development of future experimental studies, for instance, the need to perform similar study in a face-to-face setup and using other robot embodiments beyond humanoid robots.

The study of creativity in robots shows a research gap when addressed in playful, creative, and collaborative activities such as board games. We chose a playful task (a storytelling game) to empirically evaluate the extent to which a robot's physical embodiment may cause humans to attribute creative agency to a robot. The Story Cubes game offers a means to further assess the display of creativity in robots considering the applicability of this setup as entertainment and for the development of cognitive skills, spacial memory, decision making, and collaborative skills ([Wu et al., 2012](#); [Unbehaun et al., 2019](#)).

Furthermore, we consider that our approach is useful in aiming strategies for long-term interaction and as an alternative to avoid screen addiction and contribute to a better mental health in the digital age ([Aboujaoude and Starcevic, 2015](#); [Sandoval, 2019](#)). Even in the Story Cubes mobile app, the user experience is visibly compromised when compared to the physical cubes. Considering this, we highlight the importance of perceived creativity in social robots to further develop the early work in artificial creativity. It looks like it is critical to explore advantages and disadvantages among robotic interfaces that display and support creative interactions. To this end, when users are exposed to stimuli related to creative robots it seems critical to set their expectations in this type of studies. One participant said, for instance: "the study itself was fine. The premise, however, is something that has not been fully explained." For example, is this a prototype of a children's toy? Is it a learning device? Is it a diagnostic tool?" Finally, one of the comments of the participants is encouraging to continue the development of studies in creative robots playing storytelling games. Assuming the robot really came up with the story using its own creativity and it was not programmed into it, the participant was very impressed with the level of depth which the story had. In that sense, the robot could even be more creative than a lot of humans. The participant also thinks that its use of words and its manner of speech is properly human-like, that is not to say, we could not feel the robotic nature behind it at all. In my opinion, some work should be performed on the robot's movement, and how it connects to whatever it is saying in a way that makes more sense. The participant wished us good luck with the study and with further development of creative robots."

7.1 Limitations and future work

The main technical limitations to implement a robot board game have been discussed before by [Sandoval et al. \(2021\)](#). The game Story Cubes in particular has a significant element of improvisation and randomness, and the translation of the cubes to create a consistent story is a technical challenge for robots that genuinely synthesize stories from these stimuli. Similarly, the vision system required to read the cubes accurately under a range of lighting conditions and angles

would require significant technical work. We are currently working on an implementation where participants and robots play Story Cubes in a shared physical setup for a future study. Choreographic movements of the robots will be programmed and displayed on a bottom-up strategy that starts by incorporating movements in an increasing level of sophistication and detail. Then, we will compare with the random movements of this experiment. We consider that this would inform robot designers to incorporate movement that achieves a balance between creating a meaningful human–robot creative interaction (HRCI) while drawing from a library of gestures, postures, and body movements suitable for fluid communication. In terms of data collection and analysis, we plan to conduct a thematic analysis of participants' comments. Furthermore, future experimental designs could include different robot embodiments (humanoid vs. non-humanoid), variants of Story Cube games, and different stories (original stories vs. well-known stories). Finally, a more exhaustive validation of our version of the SSCS (performing factorial and reliability analysis) may be required to assess future interaction face-to-face setup in a more robust manner.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#); further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by UNSW ethics committee, application HC200985 reviewed and approved by the HREAP Executive. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author Contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

References

- Aboujaoude, E., and Starcevic, V. (Editors) (2015). *Mental health in the digital age: Grave dangers*. illustrated edition edn (Great Promise: Oxford University Press).
- Alter, A. (2018). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. reprint edition edn. USA: Penguin Books.

Funding

This project was funded by the UNSW Scientia Fellowship PS-46183-A. This fund pays for the open access publication fees, equipment, participant's remuneration, and the rest of the research expenses associated with this study.

Acknowledgments

Sandoval wants to thank to Prof Sosa for all the intense and highly collaborative discussions that lead to this article and the rest of the authors for their valuable contributions. Also thanks to Prof Alfredo Jimenez (IPN-CEPROBI) for his advice on selecting and running additional statistical analysis. Thanks to Karam Hussain at the UNSW Makerspace Paddington campus for his support in the development of the Story Cubes used in the experiments. Also thanks to the story reviewers Carolina Posadas, Jennifer Breakelaar, and Miranda Verswijvelen. Sandoval wants to thank Sebastian for inspiring him to finish the first draft of this manuscript one day before his birth.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frobt.2022.695162/full#supplementary-material>

- Alves-Oliveira, P., Arriaga, P., Cronin, M. A., and Paiva, A. (2020). "Creativity encounters between children and robots," in Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction (New York, NY, United States: Association for Computing Machinery), 379–388.

- Alves-Oliveira, P., Arriaga, P., Paiva, A., and Hoffman, G. (2017). "YOLO, a robot for creativity: A Co-design study with children," in *Idc '17: Proceedings of the 2017 conference on interaction design and children* (New York, NY, USA: Association for Computing Machinery), 423–429. doi:10.1145/3078072.3084304
- Andersen, H. C., Letts, K., and Ello, C. (1995). *The ugly duckling*. London, England: I. E. Clark Publications.
- Animacy (2021). *Animacy*, n. Oxford University Press.
- Bae, B.-C., Seo, G., and Cheong, Y.-G. (2016). "Towards procedural game story creation via designing story cubes," in *International conference on interactive digital storytelling* (Springer), 399–402.
- Bainbridge, W. A., Hart, J. W., Kim, E. S., and Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.* 3, 41–52. doi:10.1007/s12369-010-0082-7
- Bernotat, J., Eyssel, F., and Sachse, J. (2021). The (Fe)male robot: How robot body shape impacts first impressions and trust towards robots. *Int. J. Soc. Robot.* 13, 477–489. doi:10.1007/s12369-019-00562-7
- Brent (2014). *Creative story telling with story cubes*. London, UK: EdTech.tv.
- Castro-González, I., Admoni, H., and Scassellati, B. (2016). Effects of form and motion on judgments of social robots animacy, likability, trustworthiness and unpleasantness. *Int. J. Human-Computer Stud.* 90, 27–38. doi:10.1016/j.ijhcs.2016.02.004
- Desmurget, M. (2020). *La fábrica de cretinos digitales: Los peligros de las pantallas para nuestros hijos*. Madrid: Ediciones Península.
- Digital Humans, Uneeq Ltd (2021). *Digital humans: Conversational AI solutions beyond just chatbots*. Auckland, New Zealand: Ponsonby.
- Eladhari, M. P., Lopes, P. L., and Yannakakis, G. N. (2014). "Interweaving story coherence and player creativity through story-making games," in *International conference on interactive digital storytelling* (Springer), 73–80.
- Gizzi, E., Castro, M. G., and Sinapov, J. (2019). "Creative problem solving by robots using action primitive discovery," in 2019 joint IEEE 9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob) (IEEE), 228–233.
- Goldman, E. (2004). *As others see us: Body movement and the art of successful communication*. New York: Psychology Press. Google-Books-ID: mGIK5sQOVgC.
- Gordon, A. S., and Spierling, U. (2018). "Playing story creation games with logical abduction," in *International conference on interactive digital storytelling* (Springer), 478–482.
- Harris, J., and Sharlin, E. (2011). "Exploring the affect of abstract motion in social human-robot interaction," in *2011 ro-man* (Atlanta, GA, USA: IEEE), 441–448. doi:10.1109/ROMAN.2011.6005254
- Herrera, F., Oh, S. Y., and Bailenson, J. N. (2020). Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence. (Camb.)* 27, 163–182. doi:10.1162/pres_a_00324
- Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., et al. (2015). "Robot presence and human honesty: Experimental evidence," in 2015 10th ACM/IEEE international conference on human-robot interaction (HRI) (Portland, OR, USA: IEEE), 181–188.
- Hoffman, G., and Ju, W. (2014). Designing robots with movement in mind. *J. Hum. Robot. Interact.* 3, 89. doi:10.5898/JHRI.3.1.Hoffman
- Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychol. Aesthet. Creativity, Arts* 8, 62–70. doi:10.1037/a0034898
- Karwowski, M., Lebeda, I., and Wiśniewska, E. (2018). Measuring creative self-efficacy and creative personal identity. *Int. J. Creativity Problem Solving* 28, 45–57. Place: Republic of Korea Publisher: Korean Assn for Thinking Development.
- Kinaesthesia (2021). *Kinaesthesia*, n. Oxford University Press.
- Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., and André, E. (2014). Keep an eye on the task! How gender typicality of tasks influence human-robot interactions. *Int. J. Soc. Robot.* 6, 417–427. doi:10.1007/s12369-014-0244-0
- Laerd, S. (2021). *Two-way ANOVA in SPSS statistics*. London, UK: Laerd Statistics Premium.
- Leite, I., Martinho, C., Pereira, A., and Paiva, A. (2009). "As Time goes by: Long-term evaluation of social presence in robotic companions," in RO-MAN 2009 - the 18th IEEE international symposium on robot and human interactive communication (Toyama, Japan: IEEE), 669–674. doi:10.1109/ROMAN.2009.5326256
- Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., et al. (2015). "Emotional storytelling in the classroom: Individual versus group interaction between children and robots," in *Hri '15: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (New York, NY, USA: Association for Computing Machinery), 75–82. doi:10.1145/2696454.2696481
- Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., et al. (2017). Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Front. Robot. AI* 4. doi:10.3389/frobt.2017.00029
- Lindblom, J., and Alenljung, B. (2015). "Socially embodied human-robot interaction: Addressing human emotions with theories of embodied cognition," in *Handbook of research on synthesizing human emotion in intelligent systems and robotics (IGI global)* (Pennsylvania, USA: IGI Global), 169–190.
- Miller, M. N. (2021). *Love, Death and Robots season 2, episode 1 recap*. Scotts Valley, California: Automated Customer Service.
- Myoo, S. (2019). Creative robots. *Studia de Arte Educ.* 14, 30–39. doi:10.24917/20813325.14.3
- OpenAI (2021). *OpenAI*. San Francisco: OpenAI.
- Pham, T. X. N., Hayashi, K., Becker-Asano, C., Lacher, S., and Mizuuchi, I. (2017). "Evaluating the usability and users' acceptance of a kitchen assistant robot in household environment," in 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN) (Lisbon, Portugal: IEEE), 987–992. doi:10.1109/ROMAN.2017.8172423
- Paisant, C., Drui, A., Lathan, C., Dakhane, K., Edwards, K., Vice, J. M., et al. (2000). "A storytelling robot for pediatric rehabilitation," in *Assets '00: Proceedings of the fourth international ACM conference on Assistive technologies* (New York, NY, USA: Association for Computing Machinery), 50–55. doi:10.1145/354324.354338
- Rogers, C. R. (1954). Toward A theory of creativity. *ETC A Rev. General Semant.* 11, 249–260.
- Ros, R., and Demiris, Y. (2013). "Creative dance: An approach for social interaction between robots and children," in *International workshop on human behavior understanding* (Springer), 40–51.
- Sadik, A. (2008). Digital storytelling: A meaningful technology-integrated approach for engaged student learning. *Educ. Technol. Res. Dev.* 56, 487–506. doi:10.1007/s11423-008-9091-8
- Salem, M., Rohlfing, K., Kopp, S., and Joubin, F. (2011). "A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction," in *2011 ro-man* (Atlanta, GA, USA: IEEE), 247–252. doi:10.1109/ROMAN.2011.6005285
- Sandoval, E. B. (2019). "Addiction to social robots: A research proposal," in 2019 14th ACM/IEEE international conference on human-robot interaction (HRI) (Daegu, Korea (South): IEEE), 526–527. doi:10.1109/HRI.2019.8673143
- Sandoval, E. B., Brandstatter, J., Yalcin, U., and Bartneck, C. (2020). Robot likeability and reciprocity in human robot interaction: Using ultimatum game to determinate reciprocal likeable robot strategies. *Int. J. Soc. Robot.* 13, 851–862. doi:10.1007/s12369-020-00658-5
- Sandoval, E. B., Brandstetter, J., and Bartneck, C. (2016a). "Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction," in 2016 11th ACM/IEEE international conference on human-robot interaction (HRI) (Christchurch, New Zealand: IEEE).
- Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016b). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *Int. J. Soc. Robot.* 8, 303–317. doi:10.1007/s12369-015-0323-x
- Sandoval, E. B., Shi, J., Cruz-Sandoval, D., Li, B., Cappuccio, M., and Rosenbaum, S. (2021). "A prototype of a robot memory game: Exploring the technical limitations of human-robot interaction in a playful context," in *Hri '21: Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (New York, NY, USA: Association for Computing Machinery), 195–199. doi:10.1145/3434074.3447158
- Saunders, R., Chee, E., and Gemeinboeck, P. (2013). "Evaluating human-robot interaction with embodied creative systems," in *Proceedings of the fourth international conference on computational creativity* (Sydney, Australia: IEEE), 205–209.
- Sosa, R., Lee, J. B., Albarran, D., and Otto, K. (2015). "From concept to specification maintaining early design intent," in *ICoRD'15 - research into design across boundaries volume 2*. Editor A. Chakrabarti (New Delhi, India: Springer, Smart Innovation, Systems and Technologies), 445–457. doi:10.1007/978-81-322-2229-3_38
- Svanæs, D. (2013). "Interaction design for and with the lived body: Some implications of merleau-ponty's phenomenology," in *ACM transactions on computer-human interaction (TOCHI)* (New York, NY, USA: ACM), 1–30.

Tan, H., Wang, D., and Sabanovic, S. (2018). "Projecting life onto robots: The effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism," in 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN) (Nanjing, China: IEEE), 129–136. doi:10.1109/ROMAN.2018.8525584

Tung, F.-W. (2016). Child perception of humanoid robot appearance and behavior. *Int. J. Human-Computer Interact.* 32, 493–502. doi:10.1080/10447318.2016.1172808

Unbehaun, D., Aal, K., Carros, F., Wieching, R., and Wulf, V. (2019). "Creative and cognitive activities in social assistive robots and older adults: Results from an exploratory field study with pepper," in Proceedings of the 17th European conference on computer-supported cooperative work-demos and posters (Salzburg, Austria: European Society for Socially Embedded Technologies).

Van Breemen, A. J. N. (2004). "Bringing robots to life: Applying principles of animation to robots," in Proceedings of shaping human-robot interaction workshop held at CHI (Vienna, Austria: Citeseer), 143–144.

Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., Sandini, G., et al. (2017). Detecting biological motion for human-robot interaction: A link between perception and action. *Front. Robot. AI* 4. doi:10.3389/frobt.2017.00014

WHO (2018). *Addictive behaviours*. New York, USA: Gaming disorder.

Won, A. S., Bailenson, J. N., Stathatos, S. C., and Dai, W. (2014). Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *J. Nonverbal Behav.* 38, 389–408. doi:10.1007/s10919-014-0186-0

Wu, Y.-H., Fassert, C., and Rigaud, A.-S. (2012). Designing robots for the elderly: Appearance issue and beyond. *Archives gerontology geriatrics* 54, 121–126. doi:10.1016/j.archger.2011.02.003

Support Publication #9 [SP9]

Sandoval E.B., Shi J., Sandoval D.C., Li B., Cappuccio M.L., Rosenbaum S. (2021),
“A Prototype of a Robot Memory Game: Exploring the Technical Limitations of Human
Robot Interaction in a Playful Context” In *Companion of the 2021 ACM/IEEE International
Conference on Human-Robot Interaction* (HRI '21 Companion), March 8–11, 2021,
Boulder, CO, USA. ACM, New York, NY, USA, 5 pages.

<https://doi.org/10.1145/3434074.3447158>

Reproduced with permission from IEEE. All rights reserved.

A Prototype of a Robot Memory Game

Exploring the Technical Limitations of Human Robot Interaction in a Playful Context

Eduardo Benítez Sandoval
e.sandoval@unsw.edu.au
University of New South Wales
School of Arts and Design
Sydney, NSW, Australia

Jack Shi
JackS3214@yahoo.com.au
Independent Researcher
Sydney, NSW, Australia

Dagoberto Cruz-Sandoval
dagoberto@cicese.edu.mx
CICESE
Ensenada, BC, Mexico

Binghao Li
binghao.li@unsw.edu.au
University of New South Wales
School of Minerals and Energy
Resources
Sydney, NSW, NSW, Australia

Massimiliano Cappuccio
m.cappuccio@unsw.edu.au
University of New South Wales
School Of Engineering And
Information Technology
Canberra, ACT, ACT, Australia

Simon Rosenbaum
s.rosenbaum@unsw.edu.au
University of New South Wales
School Of Psychiatry
Sydney, NSW, Australia

ABSTRACT

We explored the feasibility and limitations of designing and developing a Robot-Human interactive board game known as memory, a turn-based game of matching card pairs. Our analysis of this case study suggests significant limitations and further interactive improvements before exposing the prototype to the users. In terms of technical limitations, the variability of light and the lack of sharp camera imaging makes it challenging to identify cards uniquely. Open MANIPULATOR-X's (robotic arm used) dexterity is limited and could not mimic the interaction of card flipping. For interactive design terms, we analysed the robot morphology, expressiveness and modifications in the cards. We suggest running comparative studies with well-known humanoid robots and humans. This project is the initial step to developing more engaging and interactive games between Humans and Robots. Future experiments aim to explore the emotional, physical and mental benefit users could obtain from playing games with robots.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design theory, concepts and paradigms**; *Empirical studies in interaction design*; *HCI theory, concepts and models*.

KEYWORDS

Robot Games, Memory Game, Robot Expressiveness, Robot Morphology

ACM Reference Format:

Eduardo Benítez Sandoval, Jack Shi, Dagoberto Cruz-Sandoval, Binghao Li, Massimiliano Cappuccio, and Simon Rosenbaum. 2021. A Prototype of a Robot Memory Game: Exploring the Technical Limitations of Human Robot Interaction in a Playful Context. In *Companion of the 2021 ACM/IEEE*

International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA. ACM, New York, NY, USA, 5 pages.
<https://doi.org/10.1145/3434074.3447158>

1 INTRODUCTION

Games have been a relevant activity of the human experience since prehistorical times because they allow amusement and socialisation. There is an extensive catalogue of games played by people [2]. The artificial conflict, the consistency of the rules, tangible and immediate outcomes allow the human players to engage in a social activity with their peers. Games also contribute to the generation of mental models that allow users to develop social and cognitive abilities useful in the interaction with other humans in different contexts [13].

Games are broadly used as a research instrument in Human-Robot Interaction. A search in Scopus shows 969 papers with the words "robot" and "game" used as keywords. Decision Games, Serious Games, Exertion games, and gamification have been used in the last years in HRI to investigate different variables. To illustrate, Decision Games, as Prisoner's Dilemma and Ultimatum Game have been used to investigate variables related to robot's perception as likeability [22], reciprocity [21, 24], and bribery [23]. Adaptations of fun games have been used to investigate collaboration [19], and robot learning [3]. Similarly, different robots (NAO, AIBO, self-made robots) have been programmed to play games and cards have been used to encourage robots interactivity with humans [7, 8]. Besides, there are papers and patents describing games like Chess, Go, and Checkers played with robots [16, 17]. However, research crafting the user-experience when humans and robots play board games aiming engagement has been less explored in HRI. Our research aims to design a user experience in a playful context that triggers repeated interaction and benefits users in cognitive, emotional and social ways. To achieve this goal, we assess this low-fidelity prototype's technical feasibility as the first step towards more sophisticated HRI. Similarly, we analyse the functional and usability aspects of this prototype. We also consider that further development of robot games can be a suitable tool for mental health therapy and cultivate virtue, as is suggested in [5].

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8290-8/21/03...\$15.00
<https://doi.org/10.1145/3434074.3447158>

In this paper, we explore a memory game considering its universality and accessibility to be played by a high number of users. Our interpretative analysis suggests that significant technical limitations should be beaten to reach the interactivity level that engages the user in a naturalistic HRI.

2 MOTIVATION, PHILOSOPHY AND RELATED WORK

The first motivation to develop this prototype is to explore and understand the differences and possible advantages of social robots compared with video games and traditional board games. Usually, games are used as a research instrument in HRI as a means to explore other topics. However, to our knowledge, the research aiming to design repeatable, and engaging games in HRI intentionally looking to generate amusement in the user is limited. To illustrate, there is some pioneering work (with certain limitations) in role-playing game combining interactive storytelling and interactive table surfaces in [14] and a robotic arm playing chess using a regular chess board [17]. However, the goal of these studies is not to study the amusement of the interaction per se.

The second motivation is to explore the future possible cognitive, physical and social benefits for the users when playing with robots in uncontrolled environments. We look in the long term to offer an engaging, spontaneous, interactive robot experience comparable with the experience of playing video games and board games with similar benefits [10–12].

The third motivation is to be aware of the technical challenges aiming a robust HRI outside of the experimental room. Social Robots playing games should be understandable and engaging for the user. This type of game experience requires reliable hardware and software. Our approach to achieving these interactive goals is to frame the design of the HRI focused in just one-task; in this case, a game that is well-known and simple as the memory game.

The philosophy behind the prototype relies upon the premise that games offer possibilities to generate engaging interactions with the users. Caillois defined games by six characteristics proposed in [4]. Games have to be: fun, separate (specific in time and place), uncertain, non-productive (the game is the incentive by itself), governed by rules, and fictitious (different reality). Similar heuristics can be applied to the design of robot games[15]. We aim to design future robot games that display the characteristics suggested by Caillois and develop and the mentioned cognitive, physical and social skills in the users with this prototype.

3 DESIGN OF THE PROTOTYPE

We choose a memory game as a setup for this prototype. We analyse this game's mechanics to keep a simple but fun interaction and encourage the users to replay the game on another occasion. Aside from the Caillois heuristics[4], we consider that a simple game as memory game can keep some of the LeBlanc's Taxonomy of Game Pleasures [9]. These are Sensation, Narrative, Challenge, Fellowship with the robot, a balanced symmetric game, and a flow channel with a similar pattern that the memory game played with other humans [25, 26]. However, the central goal of this paper is to report the technical aspects of this early prototype.

3.1 Robot setup

The prototype's primary purpose of using the robot arm is to test the technology required to detect and select the memory game cards. The robot used is an Open MANIPULATOR-X from Robotis. The robot was programmed using the libraries provided by the company in the Arduino IDE to display movements. The robot displays minimal interactive features and points to the cards using a low-intensity laser. We programmed the game in Python. The artificial vision system programmed detects the cards by the colour code HSB (Hue, Saturation and Brightness). The robot is capable of playing with three sets of cards (six cards). The vision system was programmed in Python to be usable in different platforms using Open CV libraries. The identification of the cards was made using background colour as the primary cue. The selection of the colour was selected manually making the system compare with the database. A standard Logitech C920 webcam with HD 1080p in a tripod was used in a top view to detect the cards during the game.

3.2 Memory Game

We use a memory game. The goal of the game is to match cards and make as many pairs as possible alternating turns. This kind of games are simple and universal; they are known as "concentration", "Match-up", "Match-Match", or Shinkei-suijaku. This specific memory game displays and ancient iconography with written descriptions in a Native American language. The use of these icons adds complexity to the game due to the intricate designs and unfamiliarity. We modify the cards' back to easily detect the card (high contrast with the background). In the preliminary test with no team members, we determined that games should have a duration of about three minutes (approximately ten sets of cards) to maintain the user's attention and game flow. See Fig. 1



Figure 1: Cards used in the game using icons from a Native American culture.

3.3 Game play

The robot dexterity is limited at the moment; it cannot manipulate cardboard cards. Users have to assist the robot in flipping the cards once it points out the cards. A laser pointer was attached to the robot to point out the cards before flipping them. A video attached to this paper shows the prototypes of the robot memory game. The

video can be reach in <https://youtu.be/WhXqTyiX7J0>. Once the system is further developed, the code will be deposited in an open-access repository to share the project with the research community. See Figure 2.



Figure 2: The first test of the Robot Game testing the technical implementation and observing the playability.

3.4 Preliminary solutions for the low-fidelity prototype

Initially, We planned to read the words on the card using text recognition. However, the card's text was illegible from the camera's image because we had to take in the board in one shot with the camera. we could use text recognition in the future using a high-quality camera. We started with artificial room light, but as the cards were glossy, we had intense reflection from them. Hence, we switched to a room with natural light and use mate paper for the cards. Slight differences in card colour were found as the cards were placed in different locations on the board, and the light did not hit them uniformly. Then, we introduced colour ranged matching. Cards of slightly different shades of the same base colour could not be used because the colour ranged matching would identify them as the same. The natural light differed from time of day, which meant the card colour's brightness could change. Hence, we switched to HSL (Hue, Saturation, Light) colour matching for a better match instead of using RGB colour matching.

At this stage, the prototype is just focused on the game's technical implementation and its feasibility. Considerations of interactive variables involved in the HRI will be taken once we assess this initial prototype's robustness.

4 METHOD AND EVALUATION

An interpretative analysis [28] of this case study was performed by this multidisciplinary team to assess the feasibility and robustness of the low-fidelity prototype. This kind of prototypes are common in the design domain [27]; however, low-fidelity prototypes are not reported very often in HRI. The multidisciplinary approach from the design, philosophy, computer science and engineering, and health (mental and cognitive) perspectives were considered for the analysis. In other words, the authors provided qualitative insights to acknowledge the limitations of the prototype and spot future improvements for a better interactive robot game.

4.1 Technical Limitations and impact in the interaction

The main limitation in developing this robot game is the interaction in uncontrolled environments to achieve natural HRI. In this prototype analysis, we observe two main challenges: Light control and dexterity of the robot.

The change in light intensity along the sessions presents a particular challenge for the vision system to detect the cards. We seek to avoid the modification of the cards to keep the interaction as natural as possible. However, the unreliable identification of the colours as the primary cue for the robot to identify the cards is the main limitation. A dynamic range of colours for card detection could mitigate this flaw in the prototype. Similarly, identifying the cards' images by border detection and segmentation could provide better results. The control of the light's intensity in the rooms could significantly impact the card detection accuracy but a significant cost(room adaptations and improvements). Better hardware could improve the prototype's performance; for instance, an industrial camera with high-quality light compensation could increase the card's detection performance.

Another technical limitation is the dexterity of the robot. The robot does not have the physical capability to manipulate the paper cards in the game. The lack of robot dexterity has a significant impact on how the game develops with not expert users as all the authors agree in the analysis of the prototype. However, further modifications in the game play and clear instructions could mitigate this main constrain. For instance, solid cards in wood or 3D printed can be gripped by the robot. The challenge relies on programming this particular robot to position the cards in a reliable way to be picked by the gripper.

The game engine does not present high tech challenges and can be escalated in the future. In other words, playing with more than six cards is possible. The main limitation is that the robot could be perceived as "too smart" by the users. With the correct vision system, the game engine can record all the cards in their respective positions with total accuracy. However, a high robot performance would impact the user who could feel she/he does not have chances to win against the robot.

4.2 Expert's Recommendations

From a technical perspective, the next steps involve developing a more accurate vision system. It is also required to extend the game engine's capability to play with ten pairs of cards. The multidisciplinary team suggests working in multiple aspects to design the robot behaviour and improve game playability to align the prototype with the LeBlanc's Taxonomy of Game Pleasures. We propose three critical factors in designing social robots for games: a) Robot Morphology, b) Robot expressiveness and c) Game design suitable to the robot capabilities.

Morphology of the robot: The team suggests the experimental validation of robots with different morphology. Usually, social robots are designed to look like humanoids; they present head, torso, and arms in diverse designs. However, it is not clear to what extent these humanoid embodiments add benefits to the user experience specifically in games. Comparative experiments are required

to define the importance of humanoid features in the robot's task compared with a robot arm.

Robot Expressiveness: Very often, the expressiveness of robots in experimental conditions is minimised to avoid possible undesirable effects in the studied variables. However, when humans play games, they tend to express themselves according to the stimuli [18, 19]. It is required to investigate social robots' expressiveness in users when they play board games with robots. Different grades of expressiveness should be evaluated in both robots to determinate the impact on the user's perception and performance.

Game design: Social Robots are different from other interactive technologies that require a screen as the main interface. Considering that our prototype does not look to use a screen, there is a possibility to redesign existing board games suitable to be played with robots and propose new games respectively [1, 6]. Other team members aim to maintain the gameplay as the original played by Humans and further push technology development in the prototype. Opposite to the mixed reality approach proposed by Robert et al. [20], this project aims to an HRI in the game that does not require extra infrastructure encouraging the interaction. Our philosophy is closer to the approach proposed in [17], making the robot interacting naturally in existing game setups.

5 CONCLUSIONS

After the prototyping of the robot games and observation of the video, the design team concluded that there would not be a significant user's engagement if robots are incapable of show dexterity and manipulate cards or tokens. We require to explore how to compensate for the lack of dexterity in this robot. In other words, the robot looks redundant if it is not capable of physically interacting with the game accessories and then with the user. Lack of physical interaction is usually not considered in HRI experiments because it aims to investigate other variables. As a design team we have two options for continuing the development of robot games: a) Enforce technology to have physical interaction with users; otherwise, the robot is irrelevant and not add value is present compared with video-games or classic board games. b) Change the game mechanics to make it suitable for the robots' current abilities. However, we consider that option b) has the main limitation of designing a game that will be out-of-date quickly depending on specific robot technology. In other words, a game designed under that approach could reduce its chance of being a relevant game in the future as chess or checkers.

We discussed that the robot becomes irrelevant in terms of expressiveness if it cannot take turns and flip the cards by themselves. Moreover, they create confusion even in the design team testing the game due to the memory game's paradigm taking turns. Synthetic voice, pointing gestures or laser pointers can not indicate how to alternate turns efficiently. Finally, we consider that results from this work can assist in the development of systems to enhance learning in various contexts in the future such as in education, cognitive rehabilitation.

6 FUTURE WORK

The low-fidelity prototypes using real robots offer significant advantages before to carry on full experiments using quantitative metrics.

A well-developed user-experience designed by a multidisciplinary team before exposition to non-expert users can significantly impact social robots out of the experimental rooms.

We consider that further explorations in robot morphology, expressiveness are worthy. Furthermore, if used in robot games, anthropomorphic interfaces should be tested as essential for interaction requiring further experimentation. Future comparative studies can be developed investigating robot's morphology (humanoid vs non-humanoid robots), expressiveness, and strategies to improve the robot's dexterity cost-effectively. NAO robot, Pepper Robot and other popular platforms could be suitable for this purpose. However, their popularity can affect the user's interaction in unexpected ways.

Similarly, the game play has the potential to be improved. Variations of the game can be efficiently designed modifying the cards and the game play. The set of cards used during the game play have a significant role in the playability of the game. Intricate designs which are not globally understood can distract over the goal of the game. However, they offer an opportunity to culture exposition and learning, but it is not well estimated if they significantly impact cognitive skills improvement or enhance user's engagement. The current icons present severe challenges in reading the cards by the artificial vision system. In this sense, there are several design choices available for designing suitable cards, or tokens with diverse aesthetic art on them; for instance, culturally appealing icons, minimalist icons, well-know images, and memes are suitable options. Besides, it is possible to add even more complexity to the game asking the users to learn and speak an unfamiliar language.

At the moment, available robots have not the physical capability to manipulate the cards in the game. As a design team, we keep an open discussion about the modification of the tokens used in the game to keep the interaction as natural as possible. It is imperative to make the robots take turns following the normal rules of the game.

Future considerations in the subtle interactive variables of the memory game should be explored and analysed. For instance, positive emotions and empathy for the robots, frivolous talk during the game, are not considered in this first development of the prototype. This approach could make the robot game increase its impact as a potential therapeutic tool for special populations such as Dementia Patients and Autistic children. This is possible considering the simplicity of the game. Increasing the number of cards also increases the complexity of the game. Hence, multiple special groups can take advantage of the proposed setup.

Finally, comparative (including human vs robot) quantitative empirical studies using established HRI-metrics (Godspeed, Big Five and others) with different user groups is highly recommended. For technical improvements using deep learning algorithms as Convolutional Neural Networks could significantly increase the vision system's performance in a future prototype.

ACKNOWLEDGMENTS

Thanks to Xiao Fang (Zoe) for her suggestions as a research assistant and the UNSW Creative Robotics Lab members for the feedback.

REFERENCES

- [1] Ernest Adams. 2010. *Fundamentals of Game Design* (2 edition ed.). New Riders.
- [2] Leigh Anderson. 2010. *Games Bible*. Workman Publishing, New York, NY.
- [3] Andrew G. Brooks, Jesse Gray, Guy Hoffman, Andrea Lockerd, Hans Lee, and Cynthia Breazeal. 2004. Robot's play: interactive games with sociable machines. *Computers in Entertainment (CIE)* 2, 3 (2004), 10–10.
- [4] Roger Caillois. 1958. *Les jeux et les hommes*. (1958).
- [5] Massimiliano L. Cappuccio, Eduardo B. Sandoval, Omar Mubin, Mohammad Obaid, and Mari Velonaki. 2020. Can Robots Make us Better Humans? *International Journal of Social Robotics* (Oct. 2020). <https://doi.org/10.1007/s12369-020-00700-6>
- [6] Konstantinos Chorianopoulos and Michail Giannakos. 2014. Design Principles for Serious Video Games in Mathematics Education: From Theory to Practice. *International Journal of Serious Games* 1, 3 (Sept. 2014). <https://doi.org/10.17083/ijsg.v1i3.12>
- [7] Filipa Correia, Patricia Alves-Oliveira, Nuno Maia, Tiago Ribeiro, Sofia Petisca, Francisco S. Melo, and Ana Paiva. 2016. Just follow the suit! trust in human-robot interactions during card game playing. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 507–512.
- [8] Filipa Correia, Patricia Alves-Oliveira, Tiago Ribeiro, Francisco S. Melo, and Ana Paiva. 2017. A social robot as a card game player. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.
- [9] Chris Crawford. 2003. *Chris Crawford on game design*. New Riders.
- [10] Maartje De Graaf, Somaya Ben Allouch, and Jan Van Dijk. 2017. Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 224–233.
- [11] Maartje MA de Graaf, Somaya Ben Allouch, and Jan AGM van Dijk. 2016. Long-term acceptance of social robots in domestic environments: insights from a user's perspective. In *2016 AAAI Spring Symposium Series*.
- [12] Maartje M. A. de Graaf, Somaya Ben Allouch, and Jan A. G. M. van Dijk. 2019. Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance. *Human-Computer Interaction* 34, 2 (March 2019), 115–173. <https://doi.org/10.1080/07370024.2017.1312406>
- [13] Geoffrey Engelstein. 2020. *GameTek*. HarperCollins, Sydney.
- [14] Martin Fischbach, Jean-Luc Lugin, Michael Brandt, Marc Erich Latoschik, Chris Zimmerer, and Birgit Lugin. 2018. Follow the White Robot - A Role-Playing Game with a Robot Game Master. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1812–1814.
- [15] M.A. Goodrich and D.R. Olsen. 2003. Seven principles of efficient human robot interaction. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, Vol. 4. 3942–3948 vol.4. <https://doi.org/10.1109/ICSMC.2003.1244504> ISSN: 1062-922X.
- [16] Lawrence T. Jones, Ashley Howden, Mark S. Knighton, Anson Sims, David L. Kittinger, and Richard E. Hollander. 1983. Robot computer chess game (US Patent 4,398,720).
- [17] C. Matuszek, B. Mayton, R. Aimi, M. P. Deisenroth, L. Bo, R. Chu, M. Kung, L. LeGrand, J. R. Smith, and D. Fox. 2011. Gambit: An autonomous chess-playing robotic system. In *2011 IEEE International Conference on Robotics and Automation*. 4291–4297. <https://doi.org/10.1109/ICRA.2011.5980528> ISSN: 1050-4729.
- [18] Tiago Ribeiro, Iolanda Leite, Jan Kędzierski, Adam Oleksy, and Ana Paiva. 2011. Expressing emotions on robotic companions with limited facial expression capabilities. In *International Workshop on Intelligent Virtual Agents*. Springer, 466–467.
- [19] Tiago Ribeiro and Ana Paiva. 2012. The illusion of robotic life: principles and practices of animation for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 383–390.
- [20] David Robert, Ryan Wistorrt, Jesse Gray, and Cynthia Breazeal. 2010. Exploring mixed reality robot gaming. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction (TEI '11)*. Association for Computing Machinery, New York, NY, USA, 125–128. <https://doi.org/10.1145/1935701.1935726>
- [21] Eduardo B Sandoval. 2016. *Reciprocity in Human Robot Interaction*. Ph.D. Dissertation. University of Canterbury.
- [22] Eduardo Benítez Sandoval, Jürgen Brandstatter, Utku Yalcin, and Christoph Bartneck. 2020. Robot Likeability and Reciprocity in Human Robot Interaction: Using Ultimatum Game to determinate Reciprocal Likeable Robot Strategies. *International Journal of Social Robotics* (June 2020). <https://doi.org/10.1007/s12369-020-00658-5>
- [23] Eduardo Benítez Sandoval, Jürgen Brandstatter, and Christoph Bartneck. 2016. Can a Robot Bribe a Human? The Measurement of the Negative Side of Reciprocity in Human Robot Interaction. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*.
- [24] Eduardo Benítez Sandoval, Jürgen Brandstatter, Mohammad Obaid, and Christoph Bartneck. 2015. Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner's Dilemma and the Ultimatum Game. *International Journal of Social Robotics* (2015), 1–15. <https://doi.org/10.1007/s12369-015-0323-x>
- [25] Jesse Schell. 2014. *The Art of Game Design: A Book of Lenses, Second Edition* (2 edition ed.). A K Peters/CRC Press, Boca Raton.
- [26] Katie Salen Tekinbaş and Eric Zimmerman. 2003. *Rules of Play: Game Design Fundamentals*. The MIT Press, Cambridge, Mass.
- [27] Martin Tomitsch, Cara Wrigley, Madeleine Borthwick, Naseem Ahmadpour, Jessica Frawley, A. Baki Kocaballi, Claudia Núñez-Pacheco, Karla Straker, and Lian Loke. 2018. *Design. Think. Make. Break. Repeat.: A Handbook of Methods*. BIS Publishers, Amsterdam.
- [28] Dvora Yanow and Peregrine Schwartz-Shea. 2015. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. Routledge. Google-Books-ID: w_TqBgAAQBAJ.

Support Publication #10 [SP10]

Cappuccio M.L., Shepherd, S. (2013),
“Pointing Hand: Joint Attention and Embodied Symbols”,
In Zdravko Radman *The Hand, an Organ of the Mind: What the Manual Tells the
Mental*, The MIT Press, Cambridge MA.

Reproduced with permission from the MIT Press. All rights reserved.

13 Pointing Hand: Joint Attention and Embodied Symbols

Massimiliano L. Cappuccio and Stephen V. Shepherd

One of the most characteristic and notable powers of the hand is its ability to indicate via pointing—a gesture produced to draw the attention of specific observers toward a distal target. When pointing successfully aligns signaler’s and recipients’ gazes and they become mutually aware of this alignment, one may say it manipulates “joint attention” (JA), the experience of openly sharing with others a common focus of interest (Seemann 2011). JA is crucial in human development; its robustness in nine- to twelve-month-old children predicts the acquisition of cognitive skills, including language (Mundy et al. 2007; Brooks and Meltzoff 2008); its absence is diagnostic of autism (American Psychiatric Association 1994). Though JA is pervasive among humans and crucial to social intelligence, scientists and philosophers have yet to agree on its precise features, mechanisms, and evolutionary and developmental origins. In this chapter, we argue that an understanding of pointing—a gesture that explicitly incarnates and manipulates JA—is key to tackling these questions. In particular, we will discuss how coattenders become aware that they are focusing on the same object and what abilities are required.

Philosophers of mind typically dichotomize representational and dispositional skills in social cognition (Seemann 2007), presenting us with a theory of mind (ToM) on the one hand, whether derived from folk psychology or mental simulation, versus embodied cognition on the other, comprising interactionist, enactivist, and/or narrative-practice accounts. Unsurprisingly, these views favor alternate approaches to JA: Either coattenders match their representations about their attentional focus by inferring shared beliefs or they coordinate their dispositions toward an attentional focus through overtly observable behaviors. Such alternate approaches have distinct cognitive prerequisites: If the representational model is correct, coattenders must be reciprocal mindreaders and infer one

another's (possibly propositional) mental contents in order to confirm parity. But if the dispositional approach is correct, coattenders may only engage in forms of reciprocal coordination mediated by embodied intentions (e.g., posture, orientation, motion, expression, and gaze). Both stories have some developmental and primatological support. Supporting the representational account, JA produced by fourteen-month-old infants reportedly involves appreciation of beliefs shared with the recipients (Moll et al. 2007); can convey abstract, absent contents (Liszkowski et al. 2007); and can aim to achieve altruistic rather than purely self-centered goals (Liszkowski et al. 2006). Supporting the dispositional account, infants younger than four years fluently produce and understand pointing despite poorly distinguishing between their own and others' beliefs, confounding the matching of mental representations (Gallagher and Hutto 2008); similarly, many nonhuman species coordinate group behavior toward a target without (it is assumed) possessing propositional attitudes and inferential capabilities (Knoblich and Sebanz 2008).

Individuals coordinate attention in many ways, but pointing is one of the most prototypical. Although the morphological details of this gesture can change across cultures and contexts, it appears to be universally used (Kita 2003) and possibly innate (as evident from blind-born infants). However, while the gesture of pointing is not conventional in the sense of signs, neither is it reducible to a mere epiphenomenon of attention (Kendon 2004), because unlike bodily postures and gaze, it has an explicitly communicative intention.

Whether JA is representational or dispositional depends primarily on the *minimal* cognitive prerequisites needed for nonhuman primates or human infants to engage in pointing. Because adult humans unequivocally have the ability to use JA symbolically (SJA), and because this ability arises (in evolution and development) from precursors that are capable only of basic JA (BJA), we are interested in observable behaviors that seem to scaffold the transition from BJA to SJA (cf. Shepherd and Cappuccio 2011). While some form of this distinction is broadly accepted, past debate has contrasted a declarative mode of pointing (assertive and informative), said to imply JA, with an imperative mode (requesting and ordering), said to not imply it. It remains contested whether nonhumans ever acquire declarative pointing: Some authors believe that even apes lack the representational abilities required (e.g., Povinelli et al. 2002; Tomasello et al. 2005), while others stress that apes at least sometimes point declaratively, at least when enculturated by humans (e.g., Leavens 2011; Leavens and Racine 2009).

We do not take a stand as to whether apes represent minds or engage in declarative pointing. Instead, we argue that declarative pointing does not *require* but, on the contrary, intrinsically *produces* a primitive form of representational intelligence. Any well-formed instance of pointing—that is, any act of pointing that effectively aligns recipients' gaze with the signaler's—incarnates the possibility of communal attention in physical form and is simultaneously recognized by all parties as explicitly produced to coordinate awareness. The hand can thus symbolically represent the coattenders' "jointness" in a minimal, prototypical, and embodied form, and does so through direct perception. Since pointing, so understood, *invites* rather than *requires* inference, and does so without specialized cognitive resources, we argue that although representation can clearly augment JA, it comes late to the party. Specifically, we propose that SJA arises in development and evolution through observable somatic states—most notably through declarative gestures such as pointing—long before the intervention of ToM.

In the following sections, we describe (1) basic and symbolic joint attention and (2) the phenomenology of declarative pointing, arguing (3) that pointing can arise from and mediate BJA and so evidences, to signaler and receiver alike, the relevance of coordinated behavior. Pointing thus (4) scaffolds the acquisition of SJA by physically cuing information not just in the world, but in the common ground of shared knowledge and intentions. Finally, we consider how such embodied manifestations of joint intention build a state of (5) "open knowledge."

1 Basic and Symbolic Joint Attention

Following or cuing another's gaze, even when the target is understood to be the subject of the other's attention, is insufficient evidence for BJA. BJA additionally requires that the coattenders are aware of mutually targeting the same object (Tomasello 1995), which then becomes available for overtly shared consideration. If BJA is initiated whenever a signaler (S) indicates to a recipient (R) her attention toward a certain targeted object (T), it is necessary that there exist:

- [I] *Shared attention* to T:
 $S \rightarrow T \ \& \ R \rightarrow T$
- [II] *Reciprocal attention* between participants:
 $S \rightarrow R \ \& \ R \rightarrow S$
- [III] Occurrent *iterative awareness* of the others' attention:
 $R \rightarrow S \rightarrow T, S \rightarrow R \rightarrow T, S \rightarrow R \rightarrow S \rightarrow T, \text{ etc.}$

The potential for iteration is crucial because it completes the collective attentional state Peacocke (2005) called “Mutual Open-Ended Availability” (MOEA): If attention is entirely symmetrical and overt, then as soon as one subject perceives a certain relation of iterative awareness (e.g., $R \rightarrow S \rightarrow R \rightarrow T$), the opportunity arises for the other to notice, producing a higher-order iteration ($S \rightarrow R \rightarrow S \rightarrow R \rightarrow T$). Were individuals’ perceptual and cognitive abilities infinite, the length of iteration could be arbitrary, but in practice only a small number of iterations is likely and indispensable for most social interactions among human adults (four, according to Sperber 2000 and Moore forthcoming). This constraint undermines not the open-endedness of JA, but merely the ability of each partner to analyze its details: By analogy, we can recognize the infinitely nested reflections in a Hall of Mirrors with neither need nor ability to count them.

MOEA is the essential feature of JA, and indeed most authors (e.g., Schiffer 1972; Sperber and Wilson 1986; Wilby 2010) account for the mutually, manifestly overt quality of JA experience through “mutual knowledge”—a reciprocal iterative understanding of the form “S knows that R knows that S knows . . . that X, and R knows that S knows that R knows . . . that X”—but whether this knowledge is dispositional (Campbell 2005; Gallagher and Hutto 2008) or representational (Tomasello 2008) is contested, perhaps signifying a distinction between mutual “know-how” and “know-that” (cf. Ryle [1945] 1971). In our view, while *dispositional* mutual knowledge is a necessary and sufficient condition for fluency in BJA, it is only necessary and *not sufficient* for fluency in SJA.

This is because the crucial issue at stake in JA is the presence of MOEA, irrespective of whether SJA is engaged. Consider that you may notice someone observing you admiring a motorcycle and see that he saw you spot him—and *you can do this purely perceptually*. You needn’t ponder his covert thoughts or intentions; you simply directly perceive a social context. In fact, you may automatically smile appreciatively at the motorcycle—or guiltily withdraw from same—based on the other’s directly perceived affect. It is surely true that these interactions are informed by a culturally and linguistically rich context—but dispositional MOEA neither precludes a role for these factors (because learned scripts and norms may shape our direct perception of a scene) nor requires explicit reflection on them (Gallagher and Hutto 2008). Thus, dispositional MOEA appears sufficient to account for BJA, and BJA to account for most instances of JA among animals and humans alike. Nonetheless, some instances of JA remain to be explained.

By introducing representations, SJA—but not BJA—can assign meaning to arbitrary locations, explicitly reference absent or abstract contents, engage informative contents and propositional attitudes, and interact with mentalistic attributions of belief. We argue that pointing plays a crucial role in making the evolutionary and developmental transition from BJA to SJA in that it physically embodies joint mental activity. To understand how this may occur we must consider the process of declaration in the context of the pointing gesture.

2 Dynamics of Declarative Pointing

The three most influential views on declaration conceptualize its preverbal function as to either (a) assert informative contents endowed with truth-values (cf. Leonard 1959), (b) direct the recipient's attention instrumentally to obtaining a certain reaction (pointing "proto-declaratively," as in Bates et al. 1975), or (c) share attention simply for its own intrinsic, contemplative, value (Tomasello 2008). All three have caveats when applied to pointing: (a) declarative gestures don't necessarily have a true/false content; (b) gestures can reference symbolic contents in a manner incompatible with purely dispositional or instrumental accounts (McNeill et al. 2008); and (c) infants show dissatisfaction if their "altruistic" pointing fails to evoke an affective response (Liszkowski et al. 2004). We therefore prefer a different definition: a deictic signal is declarative when the immediate interactional features of T (affordances recognizable in the occurrent perceptual context) are irrelevant to understand the gesture's meaning, except instrumentally to make the recipients aware of some background information related to shared attitudes toward T.

In order to unpack this idea, we shall partially follow Tomasello (2008), who builds on Grice's (1957) theory of "nonnatural meaning" to characterize declarative communication as representational. When pointing has an imperative valence its meaning is natural, because it is entirely and perceptually disclosed to R as a social affordance of T. For example, consider a toddler who, reaching toward a milk bottle far above her, looks toward her parent and starts to cry. BJA is sufficient to explain this sort of imperative interaction because S's referential intention is entirely evident in S's behavior. However, unlike the many animals that coordinate solely in species-typical contexts and ways, humans cooperate flexibly. A great deal of ambiguity can arise when signaling deictically with a multipurpose effector, such as a hand held in an indeterminate pose. Here, because the

desired goal is evident in neither the form of the gesture nor the perceptible environment, S's reference to T is a *blank*, which R must somehow fill. R must recognize S's intention, which is underspecified by the occurrently perceived environment but, because it is expressly communicative, is necessarily embedded in the common ground of shared conventions and beliefs.

Consider an example (Kita 2003; Tomasello 2008) in which the very same gesture, produced in the same scenario with the same T, can express entirely different meanings disambiguated only through the common ground. An orator speaking in a crowded conference hall can point to the empty chair in front of him to indicate that a free seat is available (gloss: "sit there"), that it is the only seat left in the hall (gloss: "there is a big audience"), that a certain person is sitting there (gloss: "there she is!"); or that an important guest is absent (gloss: "she's missing"). Declarative pointing can thus target an occurrent T (the empty chair) to refer to an incidental, abstract, or even absent feature X(T). Declarative pointing directs not only the eyes, but also the imagination. By designating and then referencing locations in an arbitrary space, declarative pointing comes to trade in symbols rather than affordances (Kendon 2004). This association is non-natural to the extent that X(T) can be specified only through the common ground, not from the details of the perceptual environment. Notice that, while common ground information about (X)T conveyed through communicative intention may vary, it must always at least reference itself as the source of the nonnatural meaning X(T). In other words, the beliefs to be shared by the coattenders in order to make sense of pointing must always include the belief that such beliefs exist. If (and only if) this happens, the gesture is implied to have nonnatural meaning and to be a kind of symbolic declaration—it can be interpreted only through reference not to T directly, but to the information associated with T.

We agree with Tomasello that nonnatural meaning is a necessary and sufficient condition for declarative communication, and transforming a perceptual T into a symbol of the *beliefs coattenders share about T* is what specifically distinguishes SJA from BJA. However, Tomasello additionally assumes that nonnatural meaning can be conveyed only when pointing has the *unique function* of sharing attention, hence excluding the possibility that declarative communication could also aim to initiate interaction. We disagree. While a contemplative, noninteractive stance toward T may reinforce declaration by deemphasizing direct affordances of T, it is not indispensable. A communicative act is declarative whenever it references contents from the common ground, regardless of its perlocutionary goal.

What, then, are the cognitive prerequisites of declarative pointing? Two opinions can be found in the literature. According to Tomasello, the common ground must be shared by S and R, and both must be aware that they share it. Drawing on data collected from infants and apes, Tomasello (2008, 189–190, 198) argued that declarative pointing requires coattenders to perform “recursive mindreading”; as implemented by Harris’s (1996) model, such metarepresentation explicitly entails that S and R simulate one another’s mental simulations (Tomasello 1999). When S points to the empty chair, S is inclined to think that R will correctly interpret S’s gesture as an intention to refer to the person who was supposed to sit there; in turn, when R sees S pointing to the chair, R is inclined to think that S knows that R knows that that person is supposed to sit there, and only thus can correctly interpret the symbolic meaning of the gesture. According to this model, JA requires mutual knowledge that is based on a mindreading process that is *internal* (it retrieves private, nonperceptual information via an offline reflective process) and *representational* (it operates on descriptions of one’s own mental contents as projected onto a coattender).

However, this is not the only account available. Primatologists remain divided as to whether nonhuman apes can point declaratively—despite general agreement that apes have limited talent for representation. Tomasello (2006) sees no compelling evidence for declarative pointing in apes, and he attributes the absence to the lack of uniquely human simulative devices. Leavens and Racine (2009) respond that declarative pointing has been observed repeatedly among captive apes (Savage-Rumbaugh et al. 1986), and at least once among wild apes (Veà and Sabater-Pi 1998). Even sophisticated pointing behaviors—for example, pointing to absent entities—might not be human specific. While some experiments provoked pointing toward absent entities by prelinguistic children but not by apes (Liszkowski et al. 2009), language-trained apes have reportedly referred to dead individuals (Fouts and Mills 1997), solicited refills of an empty glass (Hoyt 1941), and indicated hidden objects while specifying their type through a symbol language (Menzel 1999). If wild apes don’t often point declaratively, Leavens argues, this may be because of differences in developmental enculturation rather than pedigree—in other words, to training, not talent. According to this perspective, all forms of pointing are instrumental and intended to produce social interaction (Leavens 2004). While declarative pointing gestures continue to rely on a common ground, this common ground is not composed of salient facts inferred and projected via mental simulation, *pace* Tomasello, but rather,

in accordance with Hutto's (2011) hypothesis, reflects a dispositional knowledge of how to manipulate coattender's beliefs and desires through typically experienced narratives. For Leavens and colleagues, JA requires mutual knowledge derived from a behavior-reading process that is *external* (it directly perceives social affordances in sets of goal-directed intentions engaged by reciprocal attention) and *nonrepresentational* (it enacts appropriate social affordances by means of the R's disposition to respond to S's behavior).

The controversy regarding the cognitive precursors of declarative pointing can thus be characterized as one between representational and dispositional accounts of mutual knowledge. In Tomasello's camp, proponents of a simulationist ToM claim that others' goals are known through introspectively produced models (representational know-that) of inferred covert processes; in Leavens's camp, proponents of embodied cognition and narrative practice claim that others' intentional attitudes are grasped through directly perceived affordances and scriptlike heuristic templates (dispositional know-how). In our opinion, neither approach fully accounts for the properties of human JA, but with appropriate adjustment their complementarity can be demonstrated. The simulationist approach offers *reflectivity* but lacks *reciprocity*. On the one hand, each coattender has access to one's own knowledge state, as is necessary for inference, imagination, and the representation of abstract or absent contents; on the other hand, each coattender's understanding of the other must be achieved by "telementational" methods (Leavens 2011), that is, introspection and projection, because mere reciprocal attention to the outside world is insufficient for cases of nonnatural meaning. The solipsism implicit in telementation would undermine any openness and directness in JA, as Leavens correctly notes: S and R may both have the same representational mapping X(T), but because they each access it privately with no opportunity for overt comparison, they can't truly share it as a joint focus of attention. Conversely, the embodied or interactionist approach offers *reciprocity* but lacks *reflectivity*: MOEA interlinks the coattenders' disposition toward T, but no causal role is assigned to coattenders' explicit modeling of convergent and divergent views. Such an approach is tenable only when S's and R's goals match their established know-how, restricting any declared X(T) to a small library of stereotyped interactions. Tomasello (2008) is right to note that this precludes creative referencing of abstract or absent targets, limiting opportunities for SJA.

Let's return to our example of the lecture hall. Symbolically evoking an absent person by pointing to her empty chair presupposes a knowledge of the arbitrary association between chair and person that is at the same time

open (occurently reciprocal, per MOEA) and *nonnatural* (independent of the occurrent physical context or affordances of the chair but instead cuing associations in the common ground). These features are constitutive of declarative pointing as we've defined it, but in contrast to established approaches, they seem to necessitate *both* reciprocity and reflectivity. The current debate in primatology tries to define whether representation is required for declarative pointing, and both Tomasello and Leavens assume representation mediates JA only if each coattender accesses representational mutual knowledge through internal mental devices. We are inclined to reject the assumption, common to both paradigms, that social cognition is either both representational and internal, or not representational at all.

According to Tomasello, SJA is representational in character because S's communicative intention is neither overtly expressed by her pointing finger nor uniquely specified by her referential intention, and is thus made clear only through the context of the common ground. We agree that SJA is representational and that declarative pointing is sufficient to evoke it, but we remain unconvinced that declaration requires preexisting talent for representation. Instead, as we shall develop in section 3, we suspect declarative pointing *produces* representation. According to Leavens, declarative pointing is an elaboration of developmentally enculturated scripts for pointing, rather than a novel phenomenon with its own unique mental foundations. We agree that declarative communication may arise by invoking and abstracting previously experienced narratives related to pointing rather than arising by application of a preexisting internal representational mechanism. We differ from Leavens (2011) in that we believe this cultural–developmental experience implies a substantial transition in social cognition—one that achieves manipulation of sophisticated representations of beliefs through physically embodied media.

We conclude that SJA is neither purely dispositional in character nor predicated upon a preexisting system for mindreading. In the next two sections we will explore this hypothesis, first explaining how pointing is sufficient evidence for BJA, and then describing how declarative pointing is a sufficient scaffold for SJA.

3 Pointing Is Sufficient Evidence for BJA

Pointing is one of the purest means of manipulating JA. In any community that uses pointing, capacity for BJA is assured. Pointing necessarily evokes the core constitutive features of BJA and lacks extraneous features that might muddy its message. We identify three constitutive features of BJA that must be evoked by a coordinating signal. Specifically, a signal must

be: (1) *indicative* (it expresses S's intention to deictically reference T, thus sharing attention); (2) *salient* (it makes R aware of S and leaves S able to monitor R, creating reciprocal attention); and (3) *symmetric* (it requires that each coattender, including S, holds equal possibilities of attending to the others' attentional state, hence triggering iterated awareness). By combining these features, pointing coordinates a well-defined and mutually aware community.

It is rare that a signal *necessarily* displays all three of these features at the same time, but hand gestures do. A lieutenant can point out an enemy bivouac, commanding his troopers to silently attack. His order is effective, because he can be confident his troops (whom he can see) can also see where he points, and that he sees them, and that they see one another, coordinating the timing of their response—all while the enemies, sleeping unseen, are none the wiser. The lieutenant's signal achieves its goal because it necessarily coordinates a well-defined and mutually aware community. Other audiovisual signals lack at least one of BJA's three core features. A vocal alarm doesn't need, and is not meant, to be synchronously perceived with its T, so the focus of any shared attention would be either the signal itself (when T is not discernible) or T (when the signal has passed), but rarely both as a coherent intentional unit. Vocal signals are not indicative: They may have referential function but are nondirectional and so specify categories rather than instances (Cheney and Seyfarth 1990). Gaze and heading have indicative function (they refer deictically to some T), but can only effectively communicate S's attention if R's has already been secured through other means (e.g., eye contact, abruptly changed-and-held posture, or emphatic sounds). Worse, they do not easily facilitate symmetry: S cannot generally see her own eyes. More attention-getting bodily gestures (e.g., ground-slapping) are admittedly salient, but since they cannot deictically reference an arbitrary target, they result primarily in dyadic and asymmetric relations (i.e., they draw R's attention to S, not T). While these behaviors may help coordinate BJA, none is sufficient to instantiate it. Certainly, we grant that many animals coordinate BJA through species-typical signal exchange, as when one monkey threatens a foe, looks back toward an ally, and looks back toward the foe while bobbing its head ("appeal-aggression"; Waal 2003). But we argue that pointing gesture does all this and—importantly—achieves *only* this. Pointing is blankly indicative, equally useful in diverse contexts and for different purposes, and this gives it three emergent properties.

(1) *Ostension* (it conveys to R S's attention to T, producing shared attention). An indexical gesture *par excellence*, pointing holds flexible deictic

function able to precisely target any physical object regardless of direction or distance, provided that coattenders and T are all connected by sight-lines. Pointing is more flexible than vocal signals because T is determined by spatial location rather than categorical description; it is more purely indicative than gaze because it is physically removed from the affective context of S's facial expression. Moreover, pointing has an *exclusively* deictic function: whereas postural orienting is at least somewhat action specific and animal vocalizations are associated with stereotyped emotional responses to environmental cues, pointing affords no direct physical consequence and achieves nothing *in itself* but the orienting of R's attention. In this way, the purpose of pointing is entirely (and arbitrarily) ostensive, because it is free from the confound of implicit noncommunicative goals.

(2) *Self-ostension* (it draws attention to S while S attends to R, producing reciprocal attention). Pointing is effective in getting attention because, while the extension of the arm–hand–finger highlights a distal T, it simultaneously highlights the presence of S and emphasizes her communicative intention. Unlike eye contact, hand gestures can be easily displayed to multiple Rs, becoming the T of their shared attention; unlike vocal and olfactory signals, Rs must be immediately linked to S by line-of-sight, thus defining a selected community as the intended audience (e.g., directing allies but not rivals toward a strategic location). The function of pointing is not only ostensive but self-ostensive: S herself becomes *a part* of T, making salient not just the target, T, but the explicit relation, $S \rightarrow T$. This is decisive in ontogenesis (Reddy 2005) as the infant needs to become aware of the opportunities afforded by being a *target* of attention (not just an S, but also a T) in order to reciprocate caregiver gaze and, subsequently, to learn how to orient it.

(3) *Symmetry* (both S and Rs can equivalently observe $S \& R \rightarrow T$, offering both equal access to iterated awareness). Pointing is typically present concurrently in both S's and R's visual fields, unlike most other types of deictic signals. By externalizing the ostensive and deictic signal to a viewable part of one's own body, one's "targetness" becomes perceptually available to R—and also to S. Pointing becomes a concrete, directly perceivable, stand-alone indicator of triadic intent. Whereas mammalian eyes point to one thing at a time, and vocal signals fail to designate precise Ts, pointing can be used to indicate a specific T while simultaneously looking at a specific R (cf. Waal 2003), instantaneously involving R in a triadic relation with S and T. The fact that the R's dual focus of attention (S and T) is mirrored from S's side (R and T) creates the empirical conditions to establish and

iterate symmetric attentional relations. Moreover, because the pointing finger cues S as surely as it cues R, it becomes an externalized public signal, a concrete and observable indicator of the opportunity for collective attention, and more, of the abstract intention of collective action.

As long as S and R are symmetrically situated as two equal coattenders of the pointing gesture, their communication meets Peacocke's criteria of MOEA, because pointing facilitates iterated awareness through a single communicative act. This is a condition not easily achieved by other communicative signals, which require integration of multiple acts and modalities to achieve BJA's three core features. This doesn't mean that pointing is the only signal allowing iteration. It is interesting to note, however, that when (for example) haptic communication establishes JA with blind infants (Bigelow 2003), the hands remain the crucial effector. Here, the task becomes bimanual, for example, putting a toy in the infant's hand while touching his hand with the other hand. In this case, the iteration is enabled because a grasping hand can simultaneously contain and be contained by another hand so that use of two hands can establish a triadic relation.

Furthermore, as we have admitted, BJA can exist without pointing. Many species follow gaze (Shepherd 2010), and social animals coordinate actions to maintain group cohesion, defend against predators, and compete with rival social groups (Shepherd and Cappuccio 2011). However, because pointing accomplishes nothing in itself aside from the manipulation of attentional states, it must be motivated by social goals afforded through JA. Pointing is thus sufficient to establish *at least* BJA, and furthermore, it brings it about through a single enduring, physical, self-contained act. At the same time, pointing can be freely combined with other actions and with nonvisual modalities that emphasize its communicative intentions (e.g., a quizzical eyebrow, a guttural noise, or gaze conveying affective content toward R rather than T; cf. Csibra 2010 and Moore forthcoming). In our view, the most crucial aspect of declarative pointing is that it creates a physically embodied token of the JA of a well-defined community of actors, thus comprising an open and nonnatural representation of their collective state. But first we must digress to explain just what this public representation is and why it matters to SJA.

4 Declarative Pointing Is Sufficient to Scaffold SJA

We have demonstrated that interactions scaffolded by the pointing gesture are sufficient to bring about BJA, in particular by producing MOEA via (I)

shared attention, (II) reciprocal attention, and (III) iterated awareness. According to Peacocke's (2005) analysis, these features are sufficient to produce only an "incipient" form of JA, whereas the "full" form additionally requires what he calls "Reflective Social Consciousness": S and R *explicitly* jointly attend to T, becoming reflectively aware of the complex situation in which they are involved. In our terms, SJA doesn't only require the three features of BJA, but also:

(IV) *Reflective social consciousness* that participants share a state characterized by (I)–(IV).

The fact that (IV) refers to itself generates what Peacocke calls a "self-involving situation," that is, a situation *U* that is constituted by the fact that the agents are aware *both* that they jointly attend to T and that they are in the situation *U* (ibid., 319). Consider the tale of the Emperor's New Clothes. When the little boy mockingly points to the Emperor on the street, the crowd was already jointly attending to the Emperor's nudity and was mutually aware of one another's attention (as in BJA). What changed when the boy pointed? All parties involved suddenly become *reflectively* aware of their *collective* state of attention—not only that they were all looking at the Emperor, but that they were *all* finding the emperor's state *conspicuous*. Only then does shared embarrassment escalate into public scandal.

This meta-awareness deploys both reciprocity (each coattender's state of awareness depends on and refers to the others') and reflectivity (it implies awareness of one's own state of awareness). It is key that reflective social consciousness is not merely dispositional, as it doesn't only implement an interactional know-how, but also implies at least minimal knowledge that this interactional know-how *exists* and *is mediated by* the representation of S's and R's self-involving state of awareness. Peacocke argues that the reflectivity of SJA "is possible only for beings who have some way of representing attention, mental states, and employing some form of indexical reference to mental states" (ibid., 308). This form of representing attention is minimal, and merely indexical in character; as such, it is different from the representational mutual knowledge as endorsed by simulationists and rejected by the enactivists/interactionists.

In fact, whereas mutual knowledge is representational in that the coattenders' attentional states reciprocally portray one another (R's description of T contains S's description of T *and* S's description of T contains R's description of T), reflective social consciousness is representational in that the coattenders' collective state of attention makes them recognize that

they share that very state (U contains a jointly attended target, T , and indexically refers to itself, the collective state U). In representational mutual knowledge, each attender's representational knowledge about T must fully represent the other's, and so runs into infinite regress; by contrast, in reflective social consciousness, the communal state of attention reveals some information concerning its own structure (ibid., 306–307). Hence, whereas mutual knowledge can be compared to two encyclopedia entries each fully recapitulating the other's recapitulation of their own, *ad infinitum*, reflective social consciousness is like an entry that refers to the encyclopedia that contains it.

So formulated, the fact that SJA is mediated by reflectively produced representation does not lead to a paradox—but it does shift the burden, so that some empirical conditions must simultaneously grant this representation *both* reflectivity and reciprocity. At heart, this problem relates to the foundations of symbolism in JA. Tomasello and Leavens debate whether JA can be mediated by representational processes (simulating, thinking, imagining) or not, but they implicitly assume that representations are *internal* (private, covert, nonperceptual) in character. This certainly *can* be true—but it needn't be. We see no inconsistency in modeling JA as both externally accessible to multiple individuals through occurrent perception (it relies on public, perceptual cues) and representational (the cues in question serve to reflectively access and manipulate private, offline information). We access representations of just this sort whenever we think of immaterial, absent, or abstract contents because of publically available cues. As humans who grow up in a material culture, we routinely traffic in covert ideas and goals crystallized into overt sounds and images.

What we propose is that declarative pointing is the prototypic “external representation” (Kirsh 2010)—a particular instance of extended cognitive processing that doesn't only use the body as an extracranial vehicle of thought (see Clark, this vol.; Wheeler, this vol.), but also makes this thought perceptually available for public consideration, enabling the coattenders to “think about [their own joint] thinking” (Malafouris 2007). Declarative pointing thus represents BJA by indexically referring to it, externalizing it as a communal mental state. This first, minimal, symbolic function of pointing is to openly declare the situation of JA to the participants, bringing about Peacocke's reflective social consciousness, just as the crowd's public awareness is suddenly “awakened” by the little boy's pointing toward the naked emperor. This function makes openly perceptible not only social affordances, but also the explicitly collaborative perspective, and does so independently of context-specific or stereotyped species-

typical goals (such as “mob the snake”). The pointing gesture thus makes collaborative intention an object of public knowledge and so renders it available to planning and learning mechanisms. Through the pointing gesture, a prelinguistic and prementalistic human infant can indicate to a novel adult a novel object—or have a novel object indicated to her—and *legitimately expect* they will together do something rewarding with that object, learning to cultivate JA as a useful state. This experience lays a foundation for the second symbolic function of the pointing gesture, which consists in externalizing to an actual perceptual target T the imaginative content X(T) associated with it, making X(T) something that individuals can perceptually attend toward *together*.

The role played by pointing in revealing communicative intent is preeminent for two reasons. First, unlike other indicative gestures such as posture, gaze, and olfactory marking, it is necessarily self-indexical: while pointing’s ostensive function expresses S’s referential intention, its self-ostensive function can reference its own referential intention $S \rightarrow T$ and hence solicits the attribution of nonnatural meaning to T. Second, unlike occasionally declarative, and potentially prototypical, gestures such as offering/showing an object, used both by infants (Bates et al. 1975) and apes (Plooij 1978; Russell et al. 1997), pointing operates at distance from T, deemphasizes direct manipulation, and so underlines the noninteractional motivation of the gesture. Various other signals can produce BJA, but their lack of reciprocity, arbitrariness, and explicitly communicative intent make them insufficient to produce explicit awareness of the JA state as such, hindering SJA. Pointing necessarily invites this awareness because it serves only to symmetrically indicate JA.

Because declarative pointing is intrinsically (in Peacocke’s terms) “self-involving,” reflectivity and reciprocity are both granted: reflectivity, because the ostensive and self-ostensive function of pointing offers to participants an open representation of their intentional JA; reciprocity, because its symmetrical visibility can make both S and R mutually aware that they are representing their direction of attention through the same physical medium. As epitomized by the tale of the naked emperor, declarative pointing necessarily brings about public awareness in which the attended scene indexically refers to itself, incarnating reflective social consciousness in physical form and concretely instantiating SJA. Just as SJA always includes an indication of itself, declarative pointing highlights not merely T but also its own symbolic role in cuing T. Declarative pointing’s minimal function includes symbolizing its own capacity to symbolize.

Both apes and humans ritualize gestures to initiate social interactions (“ontogenetic ritualization”; see Tomasello 1999), and, since pointing initiates JA, it is reasonable to assume that under appropriate circumstances, they can ritualize pointing as a symbol of JA. If declaration is implicit to pointing, then why is it rare or absent in wild apes? One reason may be that they don’t have the imaginative and perspective-taking skills that allow humans to represent their own bodies as objects of attention, rather than effectors of interaction. A condition for ritualization of pointing in development is that infants expect caregivers’ attention (Franco and Butterworth 1996), that is, to share attention to the fact that attention has to be shared. Apes, by contrast, may lack an innate or learned disposition to share imagery and narratives, which may limit opportunities to *notice* self-indexicality and communicative expectations. Another important fact is that most apes lack the intensely (and *arbitrarily*) collaborative child-rearing experience typical of humans, in which parents actively help infants achieve goals that *the infants are not yet aware of*. For example, human infants learn very early to gesticulate to attract and direct the caregiver’s attention, to recognize parts of their body as objectual targets of the caregiver’s attention and hence to acquire expectations about that attention (Reddy 2005), about its benefits, and about strategies for its manipulation. This self-derived template may then be extended to triadic contexts involving external targets. As we have observed, because human caregivers pervasively attend to their charge’s needs, infants learn that manipulation of attention predicts reward across diverse known *and novel* circumstances.

5 Open Knowledge

Our analysis suggests that pointing can arise in communities that possess only BJA abilities, but that once it arises, it instantiates an externalized representation of JA and so scaffolds the acquisition of SJA, which then fully develops only in tandem with other psychological skills (motivational, dispositional, and particularly representational) that promote social cooperation and coordination.

If this hypothesis is correct, declarative pointing is a sufficient condition to bring about SJA, as it both produces and symbolizes JA. But moreover, declarative pointing instantiates the capacity for S and R to share T as a symbol of X(T). Declarative pointing can evoke absent, private contents through publicly and perceptually available targets, hence extending the openness of JA to abstract communications. Let us assume that X is a

minimal set of information that allows S and R to map $X(T)$ from T, for example, associating a person of shared interest with the chair in which he or she typically sits. S, who knows X, thinks of the absent person as soon as she sees the empty chair, and wants to share with R this thought. If R knows X and sees S pointing to the chair, R too will think of the absent person. Do S and R both need to know that they both know X? Yes, but only in a minimal sense that does not necessarily involve representational mutual knowledge. A more primitive experience suffices: overtly and jointly representing T (via the pointing gesture) *as a vehicle of occurrent knowledge and common interest*. This “open knowledge” (Peacocke 2005), in our hypothesis, is achieved when pointing is employed as a symbolic medium that serves to make X perceptually salient.

In fact, even if S and R independently imagine, rather than see, the person supposed to sit on the chair, both openly enjoy a public T-centered experience of the person *as if* she were actually in front of their eyes. The symbolic content evoked by pointing has the flavor of direct observation in that it enjoys various features of occurrent perception, such as *passivity*, *spatial localization*, and virtual *autonomy* from the coattenders’ beliefs and/or judgments. R’s experience of $X(T)$ would remain anchored to the perceived T even if either S or R believed that X is false, or R judged wrongly the other’s particular beliefs about X. Note the difference: Covert contents appreciated through representational mutual knowledge depend entirely on *how* they have been actively inferred and subjectively attributed to the other’s mind; on the contrary, those intuitively accessed as open knowledge are passively recognized as objective attributes of T, whether or not they involve inferences about the other’s knowledge state. Thus, whereas mutual knowledge refers to a bustle of beliefs and judgments covertly produced by each participant about the others, open knowledge allows participating in a “mutual world,” freely accessible by any participant, but virtually independent from them.

Open knowledge is immediate, without nested inferences, because reflective social consciousness grounds in perception the representation of a mutual world: $X(T)$ can be experienced by S and R as open to both if, and only if, their situation of JA to T makes them aware of this very situation. Whereas in BJA only the social affordances related to T are experienced by S and R as openly available to them (knowledge of *how* we can sit on that chair), in SJA the representation of these very social affordances is experienced as equally open and publicly accessible (knowledge *that* somebody usually sits on that chair), provided that S and R jointly attend to T, but regardless of their actual access to each other’s minds. This is why

every coattender tends to experience $(X)T$ *as if* the others actually know X and know that X is also known by any other coattender.

Open knowledge is more flexible than dispositional mutual knowledge and less demanding than representational mutual knowledge, and occupies an intermediate state between them. Dispositional mutual knowledge requires a repertoire of stereotyped interactive behaviors shared by coattenders: This hinders the declarative character of the communication, because the more communication depends on the immediate interactional context, the more the generality of $X(T)$ is constrained, preventing the necessary reflectivity of SJA. Conversely, representational mutual knowledge requires recursive mindreading between the coattenders, mediated by telementation, which hinders the openness of JA: The more we need to reflect on the other's covert intention, the more we lose the immediacy of symbolic communication and the perceptual-like character of $X(T)$.

Open knowledge is not only phenomenologically distinct from mutual knowledge, it is also less computationally demanding and more robust in creating expectations about its content, as irrelevant recursive mindreading and behavioral affordances can be bypassed or surrogated to perceptual representation of SJA, provided that both S and R have some at least minimal acquaintance with X and free perceptual access to T . Open knowledge helps explain why infants point "declaratively" even before they gain the ability to adjust their signals to R 's attentional state (Franco and Butterworth 1996): For human infant signalers, even the most vague sensitivity to others' attention (which is seemingly innate and shared by all tetrapods; cf. Sowards and Sowards 2002) is sufficient to learn that capturing and manipulating attention is rewarding, because among humans rewards are consistently anchored to associations of the T they have designated (independent of any metarepresentation by S of these associations in R 's mind).

6 Conclusion

SJA encompasses three triadic relations that are brought about, directly and simultaneously, by any well-formed act of declarative pointing, that is, any act of pointing that references at the same time the target, T , the self-involving situation, U , and the underspecified yet communally relevant referent, $X(T)$.

(1) Perceptual relation (BJA): S and R jointly attend to T , as they directly perceive T and the other's attention toward T ;

- (2) Self-indexical relation (reflective social consciousness): S and R jointly attend to S's pointing finger and *each of them sees it as a symbol of their U because each of them is aware that they are both aware of attending to it*;
- (3) Imaginative relation (open knowledge): Both S and R attend to X(T), and *recognize some X(T) as available for joint contemplation (know-that) as a potential action target or modifier of potential action plans*.

The crucial transition from (1) to (3) is scaffolded by (2): Perceiving the pointing finger as an embodied representation of JA externalizes S's agency from herself and invites the coattenders' identification with a *triadic virtual agent* whose intentions are informed by, but partially dissociated from, coattenders' purely private dispositions. By embodying *U*, pointing comprises a symbolic template for further traffic in externalized representations, and just so, during human development babies learn verbal labels for pointed-out objects. Peacocke (2005) observes that by means of SJA, open knowledge is a precursor, rather than a by-product, of the sophisticated mental devices that usually produce representational mutual knowledge; at the same time, open knowledge is a product, rather than a precondition, of the adaptive interactive skills that constitute dispositional mutual knowledge. Open knowledge provides a representational framework that is more fundamental than representational mutual knowledge but significantly more flexible than dispositional mutual knowledge, in that it manipulates representational contents as well as affordances. In the same way, SJA can arise (in development and evolution) from BJA, for it builds on an open perceptual experience that is immediately and overtly available, but it is not *reducible* to BJA, for it builds on a nonnatural representation of hidden, private knowledge.

If the representational capacity required by reflective social consciousness is reducible neither to specialized adaptive dispositions nor to some preexisting ToM for inference of others' knowledge states, then where does it come from? Hutto's (2008) mimetic ability hypothesis posits the emergence of ToM through symbolic perception, with symbolic perception arising through JA rather than the other way around. In our terms: By symbolically incarnating a communal virtual agent that is only partially dissociated from one's own subjective experience, pointing may provide each coattender with an initial framework for analyzing the covert intentions and beliefs of others. As for the origins of symbolic cognition, Hutto refers to the evolution of a panoply of embodied imaginative-imitative skills: motor memory and reenactments of motor actions, fine-grained imitation of complex motor chains, offline rehearsal of tool-making, perspective shifting, pretended play, and protological and amodal thinking

in concrete problem solving. In this account, folk-psychological competence arises not from inferential skills, but from humanity's marked behavioral flexibility, including coordinated behaviors, complex social interactions, and collaborative tool-construction or usage. In other words, the complex imaginative skills necessary to develop hand actions with abstracted or latent goals, such as the construction of tools, are the same skills necessary for gestures to represent abstracted or latent content, as in iconic gesture, ritual dance, or pantomime abilities, which in turn provide embodied support for simple prelinguistic proto-narratives.

An important indicator of how complex behaviors are constructed comes from learning how they are implemented, across species and developmental stages, by actual brains. Mastery of declarative communication surely depends on complex social behaviors related to sharing and modeling experience and on motivation to engage in cooperative or altruistic activities (Tomasello 2009). Such behaviors are unlikely to be independent of the progressive development of motor abstraction (Gallese 2009), metaphorical categorization of social affordances and causal relations (Lakoff and Gallese 2005), or goal emulation and imitation proper (Rizzolatti et al. 2002b). At a more fundamental level, such abilities may depend on the fine intertwinement of observed and enacted brain states including motor intention (Rizzolatti et al. 2002a) and sensory attention (Shepherd et al. 2009), as leveraged by goal-directed motor imagery and remembered and imagined perspective shifts (Currie and Ravenscroft 1997). It will be interesting to see whether and how these proposed neural processes contribute to JA. Our prediction is that hand-centered skillful dispositions, more than specialized inferential mechanisms, play a pivotal role in human social coordination.

References

- American Psychiatric Association. 1994. *DSM-IV: Autistic Disorder*. Washington, DC: American Psychiatric Association.
- Bates, E., L. Camaioni, and V. Volterra. 1975. Performatives prior to speech. *Merrill-Palmer Quarterly* 21:205–226.
- Bigelow, A. E. 2003. The development of joint attention in blind infants. *Development and Psychopathology* 15 (2):259–275.
- Brooks, R., and A. N. Meltzoff. 2008. Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language* 35 (1):207–220.

- Campbell, J. 2005. Joint attention and common knowledge. In *Joint Attention: Communication and Other Minds*, 287–297, ed. N. Eilan, C. M. Hoerl, T. McCormack, and J. Roessler. Oxford: Oxford University Press.
- Cheney, D. L., and R. M. Seyfarth. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.
- Csibra, G. 2010. Recognizing communicative intentions in infancy. *Mind & Language* 25 (2):141–168.
- Currie, G., and I. Ravenscroft. 1997. Mental simulation and motor imagery. *Philosophy of Science* 64 (1):161–180.
- Franco, F., and G. Butterworth. 1996. Pointing and social awareness: Declaring and requesting in the second year. *Child Language* 23:307–336.
- Fouts, R., and Mills, S. T. 1997. *Next of Kin*. New York: William Morrow.
- Gallese, V. 2009. Motor abstraction: A neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Research* 73 (4): 486–498.
- Gallagher, S., and D. Hutto. 2008. Understanding others through primary interaction and narrative practice. In *The Shared Mind: Perspectives on Intersubjectivity*, 17–38, ed. J. Zlatev, T. P. Racine, C. Sinha, and E. Itkonen. Amsterdam: John Benjamins.
- Grice, H. P. 1957. Meaning. *Philosophical Review* 66:377–388.
- Harris, P. 1996. Desires, beliefs, and language. In *Theories of Theories of Mind*, 200–222, ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.
- Hoyt, M. 1941. *Toto and I*. New York: Lippincott.
- Hutto, D. 2008. *Folk Psychological Narrative: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.
- Hutto, D. 2011. Elementary mind minding, enactivist-style. In *Joint Attention: New Developments in Philosophy, Psychology, and Neuroscience*, 307–341, ed. A. Seemann. Cambridge, MA: MIT Press.
- Kendon, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge, MA: Cambridge University Press.
- Kirsh, D. 2010. Thinking with external representations. *AI & Society* 25:441–454.
- Kita, S., ed. 2003. *Pointing: Where Language, Culture, and Cognition Meet*. Mahwah, NJ: Erlbaum.
- Knoblich, G., and N. Sebanz. 2008. Evolving intentions for social interaction: From entrainment to joint action. *Philosophical Transactions of the Royal Society of London* 363 (1499):2021–2031.

Lakoff, G., and V. Gallese. 2005. The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology* 22:455–479.

Leavens, D. A. 2004. Manual deixis in apes and humans. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems* 5:387–408.

Leavens, D. A. 2011. Joint attention: Twelve myths. In *Joint Attention: New Developments in Philosophy, Psychology, and Neuroscience*, ed. A. Seemann. Cambridge, MA: MIT Press.

Leavens, D. A., and T. P. Racine. 2009. Joint attention in apes and humans: Are humans unique? *Journal of Consciousness Studies* 16 (6–8):240–267.

Leonard, H. S. 1959. Interrogatives, imperatives, truth, falsity, and lies. *Philosophy of Science* 26 (3):172–186.

Liszkowski, U., M. Carpenter, A. Henning, T. Striano, and M. Tomasello. 2004. Twelve-month-olds point to share attention and interest. *Developmental Science* 7:297–307.

Liszkowski, U., M. Carpenter, T. Striano, and M. Tomasello. 2006. 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development* 7:173–187.

Liszkowski, U., M. Carpenter, and M. Tomasello. 2007. Pointing out new news, old news, and absent referents at 12 months. *Developmental Science* 10:F1–F7.

Liszkowski, U., M. Schaefer, M. Carpenter, and M. Tomasello. 2009. Prelinguistic infants, but not chimpanzees, communicate about absent entities. *Psychological Science* 20:654–660.

Malafouris, L. 2007. Before and beyond representation: Towards an enactive conception of the palaeolithic image. In *Image and Imagination: A Global History of Figurative Representation*, 289–302, ed. C. Renfrew and I. Morley. Cambridge: The McDonald Institute for Archaeological Research.

McNeill, D., S. Duncan, J. Cole, S. Gallagher, and S. Bertenthal. 2008. Either or both: Growth points from the very beginning. *Interactional Studies* 9 (1):117–132.

Menzel, C. R. 1999. Unprompted recall and reporting of hidden objects by a chimpanzee (*Pan troglodytes*) after extended delays. *Journal of Comparative Psychology* 113:426–434.

Moll, H., M. Carpenter, and M. Tomasello. 2007. Fourteen-month-olds know what others experience only in joint engagement. *Developmental Science* 10 (6): 826–835.

Moore, R. Forthcoming. Cognising communicative intent.

- Mundy, P., J. Block, C. Delgado, Y. Pomares, A. V. van Hecke, and M. V. Parlade. 2007. Individual differences and the development of joint attention in infancy. *Child Development* 78 (3):938–954.
- Peacocke, C. 2005. Joint attention: Its nature, reflexivity, and relation to common knowledge. In *Joint Attention: Communication and Other Minds*, 298–324, ed. N. Eilan, C. M. Hoerl, T. McCormack, and J. Roessler. Oxford: Oxford University Press.
- Plooij, F. X. 1978. Some basic traits of language in wild chimpanzees? In *Action, Gesture, and Symbol*, 111–131, ed. A. Lock. London: Academic Press.
- Povinelli, D. J., J. M. Bering, and S. Giambrone. 2002. Chimpanzees' "pointing": Another error of the argument by analogy? In *Pointing. Where Language, Culture, and Cognition Meet*, 35–68, ed. S. Kita. Mahwah, NJ: Erlbaum.
- Reddy, V. 2005. Before the "third element": Understanding attention to self. In *Joint Attention: Communication and Other Minds*, 85–109, ed. N. Eilan, C. M. Hoerl, T. McCormack, and J. Roessler. Oxford: Oxford University Press.
- Rizzolatti, G., L. Craighero, and L. Fadiga. 2002a. The mirror neuron system in humans. In *Mirror Neurons and the Evolution of Brain and Language*, ed. M. Stamenov and V. Gallese. Amsterdam: John Benjamins.
- Rizzolatti, G., L. Fadiga, L. Fogassi, and V. Gallese. 2002b. From mirror neurons to imitation: Facts and speculations. In *The Imitative Mind: Development, Evolution, and Brain Bases*, ed. A. Meltzoff and W. Prinz. Cambridge, MA: Cambridge University Press.
- Russell, C. L., K. A. Bard, and L. B. Adamson. 1997. Social referencing by young chimpanzees. *Journal of Comparative Psychology* 111:185–193.
- Ryle, G. [1945] 1971. Knowing how and knowing that. In *Gilbert Ryle: Collected Papers*. vol. 2., 212–225. New York: Barnes and Noble.
- Savage-Rumbaugh, E. S., K. MacDonald, R. A. Sevcik, W. D. Hopkins, and E. Rubert. 1986. Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*). *Journal of Experimental Psychology: General* 115:211–235.
- Schiffer, S. 1972. *Meaning*. Oxford: Clarendon Press.
- Seemann, A. 2007. Joint attention, collective knowledge, and the "we" perspective. *Social Epistemology* 21 (3):217–230.
- Seemann, A., ed. 2011. *Joint Attention: New Developments in Philosophy, Psychology, and Neuroscience*. Cambridge, MA: MIT Press.
- Sewards, T. V., and M. A. Sewards. 2002. Innate visual object recognition in vertebrates: Some proposed pathways and mechanisms. *Comparative Biochemistry and Physiology, Part A: Molecular & Integrative Physiology* 132 (4):861–891.

- Shepherd, S. V., J. T. Klein, R. O. Deaner, and M. L. Platt. 2009. Mirroring of attention by neurons in macaque parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 106 (23):9489–9494.
- Stephen, S. V. 2010. Following gaze: Gaze-following behavior as a window into social cognition. *Frontiers in Integrative Neuroscience* 4:5.
- Shepherd, S. V., and M. Cappuccio. 2011. Sociality, attention, and the mind's eyes. In *Joint Attention: New Developments*, ed. A. Seemann. Cambridge, MA: MIT Press.
- Sperber, D. 2000. Metarepresentations in an evolutionary perspective. In *Metarepresentations: A Multidisciplinary Perspective*, 117–137, ed. D. Sperber. Oxford: Oxford University Press.
- Sperber, D., and D. Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. 1995. Joint attention as social cognition. In *Joint Attention: Its Origins and Role in Development*, 103–120, ed. C. Moore and P. Dunham. Hillsdale, NJ: Erlbaum.
- Tomasello, M. 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. 2006. Why don't apes point? In *Roots of Human Sociality: Culture, Cognition and Interaction*, 506–524, ed. N. J. Enfield and S. C. Levinson. Oxford: Berg.
- Tomasello, M. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press.
- Tomasello, M. 2009. *Why We Cooperate*. Cambridge, MA: MIT Press.
- Tomasello, M., M. Carpenter, J. Call, M. T. Behne, and H. Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28:675–735.
- Veà, J. J., and J. Sabater-Pi. 1998. Spontaneous pointing behaviour in the wild pygmy chimpanzee (*Pan paniscus*). *Folia Primatologica* 69:289–290.
- Waal, F. B. de. 2003. Darwin's legacy and the study of primate visual communication. *Annals of the New York Academy of Sciences* 1000 (1):7–31. doi: 10.1196/annals.1280.003.
- Wilby, M. 2010. The simplicity of mutual knowledge. *Philosophical Explorations* 13 (2):83–100.

Support Publication #11 [SP11]

Shepherd V.S., Cappuccio M.L. (2012),
“Sociality, Attention, and the Mind’s Eye”
in Axel Seemann (ed.), *Joint attention. Perspectives and
Developments*, MIT Press, Cambridge MA, January 2012, pp. 205-242.

Reproduced with permission from the MIT Press. All rights reserved.

9 Sociality, Attention, and the Mind's Eyes

Stephen Shepherd and Massimiliano Cappuccio

They sit together, one with his arm slung over the other's neck. They cluck over an infant—a boy—and they raise and inspect him, chattering over his sex. As they stare intently, one adult kneads the other's shoulder. Except for one detail, we could be forgiven for assuming the two *jointly attended* their newly arrived companion. However, this detail is key: the individuals in question are not humans, but Barbary macaques.

Many nonhuman animals seem sensitive to the attentional states of others, but joint attention is thought to be uniquely human, intimately tied to our ability to represent the minds of others. While the precise definition of joint attention continues to be debated, it involves (at minimum) attention to another individual and to the subject of his or her attention. In this chapter, we will explore the behavioral and neural evidence regarding animals' understanding of others' attention, in particular that implied by gaze following. We will conclude by describing a crucial aspect of human sociality—the use of pointing gestures to manipulate states of joint attention—and discuss what this gesture implies about differences between human and nonhuman social cognition. We provisionally distinguish four levels of socially cued attention:

1. *Gaze following, or the tendency to look where we see someone else looking* For ourselves—or, indeed, any animal—to know the subject of another's attention, we must evaluate how he or she directs his or her sense organs. In the visual domain, for example, this involves perceiving his or her gaze, translating between his or her reference frame and our own by replicating or simulating the other's viewpoint, and extending our attention to include the other's putative visual focus. The ability to follow attention is prerequisite to the ability to jointly attend, or moreover, to infer another's intentions and goals from his or her bodily behavior. Nevertheless, this level of socially cued attention is merely a reorienting of attention in the direction attended

by another—and can therefore occur with neither awareness nor intervening conceptual understanding of the other's mental or intentional state.

2. *Seeing gaze as intention* At the next level of sophistication, an individual A observing an individual B focusing its attention on C recognizes B and C to be linked. This linkage may be achieved through simultaneous, split attention toward both B and C or through some direct perception of an intentional relationship $B \rightarrow C$. No intrinsic content need be ascribed to this linkage, which is not representational in character. Nevertheless, linking the individual with that individual's attention focus makes possible the ascription of enactive intentions such as *seeing*, *wanting*, and *responding*. An individual attending B *looking toward* C might become aware from this that B *sees* C, that his or her expressed affective state is *about* C, and that these things predict certain types of *interaction* with C.

3. *Nonrepresentational joint attention: perception of collective attentional states* The perception that another's attention refers to the same focus as our own affords certain uniquely social behaviors. Without another individual's attending to the same object as ourselves, it is impossible to compete or cooperate. Perhaps more compellingly, the perception that another individual attends to the same focus as ourselves makes possible social affordances such as *outnumbering our foe*. Many animals work in groups to defend against and drive off threats—predators, rival social groups—and individuals may regulate their participation in these behaviors through nonrepresentational recognition of collective action (see also Gallagher, this volume, and Hutto, this volume). However, instances of joint attention recognized in this way would be constrained by each animal's repertoire of species-typical social behaviors.

4. *"Full-blooded" joint attention: participation in a shared mental state* In conventional usage, joint attention requires mutuality. This is more than mere *cooperation* or *competition*, for example: I can compete with an individual who is unaware of me, as long as I'm aware our actions conflict. Moreover, I can cooperate, taking that individual's side in a conflict while he or she is unaware that I'm behind him or her, threatening his or her foe. Humans, however, can go farther, treating joint attention as a state in which each party participates in the representation of some "common ground" (Tomasello, 1995, 1999, 2008, pp. 94–95). Humans learn to actively manipulate these mutual representational states, shaping their content through the use of declarative pointing gestures.

It is broadly agreed (but not universally agreed; see Leavens, this volume) that only humans are capable of full-blooded joint attention, but what of

these proposed subordinate components? If animals cannot fully jointly attend, can they recognize communality, mutuality, or similar social states? What do other animals understand of the dyadic relations between animals and their distal behavioral goals? Which species follow their companion's gaze, and is their gaze following behavior reflexive and unconscious or flexible and goal directed? In the following discussion, we will first survey animal gaze following behaviors, then discuss possible neural correlates of these behaviors, and finally return to the question of how human joint attention—and in particular our manipulation of joint attention through gesture—compares to those forms of socially cued attention which are widespread among nonhuman species.

Following Gaze

To respond to gaze, it must first be perceived—but at least theoretically, this perception need not be precise, conscious, or intentional. Using a simple computerized cuing task, Friesen and Kingstone (1998) demonstrated that eye gaze cues influenced subjects' attention within a tenth of a second; moreover, attention was allocated in the direction of gaze despite the fact that the gaze cues had no predictive value and were thus irrelevant. Langton and Bruce (1999) found similar results using observed head orientations, and Driver and colleagues (1999) strengthened the case for automaticity, reporting subjects follow gaze even when explicitly informed that cues are *counterpredictive* of target location (see figure 9.1). These

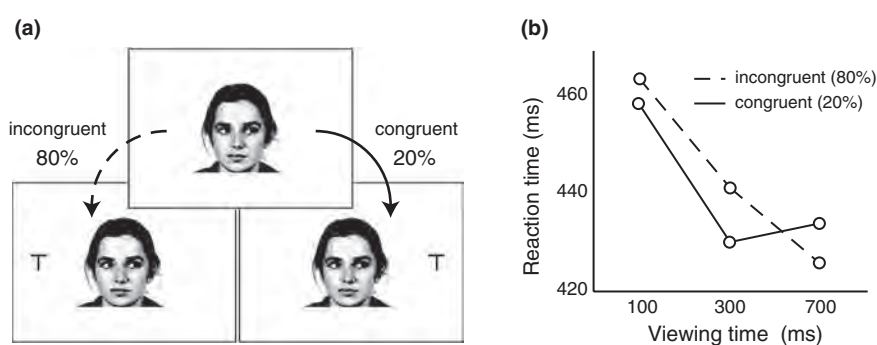


Figure 9.1

Reflexive gaze following. Humans respond faster to targets which appear congruent to previously observed gaze, even when targets consistently appear at the incongruent location (from Shepherd, 2010, based on Driver et al., 1999).

results indicate that not all gaze following responses are susceptible to willful modulation. Perhaps more importantly, these findings show that gaze following requires only those cognitive resources that arise within a few hundred milliseconds of stimulus onset.

The simplest act of gaze following would be to identify a directional visual pattern—an averted nose, or the offset concentric circles of averted eyes—and to shift attention in that direction until a thing of interest is seen. It could be fast, simple, and purely reflexive: one needn't even know that one had done it. Moreover, it might be automatic: one might be unable to stop it, even if one knew the cue would be misleading. Gaze following thus might be a fully modular behavior: reflexive, and once triggered, unalterable.

Joint attention requires more than merely following gaze or otherwise replicating another's attention: one must, at the very least, simultaneously attend to the cuing individual. This suggests we should dissociate reflexive gaze following (sometimes termed "gaze cuing") from more sophisticated understanding of another's eye movements. Further, it prompts the question of how social perception and representation relate: how "looking" grades into "seeing," or how "attending" grades into "knowing," "wanting," or "believing"—or, saliently, "sharing." Nonetheless, a middle ground intervenes between these two phenomena.

It is our impression that we do not follow gaze *solely* as an unconscious reflex. Upon seeing a group of people excitedly craning their necks, we often strain to determine what they're looking at: our attempt to follow gaze is *goal directed*, and informed by clear *expectations* as to what might and might not be the focus of their excitement—and, moreover, as to whether the focus of their excitement is something about which we truly care. To accurately follow gaze, we interpret another's bodily orientation in spatial relation to our own. This rich understanding of others' bodies and of three-dimensional visual space permits us to take their perspective, following their gaze *geometrically* to objects outside our immediate visual field.

Gaze-following behavior (level 1) in many animals likewise involves these features and so appears to transcend simple reflexes, suggesting a more flexible understanding of other's attentional states. The ability to follow gaze *geometrically*, and to modify gaze following behavior based on *expectations* and *goals*, implies some sort of understanding that others are linked to the subjects of their attention (level 2) and appears to build toward those social abilities implicit in joint attention (levels 3–4). What does the gaze following behavior of animals tell us about their understand-

ing of attention? What, if anything, can we infer about the neural substrates of this understanding?

Gaze Following in Nonhuman Animals

Social attention abilities have been differently explored in humans, primates, and nonprimate animals, respectively emphasizing connections to language, theory of mind, and learning signals. Nonetheless, responses to gaze can, to some extent, be unified in a common framework (Shepherd, 2010). Overwhelming evidence suggests that sensitivity to direct gaze—to being watched—is early developing and phylogenetically widespread, appearing soon after birth in humans (Batki, Baron-Cohen, Wheelwright, Connellan, & Ahluwalia, 2000; Farroni, Csibra, Simion, & Johnson, 2002) and exhibited by diverse fish, reptiles, birds, and mammals (Senju & Johnson, 2009; Sowards & Sowards, 2002). However, while sensitivity to being watched has clear and obvious relevance to survival, the benefit of sensitivity to averted gaze is more subtle. Four factors seem paramount: first, animals must generally coordinate their behavior with that of their predators, prey, or mates; second, social animals must coordinate their movements with one another to maintain group coherence; third, conspecific behaviors may suggest adaptive responses to yet unnoticed environmental features (cf. Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010); finally, social primates communicate and negotiate alliances through facial expression and gaze behavior and may eavesdrop on the signal exchanges of others to better monitor their social environment.

Evidence for gaze following by animals comes from the study of an individual's orienting responses to head or eye postures of an experimenter or of a covertly manipulated conspecific. Ideally, these orienting movements are spontaneous rather than reinforced since in principle any sensible stimulus can be used for conditioning (albeit more easily when stimuli are salient to natural behavior). To date, evidence has strongly supported gaze following abilities in most primate species, as well as some domesticated species, trained marine mammals, and birds.

Following Gaze

For example, Michael Tomasello and colleagues captured the attention of a cue individual by discretely showing it food, then observed the responses of an otherwise oblivious conspecific over the next ten seconds. For each of the five primate species tested (chimpanzees, mangabeys, and three types of macaque), individuals were significantly more likely to look

toward the experimenter after seeing a companion looking in the same direction (Tomasello, Call, & Hare, 1998). Two studies extend this data to other primate groups, providing evidence consistent with gaze following: one found that pair-tested cotton-top tamarins (a new world monkey) gazed in similar directions while viewing experimenter cues and that the pairs' gaze was most strongly correlated among individuals who typically resided together (Neiworth, Burman, Basile, & Lickteig, 2002); another found that ring-tailed lemurs (a prosimian or "primitive" primate) look in similar directions during natural behavior and are especially likely to co-orient with individuals toward whom they've just attended (Shepherd & Platt, 2008). Additionally, several studies have examined the tendency to follow conspecifics' gaze among nonprimate species. Using Tomasello's paradigm, Kaminski and colleagues showed that goats follow other goats' gaze (Kaminski, Riedel, Call, & Tomasello, 2005). Moreover, two types of birds—ravens (Schloegl, Kotrschal, & Bugnyar, 2007) and ibises (Loretto, Schloegl, & Bugnyar, 2010)—appear to follow one another's skyward glances, and, similarly, dogs appear able to use other dogs' attention when searching for hidden food (Hare & Tomasello, 1999).

Most commonly, however, gaze following has been studied using human-produced directional cues such as eye movement, head movement, pointing gestures, tapping, and so on. Animal subjects have sometimes been tested simply for spontaneous orienting behavior but more often have been challenged with an opportunity to seek or avoid food at cued locations. These studies suggest that apes and monkeys (see Rosati & Hare, 2009) and human-reared ravens (Bugnyar, Stöwe, & Heinrich, 2004) spontaneously follow human gaze while dogs (Hare & Tomasello, 1999) and dolphins (Pack & Herman, 2004) have both been reported to follow human gaze when rewarded. Further work has suggested that while many domesticated animals (including cats, dogs, goats, horses, seals, and dolphins; see Miklósi & Soproni, 2006) follow human cues to get food, wild animals may more typically *avoid* food that has been targeted by another (Flombaum & Santos, 2005; Hare & Tomasello, 2004). Together, these findings suggest that sensitivity to gaze direction may be widespread but that animals may learn to use this information in very different ways.

Understanding Gaze: Geometry, Expectation, and Goal Directedness

The question remains, however, what animals truly *understand* about gaze. Animals might co-orient because of an innate and modular gaze following reflex, because they have been conditioned that gaze predicts the location of rewards or salient features, or because they use others' gaze to infer

something about their shared environment or about the others' behavioral or mental states. Psychophysical comparison suggests that monkeys and humans follow head and eye gaze quickly, independently of task relevance, using neural pathways that operate within a few hundred milliseconds (Deaner & Platt, 2003). Such gaze cuing is consistent with a behavioral module, operating with neither volition nor awareness. However, like those of humans, animal gaze responses sometimes betray surprising sophistication.

First, most animals appear to understand others' gaze as a vector within a rich, three-dimensional environment. In response to humans staring behind a barrier, or to a distant object beyond a closer distractor, most tested species have been shown to look toward the true target of the humans' gaze (great apes: Bräuer, Call, & Tomasello, 2005; new world monkeys: Amici, Aureli, Visalberghi, & Call, 2009; Burkart & Heschl, 2006; even ravens: Bugnyar, Stöwe, & Heinrich, 2004). Recent investigation of great apes suggest spatially sophisticated gaze comprehension, with ability increasing from orangutans to gorillas, chimpanzees, and finally bonobos (Okamoto-Barth, Call, & Tomasello, 2007). In only one case—that of a bird, the ibis—have adults of a species been shown to follow gaze but to be confused by barriers (Loretto, Schloegl, & Bugnyar, 2010). On the other hand, developmental trajectories suggest that gaze following arises before a rich spatial representation is present. Human infants begin to co-orient with adults within the first year but only later gain the ability to follow gaze toward objects outside their immediate field of view or hidden behind barriers (Butterworth, 1991; Moll & Tomasello, 2004). Moreover, gaze processing accuracy continues to improve throughout human children's first six years (Doherty, Anderson, & Howieson, 2009). While comparative data are sparse, a report in ravens suggests nonhuman animals follow a similar trajectory (Bugnyar, Stöwe, & Heinrich, 2004).

Second, like humans, many animals appear to have expectations about what they should see when they follow another's gaze: if they find nothing, they do a double take. These expectations appear to be shaped by experience. For example, old-world monkeys and apes following gaze cues toward nothing, check back toward the cue, and then follow gaze again (Bräuer, Call, & Tomasello, 2005; Call, Hare, & Tomasello, 1998; Goossens, Dekleva, Reader, Sterck, & Bolhuis, 2008; see also Horton & Caldwell, 2006); moreover, adult apes have been shown to do so more than infants even after controlling for baseline gaze following frequency (Bräuer, Call, & Tomasello, 2005). These experiments have been performed

in few, if any, nonprimates, and even among primates the data are somewhat equivocal: while apes and old-world monkeys are reported to “double look,” negative results have been reported in new-world monkeys (Amici, Aureli, Visalberghi, & Call, 2009), and the apes vary in the sophistication of their expectations regarding gaze (Okamoto-Barth, Call, & Tomasello, 2007).

Third, despite evidence for fast, reflexive gaze cuing in primates (Deaner & Platt, 2003), it appears that both human and nonhuman animals modulate their gaze following behavior based on context: gaze is followed more readily when coming from informed, familiar, emotionally expressive, high-status, or otherwise exceptional individuals. In reality, the issue of context-sensitivity folds together two distinct questions. First, are fast, reflexive gaze responses the result of a dedicated, encapsulated module for gaze following? Second, can animals flexibly integrate observed gaze with other clues to an animal’s mental state or social significance? The first question is addressed by psychophysical findings that the strength of gaze cuing is modulated by the social status of the observed individual (monkeys: Shepherd, Deaner, & Platt, 2006; humans: Jones et al., 2010), though these effects may spare the very fastest (less than 200 milliseconds) responses (Shepherd & Platt, 2009). These findings limit the potential modularity of any gaze following mechanism (see Shepherd, 2010, for discussion) and suggest that within a split second, gaze following behaviors are modulated by the social relevance of the cue.

Moreover, co-orienting in monkeys appears to be enhanced by specific facial expressions (Goossens, Dekleva, Reader, Sterck, & Bolhuis, 2008; Teufel, Gutmann, Pirow, & Fischer, 2010; cf. Paukner, Anderson, Fogassi, & Ferrari, 2007), suggesting that gaze following in natural behavior is well integrated with other forms of social response. These findings suggest that animals may follow gaze differently depending on a cue’s social context and perceived relevance. In particular, they are reminiscent of results from human infants, in which gaze following behaviors are potentiated by “ostensive” cues suggesting specific communicative intent: infants more strongly follow gaze when associated with direct gaze or infant-directed speech (Farroni, Mansfield, Lai, & Johnson, 2003; Senju & Csibra, 2008). Gaze following behaviors may thus involve components with distinct psychophysical and developmental time courses, and while the fastest responses may be reflexive and modular, later gaze following behaviors (still arising within a second) are sensitive to contextual factors including those which suggest communicative significance.

Gaze Following: Learning and Development

These “advanced” gaze-following behaviors appear to have extended developmental time courses, both in human and nonhuman animals. This suggests that the mechanisms underlying our ability to understand others’ viewpoints, like most cognitive abilities, develop through the interaction of genetic predispositions and learning mechanisms with our experienced social environment. Complicating things further, different components may follow distinct ontogenetic trajectories. For example, monkey gaze following becomes increasingly prominent between five months and one year of age (Teufel, Gutmann, Pirow, & Fischer, 2010). While adult monkeys and chimpanzees habituate to gaze cues directed toward empty space, infants do not (Tomasello, Hare, & Fogleman, 2001), suggesting development increases the discretion and selectivity of gaze following behavior. However, from the start, gaze following is facilitated when cuing individuals issue an expressive “comment” on observed interactions—and this facilitation is strongest for monkey observers around one year of age as they begin to navigate society independently of their mothers (Teufel, Gutmann, Pirow, & Fischer, 2010). Thus, while habituation to useless gaze cues increased over development, modulation by vocal expression peaked at a crucial developmental landmark. Similar complexity has been reported in the development of gaze following by ravens: juvenile ravens habituate to irrelevant cues during simple gaze following, but older ravens following gaze geometrically do not (Schloegl, Kotrschal, & Bugnyar, 2007). In this case, the authors speculate, early-developing gaze following is specialized for predator detection while later-developing geometrical gaze following is specialized for another function: pilfering the hidden food caches of other ravens.

In summary, many primates and some birds and domestic animals appear to be capable of interpreting and responding to gaze—and many other species remain untested (see figure 9.2). Fast gaze cuing may be quick and reflexive, consistent with mediation by an encapsulated neural module shared across (at least) higher primates. However, the spatial and social sophistication of gaze following behavior suggests it is goal directed, flexible, and generally well integrated with other cognitive processes. In all species studied, gaze following behaviors appear to be constructed during early life. However, gaze cues are not used in all the contexts in which we imagine they might be helpful, and the use of gaze as a collaborative signal seems uncommon in the animal world—seemingly restricted to coordinating aggression or antipredator behavior, for example, rather than

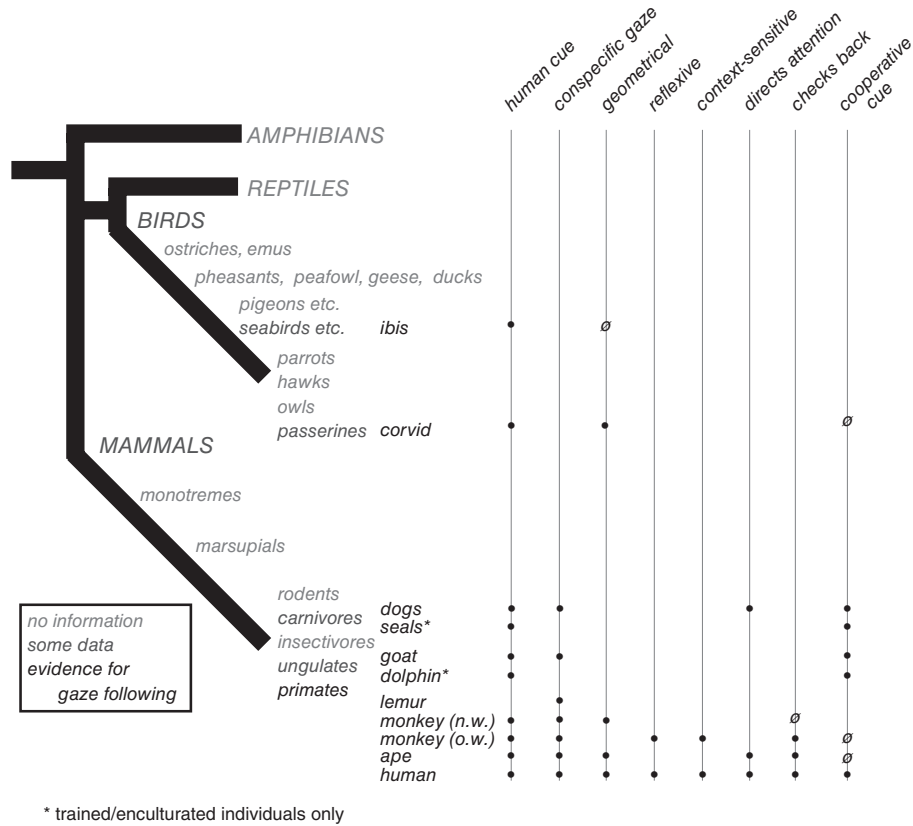


Figure 9.2

Family tree of gaze following abilities. Positive evidence for gaze following abilities is illustrated with a small black dot, negative evidence with a slashed circle. Many species remain untested (adapted from Shepherd, 2010). n.w., new world; o.w., old world.

to acquiring food, resources, or mates. Moreover, animals—and especially adult animals—lack the extraordinary and pervasive interest in others' activities displayed by human children (Tomasello, Carpenter, Call, Behne, & Moll, 2005; but see Pika & Zuberbühler, 2008; Tanner & Byrne, 2010). This suggests that while many wild-reared animals appear to perceive the relational and referential aspects of attention, they have limited ability or opportunity to learn emergent affordances of socially cued attention, with these abilities generally restricted to group defense and coalitional conflict. We shall explore these social affordances in more detail in our consider-

ation of pointing gestures. First, however, let us consider possible neural substrates for forms of socially cued attention which are widespread among nonhuman animals.

Gaze Following by Neural Circuits

When most people think of a brain, they think of the convoluted mass that comprises the human cortex (see figure 9.3). Those of us who study brains may be inclined to overemphasize the uniqueness of the cerebral cortex, a superficial and conspicuous layer of the brain useless without the important and complex subcortical structures beneath it. Moreover, while primates and especially humans have elaborately developed cortices, homologs of mammalian cortex are found within other vertebrate groups (Allman, 1999; Wang, Brzozowska-Prechtl, & Karten, 2010). It is thus

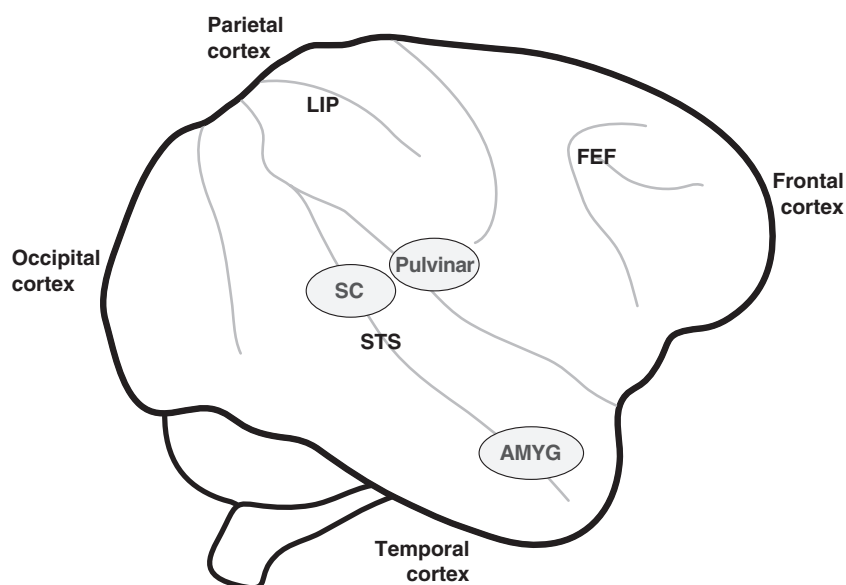


Figure 9.3

Neuroanatomical regions of interest. The pathways used in gaze following remain unknown, but several candidate areas may play a role. In this sketch, lobes are identified in plain text, cortical areas in bold (LIP, lateral intraparietal area; FEF, frontal eye fields; STS, superior temporal sulcus), and subcortical nuclei in outline (SC, superior colliculus, homolog of bird's optic tectum; Pulvinar, pulvinar nucleus of the thalamus; AMYG, amygdala).

possible that the neural mechanisms of gaze following are derived from some gaze following ancestor and are homologous across all occurrences. It appears more likely, however, that circuits for gaze following behavior have developed independently (i.e., are analogous) between divergent species such as birds and primates.

The Subcortical Pathway

The fast time course of reflexive gaze cuing suggests that a fairly short and direct circuit translates gaze observation to orienting movements. It has been theorized that the fastest reactions to basic, biologically crucial stimuli are through an innate subcortical system that is broadly conserved across vertebrates (Johnson, 2005; Sowards & Sowards, 2002). This subcortical system is thought to be interconnected with slower and more nuanced cortical networks, suggesting that while reflexive gaze cuing may be mediated by the former, the latter may mediate the more recently evolved, late developing, or computationally challenging features which characterize naturalistic gaze following.

The mammalian subcortical visual pathway is believed to pass from the retina to the superior colliculus, the pulvinar, and then the amygdala (Jiang & He, 2006; Johnson, 2005; Morris, Ohman, & Dolan, 1999). The amygdala is a sensory convergence zone associated with emotional processing and can influence attention both directly, through modulation of early sensory processing, or indirectly, via projections to social and reward cortices in the temporal and frontal lobes (Amaral, Behnia, & Kelly, 2003; Ghashghaei & Barbas, 2002; Vuilleumier, 2002a). Moreover, the superior colliculus and the pulvinar nucleus of the thalamus influence attention in their own right, serving, respectively, as a final common output pathway for gaze orienting (Platt, Lau, Glimcher, Hall, & Moschovakis, 2004) and as a partner to ongoing cortical processing (Saalmann & Kastner, 2009). While the amygdala is known to be sensitive to observed gaze (monkey electrophysiology: Gothard, Battaglia, Erickson, Spitler, & Amaral, 2007; monkey imaging: Hoffman, Gothard, Schmid, & Logothetis, 2007; human imaging: Kawashima et al., 1999), it is not known whether this information derives from cortical or subcortical signals, nor is it clear whether the amygdala distinguishes between directions of averted gaze or merely between averted gaze and direct staring. While lesions of the amygdala have been reported to disrupt human gaze responses (Akiyama et al., 2007; Okada et al., 2008), these disruptions may be secondary to damaged white matter or decreased social attention.

A homolog of the superior colliculus—the optic tectum—is the principle visual processing center of fish, amphibians, reptiles, and birds (Allman, 1999; Sowards & Sowards, 2002), but in mammals, these subcortical structures are believed to access only the coarsest and least-digested ascending visual input. This makes it unlikely that mammalian subcortical systems are sufficient to mediate spatially sophisticated behaviors such as geometrical gaze following. In contrast, two meta-analyses have identified regions sensitive to gaze direction in human cortex, most notably near the superior temporal sulcus (STS) and in the dorsal and ventral frontoparietal attention networks (Grosbras et al., 2005; Nummenmaa & Calder, 2009).

Social Processing in Mammalian Cortex

While it is possible that social processing centers include cortices homologous across all primates (Rosa & Tweedale, 2005; Tootell, Tsao, & Vanduffel, 2003) and perhaps other mammals (Kendrick et al., 2001), it is clear that great diversification has occurred across species during mammalian evolution. Two of the best studied mammalian species, macaques and humans, are believed to share similar visual systems; even so, the layout of their socially responsive cortices has proved difficult to reconcile (Op De Beeck et al., 2008; Pinsk et al., 2009; Tsao, Schweers, Moeller, & Freiwald, 2008). Broadly speaking, different social processing regions are believed to specialize in features that are relatively static (such as identity) or dynamic (such as expression, gaze, and biological motion; Hoffman & Haxby, 2000); in humans, the former is associated with the fusiform gyrus (Kanwisher & Yovel, 2006) and the latter with the posterior portion of the STS (Allison, Puce, & McCarthy, 2000). In monkeys, most socially responsive cortices appear to lie along the STS, starting deep in the sulcus posteriorly and dipping slightly ventral anteriorly, with several additional patches in the inferior prefrontal lobes (Pinsk et al., 2009; Tsao, Schweers, Moeller, & Freiwald, 2008). The functional specialization of these patches is currently undergoing intensive study and may be mapped, at least in outline, by 2015.

Already, however, anatomical and electrophysiological research in the temporal lobes provides rich information about gaze processing. STS comprises several cortical regions with distinct cellular structure and connectivity, and while posterior regions communicate both with posterior parietal and frontal areas, anterior regions communicate preferentially with frontal and visual cortices (Seltzer & Pandya, 2004; Seltzer & Pandya, 1991). The first neurons sensitive to observed gaze were reported in macaques near the STS (Hasselmo, Rolls, & Baylis, 1989; Perrett et al., 1985; Perrett, Rolls,

& Caan, 1982); moreover, Perrett and colleagues argued that the patchy distribution of gaze-responsive cells corresponded to the patchy distribution of STS neurons communicating with parietal orienting areas (Perrett, Hietanen, Oram, & Benson, 1992). Importantly, neurons in the middle anterior upper bank of the STS represent gaze direction independently of whether it arises through head or eye posture (Perrett, Hietanen, Oram, & Benson, 1992), and while more caudal neurons respond symmetrically to gaze averted to either the right or left, anterior neurons differentiate specific gaze directions (De Souza, Eifuku, Tamura, Nishijo, & Ono, 2005; see also Jellema, Baker, Wicker, & Perrett, 2000).

Human imaging studies support these findings. Though imaging studies have shown peak gaze sensitivity in the posterior STS (Allison, Puce, & McCarthy, 2000; cf. Grosbras et al., 2005; Nummenmaa & Calder, 2009), a recent adaptation study (Calder et al., 2007) showed direction-differentiating neurons concentrate anteriorly. Intriguingly, the STS is especially active in humans when we observe surprising or incongruous gaze behavior (Pelphrey, Singerman, Allison, & McCarthy, 2003): this covert activation may be analogous to overtly observed “check backs” or “double looks.” Human neural pathology provides further support for the role of the temporal lobes in perceiving gaze direction in space: two split-brain patients reflexively followed gaze in only one visual hemifield, consistent with cortical mediation by a single hemisphere (usually the right) specialized for face processing (Kingstone, Friesen, & Gazzaniga, 2000). Moreover, in a rare patient with a lesion affecting the entire right superior temporal gyrus, both gaze perception (Akiyama, Kato, Muramatsu, Saito, Nakachi, et al., 2006) and reflexively gaze following (Akiyama, Kato, Muramatsu, Saito, Umeda, et al., 2006) were absent.

In addition to temporal cortices strongly and selectively activated by faces, macaques appear to have three face-selective regions of frontal cortex not evident in humans (Tsao, Schweers, Moeller, & Freiwald, 2008). Interestingly, however, at least one human subject with lesions in orbitofrontal cortex lacked both gaze cuing and symbolic cuing (Vecera & Rizzo, 2004). While it is difficult to draw conclusions from a single subject, especially given interindividual variability in cuing behaviors, these reports may suggest that those frontal regions which process facial signals in monkeys have been generalized to process more diverse signals in humans. This possibility is especially intriguing given recent evidence that nonpredictive (but socially conventionalized) symbols, such as arrows, evoke reflexive cuing responses in adult humans (Kuhn & Kingstone, 2009; Tipples, 2002; Borjon et al., forthcoming).

It is crucial to recognize that perception of gaze is only a first step in gaze following: any extracted signal must next be converted either directly into an orienting response or indirectly into some sort of flexible knowledge. Moreover, just as attention may arise in other modalities than vision, joint attention might arise through other means than the sight of another's gaze. Our own routine gaze following responses transcend modality (Borjon et al., forthcoming), and blind individuals are capable of sharing attention, interpreting action, and attributing intent using similar brain networks to those of the sighted (Bedny, Pascual-Leone, & Saxe, 2009; Ricciardi et al., 2009; albeit with some developmental delay; see Hobson & Bishop, 2003). Both arguments suggest that gaze following must impact supramodal brain systems which coordinate attention and movement in space: most notably, frontoparietal action networks. These neural systems both direct attention and organize behavior—and have been implicated in understanding the attention and behavior of others.

The Frontoparietal Orienting System

The muscles orienting the eyes are controlled from centers in the brainstem, which receive projections from the superior colliculus and frontal eye fields. These centers, in turn, are interconnected with the lateral intraparietal area (LIP) and supplementary eye fields, and together they comprise the dorsal attention network (Platt, Lau, Glimcher, Hall, & Moschovakis, 2004). This system is likely the final step in converting gaze perception into an orienting response: that is, making the focus of another's orienting system the focus of our own.

This step could occur through a great number of pathways. In macaques, projections from gaze-sensitive regions of the STS to cortical areas 7A and LIP comprise one likely route through which observed gaze could influence orienting (Perrett, Hietanen, Oram, & Benson, 1992; see also Seltzer & Pandya, 2004; Seltzer & Pandya, 1991). LIP is particularly interesting because its neurons respond to both direct stimulation of, or orienting movements toward, their spatial response field (RF; Colby & Goldberg, 1999). Intriguingly, neurons in LIP are further modulated by the behavioral relevance of their RF, reflecting both the juice reward value of trained eye movements (Platt & Glimcher, 1999) and the intrinsic reward of acquiring social information (Klein, Deaner, & Platt, 2008). LIP constitutes a unified spatial saliency map, and neuronal activity is associated both with covert attention and overt motor intention toward a region (Colby & Goldberg, 1999).

Shepherd and colleagues (Shepherd, Klein, Deaner, & Platt, 2009) recorded the activity of LIP neurons while monkeys performed a simple test of gaze cuing. Monkeys fixated a central location, a face appeared gazing one direction or another, and then a target appeared unpredictably in either the cued (“congruent”) or uncued (“incongruent”) direction. When trials were congruent, monkeys were only slightly faster than when trials were incongruent—perhaps because the cuing effect is smaller in monkeys than in humans or perhaps because these monkeys had performed many thousands of trials and were practiced at suppressing the irrelevant gaze-cuing response. Nonetheless, gaze-cuing effects were significant. Classically, LIP neurons would be expected to ignore the gaze cue and instead fire only when targets appeared in their RF: one great burst bridging target onset and saccade initiation. Instead, we found that a large subpopulation of neurons had a small but significant response to gaze cues presented outside, but directed toward, their RF. Generally, these responses began within 100 to 200 milliseconds of the cue onset and persisted through the cue period. We saw some neurons that were enhanced and some neurons that were suppressed by gaze. The population that was enhanced responded with a time course quite similar to the behaviorally observed gaze following response, peaking between 200 and 400 milliseconds and decaying by 800 milliseconds. The inhibited population was tonically active throughout the cue period.

How are we to interpret these findings? The enhanced neurons would be expected to contribute to gaze following, the suppressed neurons would be expected to dampen it—and both effects were plausible, given that the monkeys had been extensively trained to ignore the irrelevant gaze cue. Close examination of the modulation dynamics suggested that the enhanced neurons may have been secondary players, however. If LIP were mediating the fastest and most reflexive gaze-cuing responses, enhanced activity would have led the behavioral effect by thirty milliseconds: it did not. In summary, these findings support the claim that gaze cues modulate LIP neurons from outside their classical RFs and suggest that observed gaze changes the value of orienting toward the cued location. Thus, even if LIP is not involved in the fastest gaze cuing effects, it is almost certain to contribute to the influence of social context on gaze following (e.g., Goossens, Dekleva, Reader, Sterck, & Bolhuis, 2008; Shepherd, Deaner, & Platt, 2006; Teufel, Gutmann, Pirow, & Fischer, 2010).

Consistent with these findings, human neurons that track observed gaze directions are clustered in two places: STS and posterior parietal cortex (Calder et al., 2007). However, homology between macaque and human

brains remains uncertain, with the relevant cortical areas greatly expanded in humans relative to macaques (Orban, Van Essen, & Vanduffel, 2004). Complicating things further, there is not just one but two frontoparietal attention systems: dorsal and ventral (Corbetta & Shulman, 2002). While monkey electrophysiology implicates the dorsal orienting system, human joint attention tasks often activate a right-lateralized ventral network that strongly resembles the ventral attention network identified by Corbetta and Shulman—including the right temporoparietal junction and right lateral prefrontal cortex—along with medial frontal cortices (Calder et al., 2002; Grosbras, Laird, & Paus, 2005; Nummenmaa & Calder, 2009; Saxe, 2006). Monkey homologs of these areas remain uncertain. Finally, a study of human patients suffering visuospatial neglect suggests that these orienting networks—or at least, the parietal components thereof—are not essential to reflexive gaze cuing. In fact, not only did these patients exhibit normal gaze-cuing effects, they could be induced to follow gaze from their preserved hemifield into that which they more typically neglected (Vuilleumier, 2002b).

Reflecting Gaze

To follow someone's gaze is to replicate their direction of attention. Such a shift in our attentional state can be achieved overtly and observably (by changing our bodily orientation) or covertly and privately (by changing solely our neural activity). Changes in overt orienting alter the physical coupling between our bodies, brains, and environment and, therefore, necessarily alter the contents of our sensory processing. Likewise, our covert attentional deployments alter our cognition, affecting future orienting decisions and subtly influencing our postural stance (e.g., microsaccadic drift: Engbert & Kliegl, 2003; Hafed & Clark, 2002). Moreover, it now appears that we use overlapping neural tissues to physically orient our eyes and to covertly shift our attention (e.g., Moore & Fallah, 2004). Because attention systems receive inputs from many cortical areas, there are many pathways through which social cues could exert their influence on gaze, and with no clear rubric for estimating which route might be faster or more flexible than another.

Importantly, however, the orienting system is embedded in a larger cortical network for interacting with the environment. A broad family of frontoparietal circuits link the dorsal (spatial or action-oriented) visual system and exterior body-sense areas, both associated with the parietal lobes, to frontal systems that evaluate, organize, and execute body movements. Seen through this lens, gaze following is but one of many

mechanisms which couple individuals' behavioral states. Contagion between observed and enacted behavior has been described for skeleto-motor action (Blakemore & Frith, 2005), facial expression (Dimberg & Thunberg, 1998), affective state (Hatfield, Cacioppo, & Rapson, 1993), sensory orienting and attention (Shepherd, 2010), and even perceptual states (Borjon et al., forthcoming). It is important to note that while this coupling is mediated through observable body states, it operates at a certain level of abstraction: our goals and emotions are more faithfully replicated than the postures or communication signals which conveyed them (Tamietto et al., 2009).

Some of the most interesting neural devices which may play a role in relating observed and executed behavioral states are the "mirror" neurons found in the prefrontal (area F5; Rizzolatti et al., 1996; Gallese et al., 1996) and parietal (inferior parietal lobule, mainly area PF; Fogassi et al., 2005; Gallese et al., 2002) cortices of macaque monkeys. These neurons respond to the execution of hand and mouth actions at various levels of abstraction, often sensitive to action goals rather than mere postural shifts (Umiltà et al., 2001, 2008). Moreover, these neurons are accompanied by "canonical neurons," which fire both during actions and when action-relevant objects are presented: they appear to categorize affordances, or potential actions, rather than actions currently being performed (Murata et al., 1997). Since they conflate self-produced and witnessed behaviors, mirror neurons are sometimes claimed as a crucial mechanism through which we interpret the behavior of others (e.g., by using our own motor capabilities to support a first-person subjective understanding of another's action goals; Di Pellegrino et al., 1992; Rizzolatti et al., 2001; Gallese, 2005a). This hypothesis is hotly debated. While there are various lines of evidence that that one's own motor expertise supports the understanding of others' actions (Beets et al., 2010) and that mirror neurons are both selectively and flexibly recruited during the recognition of goal-oriented actions (Gazzola et al., 2007; Cattaneo et al., 2007), objections have been raised on both philosophical and empirical grounds. These objections cite a lack of evidence that mirror neuron play a causal role in action recognition (Hickok, 2009) and claim no existing model can adequately explain how neurons mirroring another's action can support understanding of its intent (Borg, 2007; Csibra, 2007). Alternative explanations have also been suggested for mirror neurons: they may be produced by Pavlovian association, for example (Mahon & Caramazza, 2008; Catmur, Gillmeister, Bird, Liepelt, Brass, & Heyes, 2008) or may predict the next move an agent will perform to achieve a previously represented

goal (Jacob, 2008). (A rebuttal to these objections can be found in Rizzolatti & Sinigaglia, 2010.)

Shepherd et al.'s 2009 report of gaze "mirror" neurons makes this debate immediately relevant to the phenomenon of joint attention. Shepherd and colleagues identified neurons which were active both when a monkey looked a certain direction and when it saw a photographed monkey looking the same direction. If these neurons are, in fact, a kind of mirror neuron, then mirror processes may operate both for directly manipulative acts and for those (e.g., gaze) whose physical enaction is subtle and distal and whose consequences are covert. In our discussion of comparative gaze following abilities, we identified three markers of sophistication in gaze following: *geometry*, *expectation*, and *goal directedness*. Gaze mirror neurons could, in principle, accommodate each of these features. Like frontoparietal motor mirror neurons, gaze mirror neurons would be expected to operate at a certain level of abstraction: they should be sensitive to the environmental locus of attention rather than to the specific physical posture implementing a direction of gaze (e.g., eyes alone vs. head and eyes together). Similarly, because mirror neurons inhabit motor cortices focused on sensorimotor transformation across diverse bodily reference frames, it would be surprising if they were incapable of mediating geometrical gaze following. If mirror neurons, like canonical neurons, identify environmental affordances, then we should expect that they are associated with strong expectations. Finally, because the frontoparietal orienting networks intervene between perceptual and motoric cortices, we would expect minimal encapsulation, suggesting mirror-mediated gaze following would be highly context-sensitive and minimally reflexive.

To review, many animals have a rich capacity to respond to others' attention, as evident in their tendency to follow gaze (level 1) with their geometrical sophistication, expectations regarding where other animals will look, and context-dependent and goal-directed utilization of gaze cues (level 2). These abilities arise early in development, are sensitive to experience, and may be mediated by several neural pathways. Monkeys, like humans, exhibit both fast and reflexive gaze cuing and slower and more nuanced gaze following behavior. In monkeys, neurons in the LIP appear to differentiate gaze directions, with some acting to "mirror" gaze and some acting in opposition to gaze following. These data suggest that gaze is represented in the frontoparietal orienting system, which seems well-suited to mediate geometrical gaze following in a flexible and goal-directed manner (level 2). It is worth noting that if theorists are

correct in claiming mirror neurons permit a preconceptual and nondeclarative attribution of motor intentions (Gallese, 2005b), then they could support an understanding of nonrepresentational gaze as a form of motor intention. Gaze toward an object significantly increases the likelihood it will be acted upon, and while attention is not a strict precondition for transitive action, it is strongly associated with various forms of motor intent. Furthermore, gaze plays a significant role in the dyadic signaling of nonhuman primates and as such plays an *instrumental* role in signaling intention to interact with a target, typically in the context of aggression (gaze enacts a *threat*) or mating (gaze enacts *courtship*). For these reasons, it is not surprising that systems processing motor intention could be generalized to gaze signals (level 2). Crucially, by tracking the social contingencies necessary for species-typical triadic behaviors, these neurons could provide a nonrepresentational pathway for joint attention (level 3): *challenge* individual B *for* C, *team up with* individual B *against* individual C, *share* object C with individual B. The first two situations arise frequently (and the latter, rarely) in the social lives of group-living animals and comprise an important part of their natural behavioral repertoire.

If uniquely social affordances suggest a nonrepresentational mechanism for joint attention in animals, then how does this form of joint attention compare with that found in humans? A key feature of joint attention, as expressed by humans, is the ability to manipulate the focus of joint attention through gesture. However, this ability is not completely unique: animals, too, direct the attention of others through communicative signals. This aspect of socially cued attention—its active *exploitation* to meet behavioral goals—will be the subject of our final section.

Sharing Attention through Gesture

Generally, some form of joint attention is said to arise whenever two observers are simultaneously and symmetrically aware of a common stimulus and of the other's attention toward that stimulus (e.g., Moore & Dunham, 1995). Because gaze following can synchronize attentional shifts toward a common stimulus, it is likely to play a pivotal role in achieving joint attention, provided two additional conditions can be met: first, the observed attention must be recognized as intentional in character; second, mutuality must somehow be recognized, for example, through unique social affordances or mentalistic representations permitting an inference of symmetry.

Developmental studies suggest that a human infant's ability to follow and subsequently to direct an adult's gaze predicts the infant's acquisition of full-blooded joint attention abilities over the following months (Carpenter, Nagell, & Tomasello, 1998). Joint attention among humans—even among young children—appears to involve rich representational systems and cooperative motivations that are rarely claimed of animals. While many animals appear to follow gaze, and while many appear to do so in a flexible and sophisticated manner suggesting they recognize gaze as intentional, they seem to lack the pervasive interest in others' attention which characterizes humans, and they appear to recognize social affordances only in limited and stereotyped contexts. Similarly, human infants younger than nine to twelve months of age seem incapable of jointly attending despite possessing the relevant perceptual and motor competencies from about the fifth month (Tomasello, 1999, p. 66). Something dramatic seems to change early in typical human development—and only atypically (if ever) in animal development—to promote the sharing of attention. How close do animals come, then, to achieving “full-blooded” joint attention? What additional features are necessary to construct human-like abilities from those building blocks we share with other species? Though the ancestral origin of our joint attention is highly uncertain, our ritualization of reaching gestures to manipulate others' attention (along with increasing reliance on imitative learning) seems to have played a crucial role (Tomasello, 1999, pp. 87–88).

Joint Attention and Deixis in Humans

The human experience of joint attention, after all, entails much more than merely *following* it. The key difference is that the focal stimulus becomes “mutually manifest” to both coattenders (Tomasello, 1995; Clark, 1996). This mutuality implies that the attentional focus is openly accessible to each coattender and becomes salient by reason of a “common ground”: a shared context defined by the personal and cultural background of the individuals involved. In Tomasello's view (1995, 1999, 2008), both recognition of mutuality and conceptualization of the common ground are mentalistic in character, requiring perspective-taking and simulation skills believed to be unavailable to nonhuman primates (see also Meltzoff & Gopnik, 1993): this account builds on the intuition that while apes might only understand the meaning of a deictic gesture from the point of view of the signaler (“what the communicator wants for herself”), humans also understand the gesture from a self-reflexive point of view (“what the communicator thinks the recipients want for themselves”). Such an

understanding necessitates a mutually known framework of goals and interests: the common ground. This notion of joint attention is recursive in nature (Tomasello, 2008, pp. 94–95, regarding Grice, 1957, and Sperber & Wilson, 1986) and has been challenged as requiring infinite regress (Schiffer, 1972, and Campbell, 2005).

Tomasello argues (1999, 2008) that the crucial behavioral signature of joint attention, as experienced by humans, is the use of indicative gesture for the sole purpose of sharing attention. These gestures are arbitrary and produced merely to inform or declare: pointing something out or holding an object just to “show” it. The prototypical gesture in this family—index-finger pointing—is present in all known human cultures (Kita, 2003) and emerges in infancy at twelve months (Liszkowski, Carpenter, Henning, Striano, & Tomasello, 2004). The extension of a finger toward a distant object is purely deictic and expressly communicative: it suggests no goal but communication, achieves no independent aim, and conveys no intrinsic content or affective valence. In contrast, both deictic gaze behaviors and proximal reaches accomplish personal objectives, and gaze, in particular, has affective gloss on account of its instrumental role in dyadic interaction and its integration into facial expressions. Human pointing is thus ostensive and arbitrary, given significance only by virtue of situational features and the presence of a common ground.

In this manner, human pointing can convey just about anything; it can even be performed to refer symbolically to an absent object or, when targeting a present object, to indicate an incidental feature or abstract association. In fact, human pointing is so conventionalized that it can be fully abstracted from the details of the immediate environment and given meaning solely through manipulation of the common ground: pointing can direct the imagination, designating and referencing locations in an arbitrary symbolic space (Kendon, 2004). Such gestural communication has clear ties to language (Corballis, 2003; Deacon, 1997), and indeed, early fluency with pointing and showing gestures predicts language competency later in infancy (Mundy et al., 2007); conversely, early disinterest in pointing and showing is a crucial diagnostic feature of childhood autism (American Psychiatric Association, 1994, p. 66).

In summary, human pointing gestures can be purely deictic, expressly communicative, arbitrarily conventionalized, altruistically/communally motivated, informative/declarative in character, and given meaning by dint of a representational common ground requiring full-blooded joint attention (level 4). Such manipulation of joint attention grades naturally into symbolic communication: use of conventionalized, arbitrary, and public signals.

Joint Attention and Deixis in Animals

Animals, too, direct the attention of others. Many animals respond to specific threats through coordinated counterattack: for example, blue jays and mockingbirds mob larger birds such as owls and ravens in order to drive them off. Coordinating these responses, many animals have sophisticated systems of alarm calls: vervet monkeys, for example, distinguishing between (roughly) avian, mammalian, and ophidian threats (Cheney & Seyfarth, 1990). Producing certain types of alarm calls—upon seeing a snake or rival social group, for example—monkeys perform stereotyped “pointing” movements in which they shift position and abruptly freeze, staring intently at the target (personal observation, SS). Macaques similarly solicit third-party support in intragroup conflicts, using a series of alternating glances while loudly threatening a foe, and do so more often when the result of one-on-one combat is uncertain (“appeal-aggression”; de Waal, 2003). Because they help coordinate defense against adversaries, these signals are likely widespread among group-living animals.

However, while these signals appear to be both ostensive and deictic in character, they contrast with human pointing in several respects (Tomasello, 2008, pp. 34–41). Crucially, they are not arbitrary, as each signal intrinsically and specifically implies some immediate goal or motivation: for example, the ostensive signals generally convey affect (e.g., alarm calls), and the deictic signals, motor intention (e.g., staring for surveillance, tensely freezing in readiness to move). In several nonhuman primate species, evidence exists that hand gestures have been ritualized: capuchin monkeys make a food-cupping gesture to beg, even when the food is out of reach (de Waal, 2003). Moreover, the great apes appear to use open-handed reaching as a begging gesture even in nonfood contexts or situations in which the desired object is mechanically abstracted from the gestural target (de Waal, 2003). Chimpanzees, in particular, can point to direct another's attention (Leavens & Hopkins, 1998; Leavens, Hopkins, & Bard, 2005), but this behavior has been observed primarily in soliciting a human recipient to bring them some food they can't reach. Orangutans can also point to indicate a tool necessary to retrieve food (Call & Tomasello, 1994), but—differently from human children—they do it only if the object is of interest to their own behavioral goals rather than to their companion's (Bullinger, Zimmermann, Kaminski, & Tomasello, 2011). These findings suggest that even among the apes, deictic gestures rarely serve as an altruistically informative or communally declarative gesture (but see Leavens, this volume). Instead, these behaviors appear to be unified in their specific linkage to concrete, well-defined behavioral

responses. Such deictic behaviors can be termed “instrumental” or “imperative,” in that they result in specific physically observable (and generally selfish) end points.

This absence of altruistic/communal, informative/declarative gesture in nonhuman signalers may explain otherwise curious behavior. While many animals follow gaze, few use gaze appropriately in choice tasks. Apes appear to be able to read a human’s direction of selfish interest (Tomasello, Hare, & Agnetta, 1999), but in object choice tasks, in which food is hidden in one out of many possible locations, a human’s informative pointing is simply not understood (Tomasello, Call, & Gluckman, 1997). This fact is imputed by Tomasello (2008, p. 41, 2009) to the apes’ inability to attribute altruism. Interestingly, when informative/declarative pointing is distinguished from imperative/instrumental pointing as a predictor of human language ability, it is only the former sort—the sort rare or absent among nonhuman animals—that predicts language acquisition (Mundy et al., 2007).

Nonprimates, too, exhibit pointing behaviors. For example, dogs appear to actively direct human attention (Miklósi, Polgardi, Topal, & Csanyi, 2000; see also Miklósi et al., 2003), perhaps using skills crucial to group-hunting wolf ancestors (Miklósi & Soproni, 2006) and repurposed by human breeding (Hare, Brown, Williamson, & Tomasello, 2002) or early-life enculturation (Udell, Dorey, & Wynne, 2010). While formal studies have focused on pointing to acquire food rewards, dogs have been bred and trained to point for more abstract purposes (e.g., to spot birds or bombs).

Moreover, we may easily underestimate the use of deictic signals: it is challenging to distinguish the communicative significance of a signal which is not consistently coupled with triggering situations or behavioral responses. Instead, we might judge the signal to be merely a generic “greeting,” “contact call,” or “territorial marker.” For example, through a lengthy investigation, Teufel and colleagues (Teufel, Gutmann, Pirow, & Fischer, 2010) found that a generally overlooked “comment” call—most often produced by third parties witnessing a social interaction—appears to increase the likelihood their gaze will be followed by a monkey audience.

If it is true that we underestimate the rate of ostensive signaling by nonhuman animals, might we also underestimate the extent to which they have cooperative intent? Consider, for example, the case with which we opened this chapter. Among Barbary macaques, an extended audiovisual facial expression involving repetitive lip/jaw movements (the “chatter”) is “part of a typical behavioral pattern called triadic interaction or agonistic

buffering ... where usually two adults sit in body contact, hold one infant or little juvenile, and *chatter*. The adults approach their heads normally to the child while *chattering*, often lifting and lowering the head or turning the head from side to side (*head flag*). Frequently the animals *hug* and *knead* each other. Often they *chatter at* the genitalia of the child and touch the genitalia of the infant with lips or hands (*touch genitalia*)" (Hesler & Fisher, 2007, pp. 170–171). This behavioral pattern (see figure 9.4) is particularly interesting because in contrast to many triadic behaviors, it appears to have an abstractly prosocial goal. Nonetheless, these cases of triadic signaling among monkeys appear to lack arbitrariness and to be deployed in relatively limited and species-typical contexts rather than being used flexibly for informative or declarative purposes. In particular, there is little evidence to date that signaling behaviors are sensitive to the knowledge state of potential recipients (Cheney & Seyfarth, 1990). This negative finding challenges the notion that nonapes signal with the intention of informing or declaring to others. By contrast, great apes do appear somewhat sensitive to recipient knowledge: they generate more deictic gestures to guide a



Figure 9.4

Barbary macaques sit together, inspecting a nursing infant. Image courtesy Sébastien Wautelet.

human's behavior if there is evidence that the human is ignorant (Zimmermann, Zemke, Call, & Gomez, 2009). However, typically reared apes appear to do so only to obtain food rewards rather than as part of an altruistic, communal, or otherwise prosocial goal. The core finding seems to be that nonhuman animals lack the expectation that their peers will collaborate with them in flexible, rather than limited and species-typical, contexts. Even where sensitivity to knowledge states exists, nonhuman animals appear to have little incentive to shape others' knowledge for communal (as opposed to personal) ends.

Finally, the distinction between imperative and informative/declarative *production* of pointing implies a subtle difference in how humans and nonhumans *respond* to social attention cues. Human signalers actively shape their gestures so that they may be easily decoded by their target. As a result, an optimal human recipient needs to understand not only the *communicator's* intent but also the communicator's understanding of the recipient's own intent! According to Tomasello, our pervasive tendency to see ourselves from other points of view is crucial to our ability to cooperate in arbitrary endeavors. It is interesting to consider the possibility that attentional coordination through hand gesture may have been a crucial step in the evolution of "full blooded" joint attention (level 4). Manual gestures, unlike gaze gestures, can be simultaneously perceived by both signaler and recipient, affecting both in a similar (one might say mutual) fashion. Pointing, unlike gaze, is expressly ostensive in that it does not directly achieve a personal goal. Finally, pointing frees the eyes to move independently between the attentional target and any coattender (see Hopkins & Tagliatela, this volume, for more on this theme).

Conclusions

Human joint attention abilities are crucial to our social behavior and appear fully integrated with mentalistic representation, language, and other cognitive abilities thought uniquely human. These abilities did not appear *de novo*, however, and are constructed atop underlying behaviors and neural circuits we share with other species. In particular, many organisms have the ability to respond behaviorally to another individual's attention—for example, by following its gaze into distant space. Moreover, many species (including corvid birds, domesticated mammals, and nonhuman primates) appear to follow gaze with geometrical sophistication, and, moreover, can do so in a flexible, goal-directed, and context-sensitive fashion. Perception of gaze may guide behavior, permitting uniquely social

affordances to arise: social animals often act jointly to drive away adversaries. These features suggest animals recognize gaze as a kind of intentional relationship and can coordinate species-typical social behaviors in response to these relationships.

Different animals may use very different neural systems to perceive and respond to gaze. Subcortical mechanisms may play a role in fast and stereotyped gaze responses. However, mammalian cortex plays a clear role in perceiving, interpreting, and responding flexibly to observed gaze. In particular, Shepherd and colleagues (2009) identified parietal sensorimotor neurons which fire both when attending to a region of space and when seeing another individual gaze there. Neurons in this area would be well situated to interpret gaze in three-dimensional space and to flexibly integrate information about observed gaze into ongoing attentional decisions. These “mirror” responses may play a role in translating social contingencies into species-typical behaviors such as *mob our attacker*. Such uniquely social affordances suggest a kind of nonrepresentational joint attention which coordinates joint action but is limited to species-typical social repertoires.

These mechanisms are likely the foundation upon which human joint attention evolved. While nonhuman animals can manipulate the attentional states of others, humans do so through specialized gestures used for no other purpose. While other animals appear to engage in joint attention, only humans have a flexible repertoire of cooperative behaviors facilitating informative and declarative pointing gestures. The development of the human pointing gesture appears intimately connected with representational joint attention abilities and likely played a pivotal role in the development of symbolic signaling including human speech.

Acknowledgments

We thank Ipek Kulahci and Matthew Slayton for their helpful comments. SS is supported by a Princeton University Training Grant in Quantitative Neuroscience, NRSA T32 MH065214–1. MC is supported by a Postdoctoral Research Fellowship from Valente Center for Arts and Sciences, Bentley University—Waltham, Massachusetts.

References

Akiyama, T., Kato, M., Muramatsu, T., Saito, F., Nakachi, R., Kashima, H., et al. (2006). A deficit in discriminating gaze direction in a case with right superior temporal gyrus lesion. *Neuropsychologia*, 44, 161–170.

Akiyama, T., Kato, M., Muramatsu, T., Saito, F., Umeda, S., Kashima, H., et al. (2006). Gaze but not arrows: A dissociative impairment after right superior temporal gyrus damage. *Neuropsychologia*, *44*, 1804–1810.

Akiyama, T., Kato, M., Muramatsu, T., Umeda, S., Saito, F., Kashima, H., et al. (2007). Unilateral amygdala lesions hamper attentional orienting triggered by gaze direction. *Cerebral Cortex*, *17*, 2593–2600.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, *4*, 267–278.

Allman, J. M. (1999). *Evolving brains*. New York: Freeman.

Amaral, D. G., Behniea, H., & Kelly, J. L. (2003). Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. *Neuroscience*, *118*, 1099–1120.

American Psychiatric Association. (1994). *Autistic disorder*. Washington, DC: American Psychiatric Association.

Amici, F., Aureli, F., Visalberghi, E., & Call, J. (2009). Spider monkeys (*Ateles geoffroyi*) and capuchin monkeys (*Cebus apella*) follow gaze around barriers: Evidence for perspective taking? *Journal of Comparative Psychology*, *123*, 368–374.

Batki, A., Baron-Cohen, S., Wheelwright, S., Connellan, J., & Ahluwalia, J. (2000). Is there an innate gaze module? Evidence from human neonates. *Infant Behavior and Development*, *23*(2), 223–229.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 11312–11317.

Beets, I. A., Rösler, F., & Fiehler, K. (2010). Nonvisual motor learning improves visual motion perception: Evidence from violating the two-thirds power law. *Journal of Neurophysiology*, *104*, 1612–1624.

Blakemore, S. J., & Frith, C. (2005). The role of motor contagion in the prediction of action. *Neuropsychologia*, *43*, 260–267.

Borg, E. (2007). If mirror neurons are the answer, what was the question? *Journal of Consciousness Studies*, *14*(8), 5–19.

Borjon, J., Shepherd, S. V., Todorov, A., & Ghazanfar, A. A. (in press). Eye-gaze and arrow cues influence elementary sound perception. *Proceedings, Biological Sciences*.

Bräuer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology*, *119*, 145–154.

Bugnyar, T., Stöwe, M., & Heinrich, B. (2004). Ravens, *Corvus corax*, follow gaze direction of humans around obstacles. *Proceedings, Biological Sciences*, *271*, 1331–1336.

- Bullinger, A., Zimmermann, F., Kaminski, J., & Tomasello, M. (2011). Different social motives in the gestural communication of chimpanzees and human children. *Developmental Science*, 14, 58–68.
- Burkart, J., & Heschl, A. (2006). Geometrical gaze following in common marmosets (*Callithrix jacchus*). *Journal of Comparative Psychology*, 120, 120–130.
- Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), *Natural theories of minds* (pp. 223–232). Oxford: Blackwell.
- Calder, A. J., Beaver, J. D., Winston, J. S., Dolan, R. J., Jenkins, R., Eger, E., et al. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology*, 17, 20–25.
- Calder, A., Lawrence, A., Keane, J., Scott, S., Owen, A., Christoffels, I., et al. (2002). Reading the mind from eye gaze. *Neuropsychologia*, 40, 1129–1138.
- Call, J., Hare, B. A., & Tomasello, M. (1998). Chimpanzee gaze following in an object-choice task. *Animal Cognition*, 1, 89–99.
- Call, J., & Tomasello, M. (1994). The social learning of tool use by orangutans (*Pongo pygmaeus*). *Human Evolution*, 9(4), 297–313.
- Campbell, J. (2005). Joint attention and common knowledge. In N. M. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds: Issues in philosophy and psychology* (pp. 205–229). Oxford: Clarendon Press.
- Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, 63(4), i–vi, 1–143.
- Catmur, C., Gillmeister, H., Bird, G., Liepelt, R., Brass, M., & Heyes, C. (2008). Through the looking glass: Counter-mirror activation following incompatible sensorimotor learning. *European Journal of Neuroscience*, 28, 1208–1215.
- Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., et al. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 17825–17830.
- Cheney, D. L., & Seyfarth, R. M. (1990). *How monkeys see the world: Inside the mind of another species*. Chicago: University of Chicago Press.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22, 319–349.
- Corballis, M. C. (2003). From mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences*, 26, 199–208, discussion 208–260.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, 3, 201–215.

Csibra, G. (2007). Action mirroring and action understanding: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Attention and performance* (Vol. 22, pp. 435–459) Oxford: Oxford University Press.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: Norton.

Deaner, R. O., & Platt, M. L. (2003). Reflexive social attention in monkeys and humans. *Current Biology*, 13, 1609–1613.

De Souza, W. C., Eifuku, S., Tamura, R., Nishijo, H., & Ono, T. (2005). Differential characteristics of face neuron responses within the anterior superior temporal sulcus of macaques. *Journal of Neurophysiology*, 94, 1252–1266.

de Waal, F. B. (2003). Darwin's legacy and the study of primate visual communication. *Annals of the New York Academy of Sciences*, 1000(1), 7–31.

Dimberg, U., & Thunberg, M. (1998). Rapid facial reactions to emotional facial expressions. *Scandinavian Journal of Psychology*, 39(1), 39–45.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, 91, 176–180.

Doherty, M. J., Anderson, J. R., & Howieson, L. (2009). The rapid development of explicit gaze judgment ability at 3 years. *Journal of Experimental Child Psychology*, 104, 296–312.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, 6, 509–540.

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, 43, 1035–1045.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 9602.

Farroni, T., Mansfield, E. M., Lai, C., & Johnson, M. H. (2003). Infants perceiving and acting on the eyes: Tests of an evolutionary hypothesis. *Journal of Experimental Child Psychology*, 85, 199–212.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, 15, 447–452.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science*, 308, 662–667.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5, 490–495.

Gallese, V. (2005a). The intentional attunement hypothesis: The mirror neuron system and its role in interpersonal relations. In S. Wermter, G. Palm, & M. Elshaw (Eds.), *Biomimetic neural learning for intelligent robots: Intelligent systems, cognitive robotics, and neuroscience* (pp. 19–30). Berlin: Springer-Verlag.

Gallese, V. (2005b). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4, 23–48.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593–609.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance* (Vol. 19, pp. 247–266). Oxford: Oxford University Press.

Gazzola, V., Van Der Worp, H., Mulder, T., Wicker, B., Rizzolatti, G., & Keysers, C. (2007). Aphasics born without hands mirror the goal of hand actions with their feet. *Current Biology*, 17, 1235–1240.

Ghashghaei, H. T., & Barbas, H. (2002). Pathways for emotion: Interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Science*, 115, 1261–1279.

Goossens, B., Dekleva, M., Reader, S., Sterck, E., & Bolhuis, J. (2008). Gaze following in monkeys is modulated by observed facial expressions. *Animal Behaviour*, 75, 1673–1681.

Gothard, K., Battaglia, F., Erickson, C., Spitler, K., & Amaral, D. (2007). Neural responses to facial expression and face identity in the monkey amygdala. *Journal of Neurophysiology*, 97, 1671–1683.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66, 377–388.

Grosbras, M., Laird, A. R., & Paus, T. (2005). Cortical regions involved in eye movements, shifts of attention, and gaze perception. *Human Brain Mapping*, 25(1), 140–154.

Hafed, Z. M., & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research*, 42, 2533–2545.

Hare, B., & Tomasello, M. (1999). Domestic dogs (*Canis familiaris*) use human and conspecific social cues to locate hidden food. *Journal of Comparative Psychology*, 113, 173–177.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, 68, 571–581.

- Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science*, 298, 1634–1636.
- Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32, 203–218.
- Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, 2(3), 96–99.
- Hesler, N., & Fisher, J. (2007). Gestural communication in Barbary macaques (*Macaca sylvanus*): An overview. In J. Call & M. Tomasello (Eds.), *The gestural communication of apes and monkeys* (pp. 159–195). Mahwah, NJ: Erlbaum.
- Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21, 1229–1243.
- Hobson, R. P., & Bishop, M. (2003). The pathogenesis of autism: Insights from congenital blindness. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358, 335–344.
- Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, 3, 80.
- Hoffman, K. L., Gothard, K. M., Schmid, M. C., & Logothetis, N. K. (2007). Facial-expression and gaze-selective responses in the monkey amygdala. *Current Biology*, 17, 766–772.
- Horton, K. E., & Caldwell, C. A. (2006). Visual co-orientation and expectations about attentional orientation in pileated gibbons (*Hylobates pileatus*). *Behavioural Processes*, 72, 65–73.
- Jacob, P. (2008). What do mirror neurons contribute to human social cognition? *Mind & Language*, 23(2), 190–223.
- Jellema, T., Baker, C. I., Wicker, B., & Perrett, D. I. (2000). Neural representation for the perception of the intentionality of actions. *Brain and Cognition*, 44, 280–302.
- Jiang, Y., & He, S. (2006). Cortical responses to invisible faces: Dissociating subsystems for facial-information processing. *Current Biology*, 16, 2023–2029.
- Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews. Neuroscience*, 6, 766–774.
- Jones, B. C., DeBruine, L. M., Main, J. C., Little, A. C., Welling, L. L., Feinberg, D. R., et al. (2010). Facial cues of dominance modulate the short-term gaze-cuing effect in human observers. *Proceedings. Biological Sciences/The Royal Society*, 277, 617–624.

Kaminski, J., Riedel, J., Call, J., & Tomasello, M. (2005). Domestic goats, *Capra hircus*, follow gaze direction and use social cues in an object choice task. *Animal Behaviour*, 69, 11–18.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 361, 2109–2128.

Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Ito, K., et al. (1999). The human amygdala plays an important role in gaze monitoring: A PET study. *Brain*, 122, 779–783.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kendrick, K. M., da Costa, A. P., Leigh, A. E., Hinton, M. R., & Peirce, J. W. (2001). Sheep don't forget a face. *Nature*, 414(6860), 165–166.

Kingstone, A., Friesen, C., & Gazzaniga, M. (2000). Reflexive joint attention depends on lateralized cortical connections. *Psychological Science*, 11, 159–166.

Kita, S. (2003). *Pointing: Where language, culture, and cognition meet*. Mahwah, NJ: Erlbaum.

Klein, J. T., Deaner, R. O., & Platt, M. L. (2008). Neural correlates of social target value in macaque parietal cortex. *Current Biology*, 18, 419–424.

Kuhn, G., & Kingstone, A. (2009). Look away! Eyes and arrows engage oculomotor responses automatically. *Attention, Perception & Psychophysics*, 71, 314–327.

Langton, S., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition*, 6, 541–567.

Leavens, D. A., & Hopkins, W. D. (1998). Intentional communication by chimpanzees: A cross-sectional study of the use of referential gestures. *Developmental Psychology*, 34, 813–822.

Leavens, D. A., Hopkins, W. D., & Bard, K. A. (2005). Understanding the point of chimpanzee pointing: Epigenesis and ecological validity. *Current Directions in Psychological Science*, 14(4), 185–189.

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science*, 7, 297–307.

Loretto, M., Schloegl, C., & Bugnyar, T. (2010). Northern bald ibises follow others' gaze into distant space but not behind barriers. *Biology Letters*, 6(1), 14–17.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology, Paris*, 102(1–3), 59–70.

Meltzoff, A. N., & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds* (pp. 335–366). Oxford: Oxford University Press.

Miklósi, A., & Soproni, K. (2006). A comparative analysis of animals' understanding of the human pointing gesture. *Animal Cognition*, 9, 81–93.

Miklósi, Á., Kubinyi, E., Topál, J., Gácsi, M., Virányi, Z., Csányi, V., et al. (2003). A simple reason for a big difference: Wolves do not look back at humans, but dogs do. *Current Biology*, 13, 763–766.

Miklósi, A., Polgardi, R., Topál, J., & Csányi, V. (2000). Intentional behavior in dog-human communication: An experimental analysis of “showing” behaviour in the dog. *Animal Cognition*, 3, 159–166.

Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, 7, 1–9.

Moore, C., & Dunham, P. J. (1995). Joint attention: Its origins and role in development. Hillsdale, NJ: Lawrence Erlbaum.

Moore, T., & Fallah, M. (2004). Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology*, 91, 152–162.

Morris, J. S., Ohman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating “unseen” fear. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 1680–1685.

Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., Parlade, M. V., et al. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, 78, 938–954.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., Rizzolatti, G., et al. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology*, 78, 2226–2230.

Neiworth, J. J., Burman, M. A., Basile, B. M., & Lickteig, M. T. (2002). Use of experimenter-given cues in visual co-orienting and in an object-choice task by a new world monkey species, cotton top tamarins (*Saguinus oedipus*). *Journal of Comparative Psychology*, 116, 3–11.

Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13, 135–143.

Okada, T., Sato, W., Kubota, Y., Usui, K., Inoue, Y., Murai, T., et al. (2008). Involvement of medial temporal structures in reflexive attentional shift by gaze. *Social Cognitive and Affective Neuroscience*, 3(1), 80–88.

Okamoto-Barth, S., Call, J., & Tomasello, M. (2007). Great apes' understanding of other individuals' line of sight. *Psychological Science*, 18, 462–468.

Op De Beeck, H. P., Dicarlo, J. J., Goense, J. B., Grill-Spector, K., Papanastassiou, A., Tanifuji, M., et al. (2008). Fine-scale spatial organization of face and object selectivity in the temporal lobe: Do functional magnetic resonance imaging, optical imaging, and electrophysiology agree? *Journal of Neuroscience*, 28, 11796–11801.

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, 8, 315–324.

Pack, A. A., & Herman, L. M. (2004). Bottlenosed dolphins (*Tursiops truncatus*) comprehend the referent of both static and dynamic human gazing and pointing in an object-choice task. *Journal of Comparative Psychology*, 118, 160–171.

Paukner, A., Anderson, J. R., Fogassi, L., & Ferrari, P. F. (2007). Do facial gestures, visibility or speed of movement influence gaze following responses in pigtail macaques? *Primates*, 48(3), 241–244.

Pelphrey, K. A., Singerman, J., Allison, T., & McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: The influence of context. *Neuropsychologia*, 41, 156–170.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 335, 23–30.

Perrett, D., Rolls, E., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47, 329–342.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., et al. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 223, 293–317.

Pika, S., & Zuberbühler, K. (2008). Social games between bonobos and humans: Evidence for shared intentionality? *American Journal of Primatology*, 70(3), 207–210.

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., et al. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative fMRI study. *Journal of Neurophysiology*, 101, 2581–2600.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400, 233–238.

Platt, M. L., Lau, B., Glimcher, P. W., Hall, W. C., & Moschovakis, A. (2004). Situating the superior colliculus within the gaze control network. In W. C. Hall & A. Moschovakis (Eds.), *The superior colliculus: New approaches for studying sensorimotor integration* (pp. 1–34). Boca Raton: CRC Press.

Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J. V., et al. (2009). Do we really need vision? How blind people “see” the actions of others. *Journal of Neuroscience*, 29, 9719–9724.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews. Neuroscience*, 11, 264–274.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research*, 3, 131–141.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews. Neuroscience*, 2, 661–670.

Rosa, M. G. P., & Tweeddale, R. (2005). Brain maps, great and small: lessons from comparative studies of primate visual cortical organization. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 360(1456), 665–691.

Rosati, A. G., & Hare, B. (2009). Looking past the model species: Diversity in gaze-following skills across primates. *Current Opinion in Neurobiology*, 19, 45–51.

Saalmann, Y. B., & Kastner, S. (2009). Gain control in the visual thalamus during perception and cognition. *Current Opinion in Neurobiology*, 19, 408–414.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16, 235–239.

Schloegl, C., Kotrschal, K., & Bugnyar, T. (2007). Gaze following in common ravens, *Corvus corax*: Ontogeny and habituation. *Animal Behaviour*, 74, 769–778.

Seltzer, B., & Pandya, D. N. (1991). Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. *Journal of Comparative Neurology*, 312, 625–640.

Seltzer, B., & Pandya, D. (2004). Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. *Journal of Comparative Neurology*, 281, 97–113.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology: CB*, 18, 668–671.

Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, 13, 127–134.

Sewards, T. V., & Sewards, M. A. (2002). Innate visual object recognition in vertebrates: Some proposed pathways and mechanisms. *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, 132, 861–891.

Shepherd, S. V. (2010). Following gaze: Gaze-following behavior as a window into social cognition. *Frontiers in Integrative Neuroscience*, 4(5), 1–13.

Shepherd, S. V., & Platt, M. L. (2008). Spontaneous social orienting and gaze following in ringtailed lemurs (*Lemur catta*). *Animal Cognition*, 11, 13–20.

Shepherd, S. V., & Platt, M. L. (2009). Neuroethology of attention in primates. In L. Tommasi, M. A. Peterson, & L. Nadel (Eds.), *Cognitive biology: Evolutionary and developmental perspectives on mind, brain, and behavior (Vienna Series in Theoretical Biology)* (Vol. d, pp. 281–308). Cambridge, MA: MIT Press.

Shepherd, S. V., Deaner, R. O., & Platt, M. L. (2006). Social status gates social attention in monkeys. *Current Biology*, 16, R119–R120.

Shepherd, S. V., Klein, J. T., Deaner, R. O., & Platt, M. L. (2009). Mirroring of attention by neurons in macaque parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 9489–9494.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology*, 20, 649–656.

Schiffer, S. (1972). *Meaning*. Oxford: Clarendon Press.

Sperber, D., & Wilson, D. (1986). *Relevance, communication and cognition*. Cambridge, MA: Harvard University Press.

Tamietto, M., Castelli, L., Vighetti, S., Perozzo, P., Geminiani, G., Weiskrantz, L., et al. (2009). Unseen facial and bodily expressions trigger fast emotional reactions. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 17661–17666.

Tanner, J. E., & Byrne, R. W. (2010). Triadic and collaborative play by gorillas in social games with objects. *Animal Cognition*, 13, 591–607.

Teufel, C., Gutmann, A., Pirow, R., & Fischer, J. (2010). Facial expressions modulate the ontogenetic trajectory of gaze-following among monkeys. *Developmental Science*, 13, 913–922.

Tipples, J. (2002). Eye gaze is not unique: Automatic orienting in response to uninformative arrows. *Psychonomic Bulletin & Review*, 9, 314–318.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (Vol. 16, pp. 103–130). Hillsdale, NJ: Erlbaum.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, 68, 1067–1080.

Tomasello, M., Call, J., & Hare, B. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour*, 55, 1063–1069.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675–691.

Tomasello, M., Hare, B., & Agnetta, B. (1999). Chimpanzees, *Pan troglodytes*, follow gaze direction geometrically. *Animal Behaviour*, 58, 769–777.

Tomasello, M., Hare, B., & Fogleman, T. (2001). The ontogeny of gaze following in chimpanzees, *Pan troglodytes*, and rhesus macaques, *Macaca mulatta*. *Animal Behaviour*, 61, 335–343.

Tootell, R. B. H., Tsao, D., & Vanduffel, W. (2003). Neuroimaging weighs in: humans meet macaques in “primate” visual cortex. *Journal of Neuroscience*, 23, 3981–3989.

Tsao, D. Y., Schweers, N., Moeller, S., & Freiwald, W. A. (2008). Patches of face-selective cortex in the macaque frontal lobe. *Nature Neuroscience*, 11, 877–879.

Udell, M. A., Dorey, N. R., & Wynne, C. D. (2010). What did domestication do to dogs? A new account of dogs’ sensitivity to human actions. *Biological Reviews of the Cambridge Philosophical Society*, 85, 327–345.

Umiltà, M. A., Escola, L., Intskirveli, I., Rochat, M., Caruana, F., Jezzini, A., et al. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 2209–2213.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing. A neurophysiological study. *Neuron*, 31(1), 155–165.

Vecera, S. P., & Rizzo, M. (2004). What are you looking at? Impaired “social attention” following frontal-lobe damage. *Neuropsychologia*, 42, 1657–1665.

Vuilleumier, P. (2002a). Facial expression and selective attention. *Current Opinion in Psychiatry*, 15, 291–300.

Vuilleumier, P. (2002b). Perceived gaze direction in faces and spatial attention: A study in patients with parietal damage and unilateral neglect. *Neuropsychologia*, 40, 1013–1026.

Wang, Y., Brzozowska-Prechtl, A., & Karten, H. J. (2010). Laminar and columnar auditory cortex in avian brain. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 12676–12681.

Zimmermann, F., Zemke, F., Call, J., & Gomez, J. C. (2009). Orangutans (*Pongo pygmaeus*) and bonobos (*Pan paniscus*) point to inform a human about the location of a tool. *Animal Cognition*, 12, 347–358.

Support Publication #12 [SP12]

Alnajjar F., Renawi A.M., Cappuccio M.L., & Mubin O. (2019),
“A low-cost autonomous attention assessment system for robot intervention with autistic
children”, *Proceedings Of The Tenth Ieee Global Engineering Education Conference*
(Educon 2019), 9-11 April 2019, American University in Dubai, Dubai, UAE, 787-792.
<https://doi.org/10.1109/EDUCON.2019.8725132>

Reproduced with permission from IEEE. All rights reserved.

A Low-Cost Autonomous Attention Assessment System for Robot Intervention with Autistic Children

Fady S. Alnajjar
Department of Computer Science and
Software Engineering, UAE University
Al Ain, UAE
fady.alnajjar@uaeu.ac.ae

Abdulrahman Majed Renawi
Department of Computer Science and
Software Engineering, UAE University
Al Ain, UAE
abdulrahman.m@uaeu.ac.ae

Massimiliano Cappuccio
Department of Philosophy,
UAE University
Al Ain, UAE
m.lorenzo@uaeu.ac.ae

Omar Mubain
Western Sydney University
Australia
o.mubain@westernsydney.edu.au

Abstract— Attention is an essential mental process that is important to achieve learning progress. We cannot get better in our academic learning unless we concentrate our attention on the person giving the educational material, such as the teacher or trainer. Children with Autism Spectrum Disorder (ASD) may have attention difficulties that can directly influence their academic skills. In recent years, robot intervention in autism therapy and assessment is becoming a popular research topic due to its role in enhancing children's attention more than a regular human therapist, as well as, the increasing number of autism children compared to the availability of professional therapists. Robot intervention helps in reducing therapy time and makes early therapeutics sessions easier and much promising. Many researches have been conducted to develop robot intervention techniques for ASD children, and some methods have already been used to assess autistic individuals' attention during the robot intervention sessions. Yet, the existing attention assessment methods are either very complex or simple with one measured interaction cue only. This paper presents a practical and low-cost automatic approach to assess autistic individuals' attention during robot intervention; addressing multiple interaction cues. Experimental results show that the proposed attention assessment system could accurately measure the child attention and enhance therapy progress. This automatic attention system can open a new era for utilizing technologies to monitor students' attentions in the class to enhance educational systems.

Keywords—Autism Therapy, Robot Intervention, NAO Robot, Assessment System, Autism Diagnosis

I. INTRODUCTION

Children with autism are usually experiencing neurodevelopmental disorders characterized by attention deficit and learning difficulties [1]. Earning the autistic children attention (i.e., perform eye/face contact, focus on the information they hear while filtering the noises and perform the task assigned to them), in a complex environment is, therefore, an extreme challenge. This challenge can affect their learning and academic skills and their potential for long-term success.

Autism therapy and assessments with robots intervention is a trending research topic, since the statistics show noticeable increasing rates in the number of autistic individuals in UAE compared to professional therapists [2]. There are plenty of researches addressing autism therapy and assessments techniques with the aid of robotic interventions,

yet they are applied only in controlled experimental setups and not intensively focusing on automatic attention level assessment [3,4].

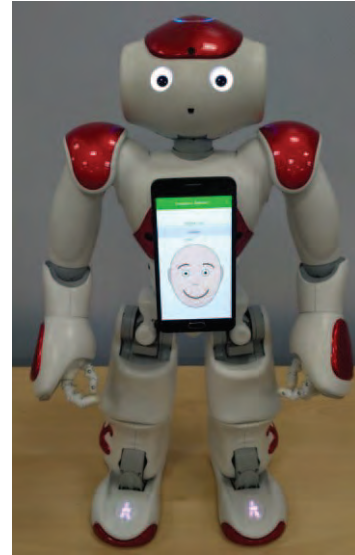


Fig. 1. The NAO robot with our customized chest holder to carry a mobile to enhance the robot's sensory & motor capabilities.

Recent researches on robotics intervention present several techniques to diagnose autistic individuals using robots and accompanying sensors [4,5]. Some of these techniques present a score-like assessment to represent the interaction level between the robot and the autistic individual [6]. These scores have several forms to represent multiple interaction and attention cues. Previous literature informs us that attention definition in the context of robotic autism assessment could be modeled to include various factors such as:

(1) **eye/face attention:** represents how much the individual is paying eye/face contact to the robot. This can be measured using a variety of measurement techniques, ranging from eye-pupil diameter measurement [7] to image processing utilizing the robot's onboard camera [8].

(2) **Joint attention:** represents the ability of the individual to react to robot's motoric actions, like looking at the same object that the robot is pointing to. This can be measured using the same hardware used to measure attention with different

data processing techniques [3] or it can be measured separately using a motion tracking method [9].

(3) **Facial emotions:** detection of the facial reaction of the individual interacting with the robot. Detecting facial features can be used to detect emotions and other aversion and attention cues [10]. A camera could be enough to detect facial features, and a depth camera [11] can be used for better accuracy. The output of facial features detection helps in understanding the emotional state of the individual to relate it to other attention measures.

(4) **Sound/Linguistic response:** the ability to answer the given questions during the interventions within a certain time.

(5) **Imitation response:** the ability to imitate the robot's movements.

Most of the existing assessment techniques in robot intervention studies either use external hardware components not mounted on the robot, which limits portability, or target a single attention cue due to robot's onboard hardware limitations. This paper presents a method that uses upgraded robot onboard/mounted hardware targeting automatic assessment of the above mentioned attention cues.

II. METHODOLOGY

This study introduces a simple and efficient numerical assessment system for robot intervention without the use of external monitoring equipment. The utilized robot for this study is the humanoid robot Nao from Softbank Robotics with our added mobile screen mounted on the robot body, shown in Fig. 1.

A. Technical and hardware configurations

For the robot intervention with autistic children, beside regular Nao robot intervention [6], we have added mobile based hardware with Facial Features Display to enhance the intervention, as seen in Fig. 1. The presented approach does not compromise the mobility of the robot and ensures the applicability in human-robot interaction scenarios. The display is assembled with the aid of a custom chest-mount designed for the NAO robot, shown in Fig. 2, to minimize the changes in the moment of inertia due to the added weight. The added display is a mobile phone that shows emotions running a custom designed mobile application named Emotions Selector. A mobile magnetic holder was customized to fit on the chest-mount to carry the mobile screen. The mobile application was designed using the MIT App Inventor [12], it receives the emotion control data from the computer via socket communication over Wi-Fi. The application is made easy to use without complications other than specifying the computer's IP address to link it with the mobile automatically.



Fig. 2 [a] The customized NAO chest holder front side with the mobile magnetic holder. [b] is the holder back side with the added strap to fasten it.

The used chest holder (chest-mount) is designed in our lab using Fusion 360 CAD Software to be a multipurpose tool for holding any object using up to four screws for rigidity. Velcro straps are used to fasten the holder from the back. It can be used by other researchers to carry a mobile, a camera, a laser sensor or any other item on the chest of NAO robot. Moreover, the USB socket on the head of NAO robot is used to charge the mobile phone. This technique maintains the robot's equilibrium ability, thus robot's built-in features can be used extensively during the diagnosis and therapy processes.

B. Autonomous Assessment system

Since the target of this paper is to particularly present the autonomous assessment system of the robot intervention, rather than the intervention technique itself, the following sections will highlight the assessments factors:

Attention Score:

The mounted mobile's camera is used for facial recognition to produce and accumulate the attention score (i.e., how much the individual is paying eye/face contact to the robot, thus a smiley face indicates maximum attention). Image frames are retrieved online from the camera of the mobile, acting as an IP camera, and processed using Haar classifiers [13] to detect the participants face. The participant's face is detected if his/her face is directed towards the robot's front body. This assumes that the participant is paying attention to the robot. Once the participant's face is detected, the attention score is incremented with a set increment value. As long as the participant's face is not detected, the attention score is decremented by a smaller set decrement value. The increment and decrement values are made variables that can be reset based on future findings from experiments with autistic participants. The attention score is designed to be a comparison factor between several interaction sessions' results for a single participant. Thus, its parameters should be constant for each participant. In this study, parameters are shown in Table I. The used increment and decrement values affecting attention score are shown in the table (experimentally tuned).

TABLE I. USED INCREMENT AND DECREMENT VALUES AFFECTING ATTENTION SCORE.

	Increment Value (points)	Decrement Value (points)
Attention	5	1
Joint Attention	15	1
Sound Response	Eq. (1)	5

Joint Attention Score:

This score part is designed to measure how the participant reacts to robot motoric actions and requests. For instance, if the robot talks about a box on the side and points towards it, the participant is expected to look towards the same box. The participant's head motion detection is simplified to detect the participant's left and right ears. If the participant looks to his/her right side, his/her left ear should appear to the robot. And the same applies for his/her left side and right ear. This detection is made using the same image frames retrieved from the mobile camera. If the detected participant's head direction matches the current robot's head direction, the joint attention score is incremented by a set increment value; else, it is decremented by a set decrement value (Table I). In addition,

the attention score is incremented or decremented by set values in each joint attention instant, since attention score is designed to report on the general attention of the participant.

Sound Response:

The mobile microphone is used as an IP mic and defined as a sound input in the laptop for sound response module. The sound response module calculates the time between the time instant when NAO robot finishes asking a question and the time instant when the participant starts answering. This time difference is reported as sound response time and it is used to calculate an increment in the attention score. Lower sound response time (faster response) generates a bigger increment. The maximum increment due to the lowest response time (ideally zero) can be reset, but it should be constant for the same participant in all sessions for comparison.

This module has some parameters that can be adjusted based on the participant's normal speaking volume and background noise. These parameters are:

- *Threshold-sound-peak*: detected peak sound level after which the module assumes the person has started speaking. Set to 2000Hz in the following experiments.
- *Time-out*: the maximum time the module waits before killing the process and ignoring the sound response. This value is expected to be bigger with participants with speech difficulties. Set to 2.5 seconds in the following experiments.
- *Text-is-done-time-lapsed*: the maximum time the module waits for the robot to reply with a text-is-done, after which the process is killed if robot didn't finish speaking. This value can be set to larger than the time the robot takes to say the longest expected question. Set to 5 seconds in the following experiments.
- *Time-delay*: a delay in the code to compensate for the mic delay; the time between capturing a sound peak at the mic and receiving the same sound peak received at the computer. Set to 0.25 seconds in the following experiments.

The increment in attention score due to sound response (*Incr*) is made as a function of response time (*RespTime*) and maximum score for response time (*MaxRespScore*) as the following Eq1:

$$Incr = MaxRespScore - \frac{MaxRespScore}{Time-out} \times RespTime \quad (1)$$

Based on the above increment equation, the faster the sound response, the more attention score increment is granted. The used value of *MaxRespScore* in the following experiments is 25 points.

Moreover, the IP mic allows the operator to hear what the mobile hears; what the participant and the robot say, without interfering with the sound response module. All of these communications were made over Wi-Fi with different ports.

Emotions Detection:

The proposed assessment system is also able to estimate the participant's emotion from facial features with the aid of NaoQi [8] emotions detection Python module utilizing NAO robot onboard camera. The module returns the probability of

the participant's current emotion among five possible emotions: happy, sad, angry, surprised and neutral. The most probable emotions are logged and graphically represented as color bands, each color represents a specific emotion

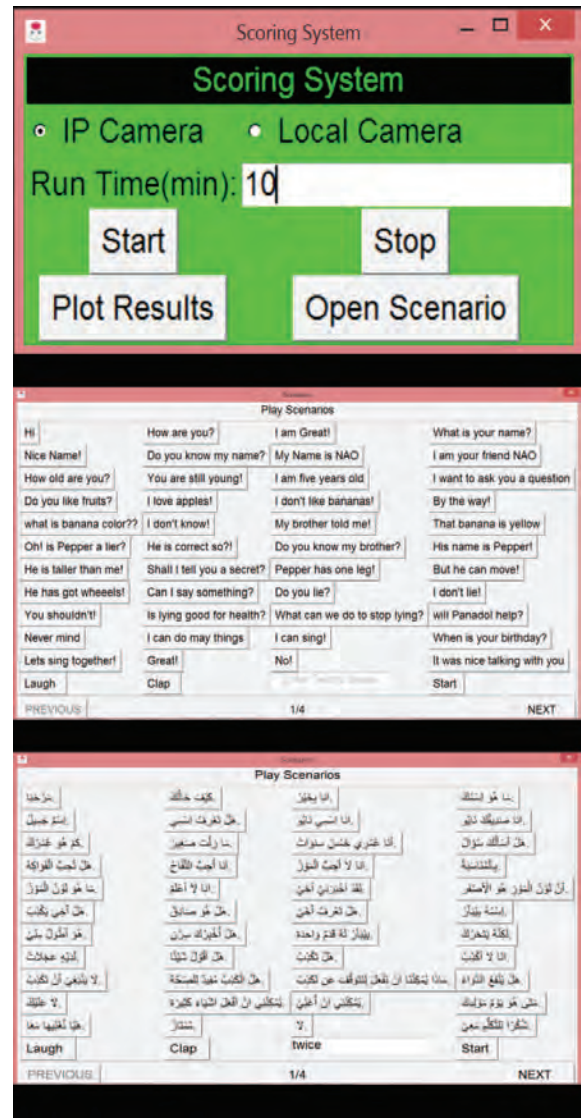


Fig. 3. Assessment System and Scenario GUIs. The first page in the GUI [a] is the assessment system control interface. It supports local and IP cameras. Plot Results will plot the latest session results. The second [b] and third [c] pages are the English and Arabic versions of the desinged scenario, respectively. They can be accessed from Open Scenario button.

C. Experimental Setup:

Our robot intervention scenario is divided into two parts [6], a control part, in which the therapist can control the robot from operator's room, Fig. 4[b] (more details are in [6]). And the other part is the experimental room in which the child interacts with the robot and our system assesses the interaction (Fig. 4[a]).

The control part is made as a graphical interface as shown in Fig. 3 to be easy-to-use by non-technical operators like therapists. The operator should specify the maximum time limit and start the interaction session, having the ability to stop at any time. During the diagnosis process, the operator can control the flow of the predesigned scenario, having the ability

to type a new response if the participant deviates from the preset scenario. The scenario includes dialogues to interact with the participant to grab his/her attention and test his/her interaction capabilities. Sound responses, joint attention requests and robot movements are automatically called in the preset scenario parts.

5 experiments were carried out with healthy participants to test the robustness of the designed control and assessment systems. As stated before, the control station was placed in a room different than the room where participant and the robot are. Participant and robot were seated close to each other as seen in Fig. 4[a]. A side camera was used to record and monitor the experiments remotely.

III. RESULTS

The output assessment scores are accumulated in real-time and shown as a separate diagnostic real-time plot as shown in Fig. 5 to show the operator the level of interaction between the autistic individual and the robot as attention and joint attention scores. The real-time plot reports on the fast responses' values, sound response requests, attention score and joint attention requests' instants. The results are saved at the end of each session as a 'pickle' data file for further processing. The operator can directly plot the recent experiment's results from the scoring system GUI. Fig. 6 and Fig. 7 shows the results obtained from experiments done with two different participants.



[a]



[b]

Fig. 4. Robot Interaction Room [a] and Control Station [b]. Communication is made over WiFi. The operator may not need any technical trainings to use the control station. In the lower figure [b], the right screen is the monitor showing the Robot Interaction Room to the operator, the middle screen is the scenario control GUI and the left screen shows the real-time plot and the robot general motions and behaviors controller.

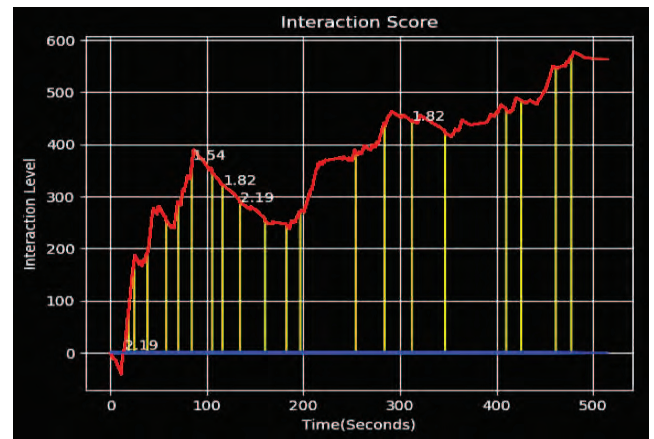


Fig. 5. Real-Time Plot. The red line is the real-time attention score. The blue line is the real-time joint attention score (this experiment had no joint attention requests). Yellow ribbons mean a sound response request; a sound response was expected at this time instant. The numbers in white are the fast sound responses; less than the set time_out value.

Final results after the end of the sessions are presented in three subplots as:

- 1- Attention mixed scores plot: emotions color bands, attention and joint attention scores
- 2- Sound Response Plot: shows the participants' response time versus interaction time in seconds.
- 3- Emotions counter: shows how many times each emotion has been detected.

Assessment system results show the expected interaction level; after 100 seconds, the participant in Fig. 6 was not looking at the robot for a duration of 95 seconds, which was a clear drop in the attention score. Yet, the other participant, Fig. 7, showed a better attention to the robot with smaller drops in attention score.

Second participant, Fig. 7, didn't have any joint attention requests from the robot, while the first participant, Fig. 6, was expected to show some joint attention cues, which is shown by the increase in the joint attention score.

Sound Response results show that both participants were partially responsive at all times of interaction, which is shown in real-time plots of yellow bands without the response time written on top of the yellow band. Not writing the response time means that the response time was more than the set timeout value. This is approved in the sound response plots, where participants had many big values in response times, which means that they were slow in their vocal responses at those instants.

Emotion bands' plots show the most probable emotion at each instant compared to attention and joint attention scores versus interaction time. This helps in matching participants' interaction levels to emotional states. The emotions counter makes it easier to know the dominant emotion compared to other emotional states in the entire interaction session. Moreover, some autistic traits could be tested with the knowledge of participant's emotional state. Both participants show more 'Happy' emotion predictions at the beginning of interactions than 'Sad', and more 'Sad' than 'Happy' at the end of interaction. This shows a decay in the level of excitement as the interaction time increases.

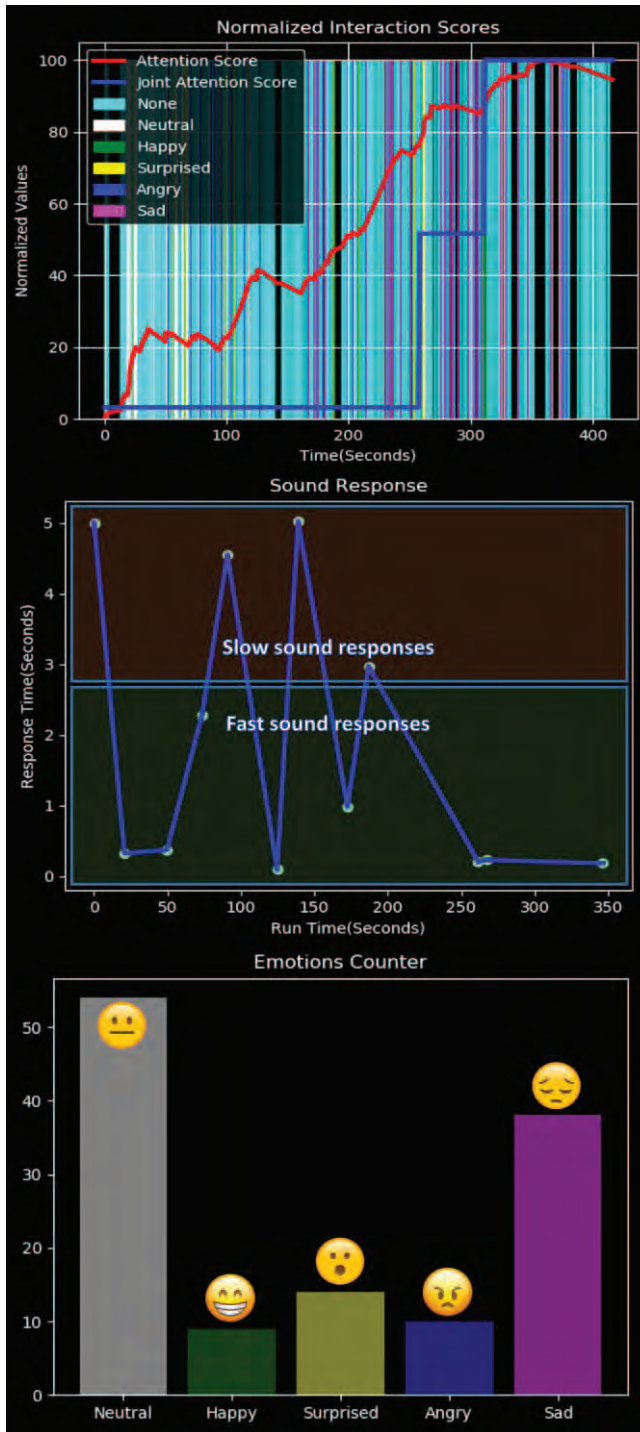


Fig. 6. Results of the experiment from the first participant. Upper plot are the attention scores (red) and joint-attention scores (blue) normalized and plot over emotions color bands. Attention and emotional states can be mapped at each time instant. “None” emotion means that no emotion data could be generated at that instant. Middle plot shows all sound responses recorded. Slow sound responses are more than timeout value and they are not plot in real-time. Fast sound responses are equal to or less than the timeout value. Timeout value can be reset for each participant to match their cases. Several parameters can be adjusted by the operator to customize the scoring system based on participant’s case, this enhances attention cues detection. Lower plot shows participants emotions counter; it is the overall count of emotions detected during the session. These results are comparable for the same participant in different sessions given the same parameters values.

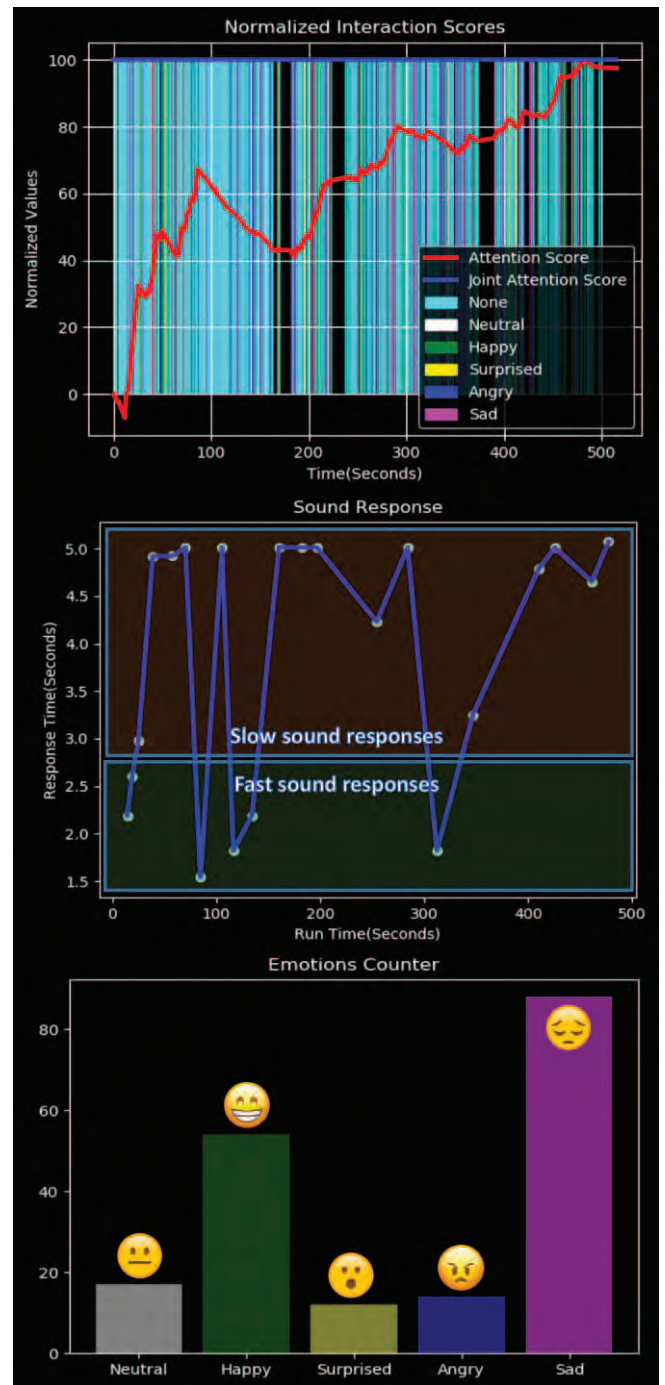


Fig. 7. Results of The Experiment With The Second Participant.

IV. CONCLUSION

A novel low-cost scoring system for robot intervention in autism therapy and diagnosis was designed and tested on two participants. Results show a match between the interaction trend captured by the scoring system and observations. The proposed scoring system uses a simplified hardware and software setup that is easy to use by non-technical autism therapists, as well as parents in domestic environments. This scoring system facilitates using robots as daily assessment tools and therapy assistants in autism schools, hospitals and in domestic use. Our future direction is to apply the attention testing in clinical setups to a cohort of autistic children with attention deficit and compare the results with therapists’

assessment results to test the validity of the proposed system clinically. Having clinically validated results would help develop a permanently available clinical system that can be used in domestic setups as well.

REFERENCES

- [1] J. L. Matson and M. Shoemaker, "Intellectual disability and its relationship to autism spectrum disorders," *Research in developmental disabilities*, vol. 30, no. 6, pp. 1107-1114, 2009.
- [2] V. Eapen, A. A. Mabrouk, T. Zoubeidi, and F. Yunis, "Prevalence of pervasive developmental disorders in preschool children in the UAE," *Journal of Tropical Pediatrics*, vol. 53, no. 3, pp. 202-205, 2007.
- [3] S. M. Anzalone *et al.*, "How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot," *Research in Autism Spectrum Disorders*, vol. 8, no. 7, pp. 814-826, 2014.
- [4] K. S. Lohan, E. Sheppard, G. Little, and G. Rajendran, "Towards improved child robot interaction by understanding eye movements," *IEEE Transactions on Cognitive and Developmental Systems*, 2018.
- [5] J.-J. Cabibihan, H. Javed, M. Aldosari, T. W. Frazier, and H. Elbashir, "Sensing Technologies for Autism Spectrum Disorder Screening and Intervention," *Sensors*, vol. 17, no. 1, p. 46, 2016.
- [6] M. Alahbabi *et al.*, "Avatar based interaction therapy: A potential therapeutic approach for children with Autism," in *Mechatronics and Automation (ICMA), 2017 IEEE International Conference on*, 2017: IEEE, pp. 480-484.
- [7] *Codamotion CXI-800* [Online]. Available: <https://codamotion.com/3d-measurement/>.
- [8] *Aldebaran Nao robot* [Online]. Available: <http://www.aldebaranrobotics.com/>.
- [9] *Kinect for Windows features* [Online]. Available: <http://www.microsoft.com/en-us/kinectforwindows/meetkinect/features.aspx>.
- [10] Y. Feng, Q. Jia, M. Chu, and W. Wei, "Engagement Evaluation for Autism Intervention by Robots Based on Dynamic Bayesian Network and Expert Elicitation," *IEEE Access*, vol. 5, pp. 19494-19504, 2017.
- [11] StereoLabs. zed-ros-wrapper [Online] Available: <https://github.com/stereolabs/zed-ros-wrapper>
- [12] S. C. Pokress and J. J. D. Veiga, "MIT App Inventor: Enabling personal mobile computing," *arXiv preprint arXiv:1310.2830*, 2013.
- [13] P. I. Wilson and J. Fernandez, "Facial feature detection using Haar classifiers," *Journal of Computing Sciences in Colleges*, vol. 21, no. 4, pp. 127-133, 2006.

Support Publication #13 [**SP13**]

Alahbabi M., Almazroei F., Almarzoqi M., Almeheri A., Alkabi M., Nuaimi A.A., Cappuccio
M.L., Alnajjar F. (2017),
“Avatar Based Interaction Therapy: A Potential Therapeutic Approach for Children”
In X IEEE International Conference on Mechatronics and Automation (ICMA), 480–496.

Reproduced with permission from IEEE. All rights reserved.

Avatar Based Interaction Therapy: A Potential Therapeutic Approach for Children with Autism

Moza Alahbabi¹, Fatima Almazroei¹, Mariam Almarzogi¹, Aysha Almeheri¹, Mariam Alkabi¹, A. Al Nuaimi²,
Massimiliano Cappuccio¹, Fady Alnajjar¹,

¹UAE University, UAE

²Systems Department, Veus Technology, Brazil
fady.alnajjar@uaeu.ac.ae

Abstract – Autism is a neurodevelopment disorder characterized by severe deficiency in social interaction and communication skills. In this study, we investigated response levels of 5 healthy children to a robotic-avatar therapeutic system specially designed for children with Autism, and compared it to a pre-programmed robot designed to interact with autistic child and conventional interactive therapy conducted by a therapist. Our preliminary results show that our proposed avatar robot can increase the response and the interaction time of the child, thus increasing the attention span. On the other hand, our system can also automate and facilitate measuring the progress of interaction. The avatar robot can replicate the therapist's presence, mimicking his/her motion, transferring his/her voice, as well as, delivering feedbacks to the therapist with visuals and audio from the child environment.

Index Terms – Autism, Avatar-robots, Therapy

I. INTRODUCTION

Autism is commonly defined as a markedly abnormal or impaired development in social interaction and communication and a significantly restricted repertoire of activities and interests [1]. Diagnosis requires the presence of multiple abnormalities, with at least two being symptoms of impaired social interactions, at least one being a symptom of impaired communication, and at least one being a symptom of a restricted repertoire of activities and interests [2][3]. The etiology of Autism and its neurological bases are still debated and are currently investigated by the experts without consensus [4].

To address the main symptoms of Autism in kids, the most common therapeutic approaches have been focusing social skills training [5]. These approaches often involve practicing specific social and physical skills during regular meetings of groups of autistic children, using hypothetical or dramatized scenarios designed to solicit social learning [6]:

A. Direct personal therapy

Direct therapy refers to face-to-face communication between the patient and a trained therapist to increase the patient's social interaction skills in a measurable way [7].

Behavioral skills training (BST) is a kind of such direct training. It includes:

1) *Instruction*: the therapist uses verbal or written explanation of what a skill is and how is it performed

2) *Modeling*: the therapist demonstrates to the patient how the skill is done.

3) *Rehearsal*: the therapist allows the patient to attempt the skill independently

4) *Feedback*: the therapist explains what the learner performed accurately and where improvement can be achieved [8].

Along these lines, some researchers have utilized language paradigm training [9] while others suggested more specialized behavioral interventions [10].

Although these direct therapies have been shown to have various levels of success, they all raise the same inherent problems [11]. First, as the autistic child lacks in social and communication skills, he might develop anxiety for interaction with strangers, thus rendering him less responsive to treatment. The treatment in some cases can be felt as a chore by the child due to him being forced to interact with a stranger [12].

Another important problem is the subjective nature of the evaluation procedure carried out by therapists at the end or during the session, to quantify the relevant features of the child's behavior. A lack of objectively defined measurement protocols might create unreliable records of therapy that can affect the diagnosis of the autistic symptoms and invalidate future longitudinal analyses. [13].

B. Virtualized Therapy

In an effort to lower the distress generated by direct interaction with the therapist, and making the therapeutic sessions more appealing, various virtualization methods have been employed [14].

Recent research has been done to answer the question of whether autism patients show a favorable response to robot-based therapy. It has been shown that humanoid robots are superior to other non-humanoid virtual characters in this respect [15].

Regarding whether children generally accept a robotic therapist, it has been shown that acceptance generally is positive. Child acceptance can be fairly measurable and easy to detect: an increase in cardiac rate, for instance, has been used as an indicator of whether the child is alarmed by the presence of the robot or whether he is relaxed, thus providing an easy tool to identify the efficacy of the treatment [16].

As a measure of child response and performance, imitation speed has been measured in multiple studies and

used as an indicator of the progress of therapy. These imitation procedures can include basic operations such as grasping a ball or waving to a partner [17].

Although robotic-based treatment has shown some promising advantages, such as quantitative progress measurements and evasion of forcing the child to stay with a stranger, they suffered from some intrinsic shortcomings [18]. Most of the robotic and virtualized therapies are preprogrammed. This fact can make the learning process repetitive, potentially boring, and lacking in versatility, especially when temporal factors must be taken into account [19]. Although there have been a small number of systems with some adaptive capabilities, they have been shown to be quite limited and working only in very controlled environments [20]. Another major shortcoming is the inadequacy of quantitative behavioral analysis tools due to the absence of a professional therapist in the process of development of such robots.

We here are targeting these two important shortages. In this study, we investigated a novel robotic avatar-based therapeutic system specially designed for autism patients: the system increases the immersion of both the patient and therapist and facilitates evaluating the progress of the therapy. Our avatar robot will replace the therapist in the room. The therapist would wear portable motion sensors giving him a partial control of the robot motion, and receive a visual and auditory stream from the robot. The therapist would also be able to talk to the patient through the robot, by using a robotic voice. We believe that this work can fill the gap between therapists and robot programmers, in addition to being more fun for the patient, in comparison to pre-programmed alternatives.

II. METHOD

A. Participants

For the preliminary stage of this study, instead of Autistic patients, five healthy male children were recruited (average age 5+/-2). This preliminary study was conducted to prove the concept and to solve technical problems, delay in avatar motion, scoring system, etc. The recruited individuals have no history of neuro-disorders. An important factor considered when selecting our five participants was the level of anxiety for strangers, as reported by their parents.

B. Materials

1) *Motion captures sensors*: A Perception Neuron set of 32 sensors was used for real-time capturing of the volunteer's body movements [21], Fig.1. The sensors provide accurate measurements for the angles of the head and arms pose and velocity. The data was treated with Axis Neuron software and forwarded to Unity3D using the provided plugin. A mini-framework was built in Unity3D using C# to allow forwarding the captured movements to 3D virtual avatars and to the robot.

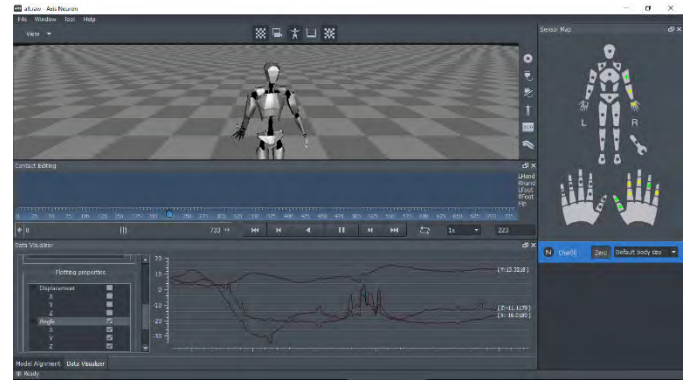


Fig. 1 Real-time motion captures by the Perception Neuron Sensors.

The created framework initializes a User Datagram Protocol (UDP) server that would broadcast the motion and audio data to any recipient accessible within the computer's network. The motion data was JSON encoded, with floating point angle measurements for the head, shoulders, elbows, and individual fingers. Audio data was broadcasted as 16-bit wav audio treated in real-time to resemble a robotic voice. The visual stream was broadcasted in the same way using the robot internal camera.

2) *The NAO robot*: for this study, a Nao Robot by Aldebaran was used [22]. NAO robot is flexible and provides comparatively smooth mimicking of the received movements. In addition to that, NAO comes with some security features such as automatic body balancing and opening the hand when pressure is felt, so as not to hurt the user in case of holding his hand (Fig.2).

The movement calculation module was implemented using a modular design that allows adapting it according to the Degrees of Freedom (DoFs) of the robot, on the basis of the required balance between accuracy of motion (large DoFs) and economy of implementation (small DoFs). We defined the DoFs of the robot based on the engagement of the user. Having lower DoF generates simpler movements that can be easily mimicked by the user. As the user progresses in the engagement /therapy, the number of DoFs can be increased by enabling disabled actuators or by adding more motion complexity using a virtual robot on the screen.

A Bluetooth heart-rate measuring band was worn by the participant and remotely connected to the therapist's computer. The heart was continuously monitored by the therapist to determine the level of anxiety of the participants when dealing with the robot, thus to act whenever an intervention was required.

Microsoft Kinect was used to analyze the motion of the participants to be matched to the movements done by the therapist. The patient required no special headphones and microphones to be attached so as to reduce any stress on the participating children.



Fig. 2 Nao Robot used for real-time interaction with Children

C. Procedure

Participants were told that they would be playing a mimicking game with a Nao robot and answering some questions in English and Arabic. Mimicking includes reproducing the robotic movements and repeating what the robot says. Before the experiment, Kinect motion capturing software was calibrated to make sure that the poses from the Perception Neuron sensors and Kinect are matching.

The experiment started by the robot singing a song for children and doing some dances while the therapist monitored the heart rate of the patient. We preprogrammed a library on a robot to complete this stage. After making sure that the participant is emotionally stable, the therapist would talk to the child through the robot asking him/her to repeat some movements and gestures performed by the robot through the avatar of the therapist. These gestures were specifically selected to teach the child how to interact with other people and to make sure that (s)he is more responsive to other people (for example, bow, shake hands, say thank you, etc.)

Participant motion was captured by Kinect and compared in real-time to the adjusted movements of the therapist. A score was calculated for each group of movements and stored in a database within the participant's record, Fig.3. These data are used in subsequent visits to analyze the continuous progress of the user and to advise the therapist on when and how to the training has to be adjusted.

Conversations were also used in tandem with movement mimicking in order to closely simulate real-life situations. The recorded audio was also stored in the user's record for further analysis. Each session lasted for 15~20 minutes.

Due to the mechanical limitations of our robot, we instructed the therapist to move slowly and avoid sudden and large motions while using the motion-capture sensors.

An important factor related to autism therapy, considered also in this study, was joint attention. Impaired development of joint attention is a well-known feature of autism [23]. Researchers have shown that autism patients produce poorer joint attention responses when exposed to complex sensorial stimuli, depending on the severity of the case [24]. Thus in this study, we could control the level of the joint movement complexity of the robot. To do that, the readings of the motion capture sensors were first mapped to the degrees of freedom (DoFs) of the robot by inverse kinematics on individual limbs. Constrained optimization was then used to convert the joint

angles and velocity. Optimization techniques were used to transform the computed motion to the capabilities of the robot by specifying an objective function and a set of constraints that preserve the characteristics of the original motion. The robot would then track the trajectories of the transformed data using a position and velocity tracking system, Fig.4.

In the case of Nao robot, the following inverse-kinematics equations were used [25]:

$$T' = (A_{Base}^0)^{-1} T(A_4^{End})^{-1} (R_z(\frac{\pi}{2}))^{-1}$$

$$T'' = (T')^{-1}$$

$$\theta_3 = \begin{cases} \text{asin}(\frac{T''_{(3,4)}}{l_1}) \\ \pi - \text{asin}(\frac{T''_{(3,4)}}{l_1}) \end{cases}$$

$$\theta_4 = \mp \text{acos}(\frac{l_3 T''_{(2,4)} + l_2 T''_{(1,4)} \cos \theta_3}{l_2^2 + l_1^2 \cos^2 \theta_3})$$

$$T''' = T' (T_2^3)^{-1} (T_3^4)^{-1}$$

$$\theta_2 = \text{atan}(\frac{T'''_{(2,1)}}{T'''_{(2,2)}}) - \frac{\pi}{2}$$

$$\theta_1 = \text{atan}(\frac{T'''_{(1,3)}}{T'''_{(3,3)}})$$

Where T is a transformation matrix, A is a parametric translation matrix, R is a rotation transformation matrix, θ is the relative angle at the joint, and 1~4 represents the coordinates of the joint-end.

Since each of the robot's arms has 6-DoFs, we started by disabling some of them to generate simpler motion and enabling them gradually as the user gets more accustomed to the current training regime.

Evaluation of the mimicking accuracy was calculated by RULA (Rapid Upper Limb Assessment) readings (Fig.3) [26].

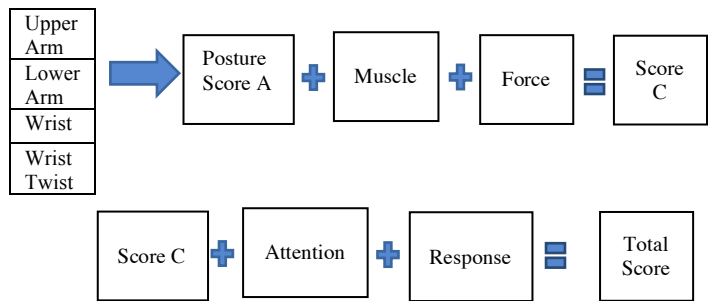


Fig. 3. Scoring system. Evaluation of the mimicking accuracy was calculated by RULA (Rapid Upper Limb Assessment) readings. *Posture Score A*: was calculated from the deviation of the angles of the limbs from the target posture. *Muscle Score*: was given for static poses held for more than 10 seconds or for repetitive actions done more than 4 times per minute. *Force*: a score given to the weight being held (if the posture includes grasping and carrying of objects). We have also integrated in the RULA's score the *Attention* and the *Response* factors. *Attention*: look at whether the subject is looking at the robot or not (based on face recognition algorithm). *Response*: is a tool to examine whether the patient is talking/answering the avatar-robot or not (sound recognition algorithm).

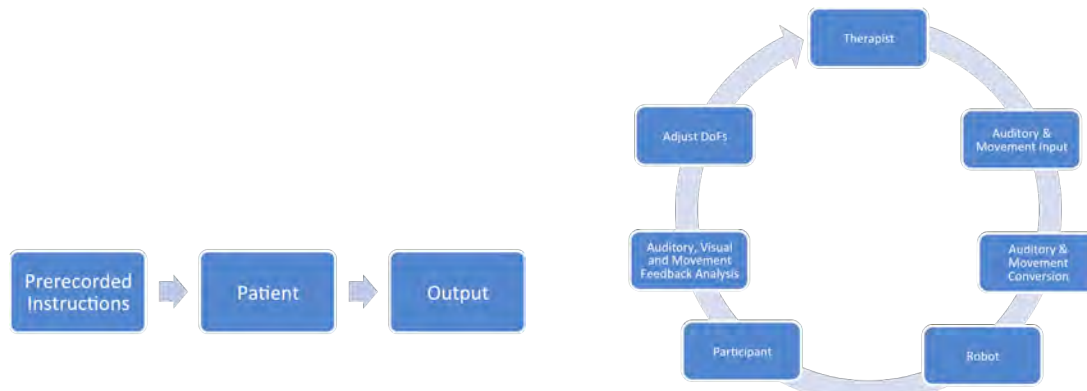


Fig. 4 Standard Robot versus Interactive Avatars Training. In standard preprogrammed robot, a one-way communication was mainly the case. In Avatar training, a two ways interaction and adaptation were offered.

In addition to motion score, two other scores were calculated: the Attention score, and the Response score. *Attention* score was calculated by giving one positive point for each second the child was looking at the robot's face whenever the robot was asking the child a question. *Response* points were given for the delay time between the moment in which the robot asks a question and moment in which the child started answering the questions. The child is given 10 points for the first 10 seconds after the robot/therapist stops talking, and one point is subtracted for each second of delay. The Total score is calculated by adding the three scores, giving them almost similar weights in the evaluation process.

III. RESULTS AND DISCUSSION

A. Human-Robot Interaction level time

In this preliminary study, our main target was to examine how much various types of interaction bodies may earn the participants' attention and whether it's able to win more interaction time with them. A pre-programmed robot, a therapist, and the same therapist in form of robot avatar were the interaction bodies. Each child was introduced to these interaction bodies for 20 minutes. A scoring system was computed and evaluated by an evaluator every 4 minutes of interaction. The evaluator was invisible to the participant, i.e., watching the child from behind the camera, and was monitoring a real-time data.

Figure 5 shows the abstract output from the evaluator. From the figure, at the beginning of the first stages, the avatar and the pre-programmed robots had higher scores than the therapist. Gradually, the avatar's scores increased and a slight decrease on the pre-programmed robot was observed. We believe that the decrease in the score of the preprogrammed robot was due to the robot's repetitive motions and the gradual delay in the child response, children started feeling bored and were ignoring robots. We did not notice such loss during the interaction with the avatar, because the therapist who was controlling the avatar started to adapt his interaction based on the situation and the child's response.

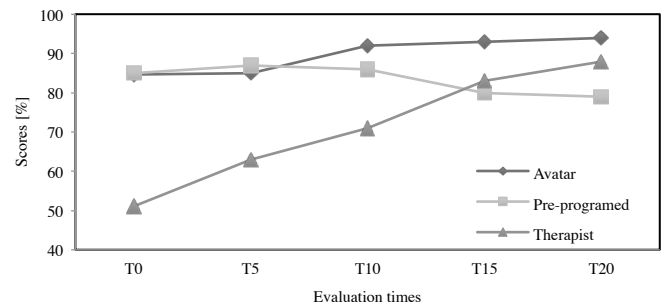


Fig. 5 Averaged scores computed for each interaction subject with the 5 children. "100%" means full response and interaction, while "0%" means neither response nor interaction. T0 reflects the score of the first moment the child meet the interaction body; T5 reflects the score of after 5 minutes of interactions, etc.

Although the therapist, as an interaction body, was having bad response from the child at the initial period of the meeting, he overcame the pre-programmed robot by around 14 minutes of interaction (after the therapist could win the trust of the child). Yet, the avatar robot kept its lead until the end of the session. The parents who attended the sessions showed their admiration to their child interaction with the robot avatar.

IV. CONCLUSIONS

In this preliminary study, we investigated response levels of five healthy children to a robotic avatar therapeutic system specially designed for children with Autism, and compared it to a pre-programmed robot and conventional therapy conducted by a therapist. Each child was introduced to each session for 20 minutes. A scoring system was designed to evaluate the interaction and response of the child for each session. The results show that our proposed avatar robot system was superior in comparison to the other two interaction bodies, and could hold better the attention and the response of the child.

Although our robot avatar was better than pre-programmed robot in engaging in interaction with normal children, but we cannot predict yet how differently autistic

children would react. Our future plan, therefore, is to introduce the avatar system to autistic children and compare it with conventional therapy/robot systems. We are also planning extra tuning of the scoring system to be more robust and accurate.

REFERENCES

- [1] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders: DSM-5," Washington, DC: American Psychiatric Association, 2013.
- [2] Stephen E. Brock, "The Identification of Autism Spectrum Disorders: A Primer for the School Psychologist," Sacramento California State University, 2006.
- [3] Maglione, M.A., Gans, D., Das, L., Timbie, J., & Kasari, C., "Nonmedical interventions for children with ASD: Recommended guidelines and further research needs. *Pediatrics*", pp. 130, 2012
- [4] Baron-Cohen, S., "The hyper-systemizing, assortative mating theory of autism. *Progress in NeuroPsychopharmacology and Biological Psychiatry*," pp. 865–872, 2006
- [5] Klin A, Lang J, Cicchetti DV, Volkmar FR., "Brief report: Interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: Results of DSM-IV autism field trial," *Journal of Autism and Developmental Disorders*, pp. 163–167, 2000.
- [6] Rao, P. A., Beidel, B. C., & Murray, M. J., "Social skills interventions for children with Aspergers syndrome or highfunctioning autism: A review and recommendations", *Journal of Autism and Developmental Disorders*, pp. 353–361, 2008.
- [7] Hanna, E., & Meltzoff, A.N., "Peer imitation by toddlers in laboratory, home, and day-care contexts: Implications for social learning and memory," *Developmental Psychology*, pp. 701–710, 1993.
- [8] Ward Horner, J., & Sturmey, P., "Component analysis of behavior skills training in functional analysis," *Behavioral Interventions*, pp. 75-92, 2012.
- [9] Gianoumis, S., Seiverling, L., & Sturmey, P., "The effects of behavior skills training on correct teacher implementation of Natural Language Paradigm teaching skills and child behavior," *Behavioral Interventions*, pp. 57–74, 2012.
- [10] Hogan, A., Knez, N., & Kahng, S., "Evaluating the use of behavioral skills training to improve school staffs' implementation of behavior intervention plans," *Journal of Behavioral Education*, pp. 242-254, (2015).
- [11] Joshua J. Diehl, Lauren M. Schmittl, Michael Villano, Charles R. Crowell, "The clinical use of robots for individuals with Autism Spectrum Disorders: A critical review," pp. 249-262, 2012.
- [12] Ricks, DJ.; Colton, MB. "Trends and considerations in robot-assisted autism therapy," *IEEE International Conference on Robotics and Automation (IRCA)*, pp. 4354-4359, 2010.
- [13] Ricks, DJ., Colton, MB., "Trends and considerations in robot-assisted autism therapy," *IEEE International Conference on Robotics and Automation (IRCA)*, pp. 4354-4359, 2010.
- [14] Pierno, A.C., Mari, M., Lusher, D., & Castiello, U., "Robotic movement elicits visuomotor priming in children with autism. *Neuropsychologia*," 46, pp. 448–454, 2008.
- [15] Pioggia G, Iglizzi R, Sica ML, Ferro M, Muratori F, Ahluwalia A, De Rossi D. "Exploring emotional and imitational android-based interactions in autistic spectrum disorders," *Journal of CyberTherapy and Rehabilitation*, pp. 49–62, 2008.
- [16] Bird G, Leighton J, Press C, Heyes C. "Intact automatic imitation of human and robot actions in autism spectrum disorders," *Proceedings: Biological Sciences*, pp. 3027–3031, 2007.
- [17] Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L., "Autism Diagnostic Observation Schedule: ADOS-2", 2012, Los Angeles, CA: Western Psychological Services.
- [18] Klin A, Lin DJ, Gorrindo P, Ramsay G, Jones W.. "Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*," 459, pp. 257–261 2009.
- [19] Cabibihan, J.J., Javed, H., Ang M. Jr., & Aljunied, S.M., "Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism," *International Journal of Social Robotics*, 5, pp. 593–618, 2013.
- [20] Rao PA, Beidel DC, Murray MJ., "Social skills interventions for children with Asperger's syndrome or high-functioning autism: A review and recommendations. *Journal of Autism and Developmental Disorders*," 38, pp. 353–361, 2008.
- [21] Perception Neuron by Noitom, website: neuronmocap.com
- [22] Aldebaran Nao Robot, website: www.aldebaranrobotics.com
- [23] Emily A. Jones, Edward G. Carr, "Joint Attention in Children With Autism," *Theory and Intervention*, Vol 19, Issue 1, 2004.
- [24] Kalas A., "Joint attention responses of children with autism spectrum disorder to simple versus complex music," *J Music Ther.*, pp. 430-52, 2012 Winter. 49(4).
- [25] Nikos K, Emmanouil O., Michail G., "Complete analytical inverse kinematics for NAO", *13th International Conference on Autonomous Robot Systems (Robotica)*, 2013
- [26] McAtamney, L., Corlett, E.N., "RULA: A survey method for the investigation of work-related upper limb disorders," *Appl. Ergon.* , pp. 91–99, 1993, 24.

Support Publication #14 [SP14]

Cappuccio M.L. (2015),
“Inference or Familiarity? The Embodied Roots of Social Cognition”
Synthesis Philosophica, 29/2, (April 2015),
Croatian Philosophical Society, Zagreb, pp. 253-272.

Reproduced with permission from the publisher.
All rights reserved.

Massimiliano L. Cappuccio

UAE University, Cognitive Science Program, Room 0018A, E5 Building, Al Ain, United Arab Emirates
M.Lorenzo@uaeu.ac.ae

Inference or Familiarity?

The Embodied Roots of Social Cognition*

Abstract

I consider two distinct deflationary theories in social cognition that aim to explain action understanding without demanding meta-representational or mindreading processes: the first one is the 'teleological stance hypothesis' (TSH), claiming that we infer the intended goal of a certain observed action based on the mere perception of its effects and of its situational constraints; I decided to dub the second one 'the embodied familiarity hypothesis' (EFH) to comprise all the theories claiming that we recognize the intended goal of a certain action based on the perceptual or motoric expertise developed within the sensorimotor contingencies associated to that action's context. TSH's main requirement is that the observer could ascribe efficiency, and therefore rationality, to the observed agent's movement, while EFH's main requirement is that the observer were somehow exposed to the perceptual or motoric details of the observed agent's action. I argue that EFH describes a more primitive and fundamental form of action understanding, i.e. one that is necessarily presupposed by TSH: in fact, while recognizing efficiency is neither a necessary nor a sufficient condition for detecting goal-relatedness, some kind of perceptual or motoric familiarity with the details of the observed action's context is always necessary for any ascription of efficiency, and therefore of rationality, to the observed agent. I conclude that, while TSH might certainly be effective in describing certain rational forms of action understanding, it implicitly requires EFH to be true, as its inferential system would be groundless without an assumed familiar background of embodied expertise.

Keywords

social cognition, teleological stance, mirror neurons, rationality, intentional stance, direct matching hypothesis

Introduction: two deflationary approaches to social cognition

Social cognition is the field of interdisciplinary research that studies how we comprehend the others. It is interested, among various things, in how we can understand¹ the goal of an observed action in an effortless, automatic and

*

This paper is in debt with the philosophical conversations I had with Corrado Sinigaglia at the University of Milan and with Dan Hutto during my period of research visit at the University of Hertfordshire. I am very thankful to Chiara Brozzo for reading an early version of this paper and for commenting on it. The responsibility of any possible mistake is just mine.

¹

After cautious meditation, I will intentionally ignore the warning, presented and reasonably argued for by various scholars (e.g., Hutto 2013), against any similar use of the concept of 'understanding': even if it were true that in the philosophical literature this word predominantly indicates intellectual comprehension, the expression 'understanding an action' actually has a much broader and less



largely reliable manner, even if we know nothing of what the agent is thinking and very little of her personal background. As opposed to many models of ‘mindreading’ that have been advanced during the last decades (Baron-Cohen 1994, Goldman 2006), and that demand meta-cognitive representations of the agent’s mind, two deflationist theories seem particularly promising today, as they attempt to provide a cognitively parsimonious account of the processes underlying action understanding: both of them assume that in our direct interaction with the others we do not need any intellectual representation of their mental states (i.e. propositional contents such as beliefs, intentions, and desires), because the key information necessary to recognize the action goal is perceptually accessible through observation of the agent’s bodily movements (Gallese et al. 2009). Both human infants (Astington & Gopnik 1991; Gopnik, 1993; Wellman 1991) and non-human primates (Call and Tomasello, 2008) manifestly understand the others’ actions goals without having propositional or mentalistic representations of the observed agents’ states at their disposal, and possibly without being capable of distinguishing their own mental states from the observed agents’ (Gallagher and Povinelli 2012). The two aforementioned deflationist accounts attempt to offer different explanations of how low-profile action understanding is available even to subjects that cannot rely on language or “mindreading”.

The first account claims that in order to make sense of the other’s action we need to infer its goal based on the expected effects of her action and what we see of her current contextual circumstances: this view assumes that the observer’s inference is reliable insofar as she realizes that the observed agent knows both what she wants to achieve and what the most efficient means to achieve it are. Gergely and Csibra (2003) refer this to as the ‘teleological stance hypothesis’, or TSH. The second account claims that our capability to understand the other’s action goal relies fundamentally on our own embodied familiarity with the observed action, including the sensorimotor contingencies in which it is situated. I will generically refer to this view, which assumes that we can make sense of the intention underlying another’s movements on the basis of our direct acquaintance with them, as ‘embodied familiarity hypothesis’ (EFH). In this paper I will intentionally characterize this view in a generic way, because – if broadly construed – EFH encompasses various embodied approaches to social cognition, and they do not always entirely agree on the kind of perceptual and motoric requirements that are necessary to recognize movements as goal-oriented (Gallagher 2008). The approaches encompassed by EFH include for example the interactionist and enactivist approaches and the motor cognition hypothesis. On the one side of this approach, interaction theory (Gallagher 2001; Gallagher & Hutto 2008; Gallagher & Zahavi 2008; De Jaegher et al. 2010; Froese and Gallagher 2012) claims that we do not infer goals beyond the others’ actions, but we directly perceive them in the others’ movements, and that this perceptual function develops with our capability to proficiently adjust our conduct to another’s behavior (the goal of the action is perceptually associated to its immediate effects, and these effects gain a perspicuous perceptual meaning during live experience by reason of the history of the unprincipled embodied engagement between interacting agents). On the other side of EFH, the ‘motor cognition hypothesis’ (Gallese et al. 2009) asserts that it is our capability to perform a bodily action to achieve a certain motor goal that allows us to recognize the movements aiming at the same motor goal, when we see them performed by someone else. Note that both the ‘direct-matching hypothesis’ (Rizzolatti et al. 2001; Rizzolatti and Sinigaglia 2010) and the ‘embodied simulation theory’ (Gallese 2005, 2010)

offer more specific versions of the motor cognition hypothesis, which we will not consider here. Therefore, according to EFH, goal recognition could either rely on the familiarity that the agent has developed in responding to the observed movements, or on the analogy between the movements observed and the movements that the agent would have produced in similar contexts: the observer maps the observed action in terms of goal-oriented movements that are either complementary or analogous to the ones observed, but in either case – unlike for TSH – it is the involvement of his embodied dispositions (and hence his direct sensorimotor experience) that allows him to make sense of the observed action's goal, not calculative skills or detached capabilities of reasoning and intellectual judgment.

Both TSH and EFH reflect the essence of traditional philosophical doctrines, which today receive extensive empirical support from experimental cognitive psychology: the first of them has been corroborated by a series of psychological experiments on the infants' attention based on a looking-time paradigm (Gergely et al. 1995; Gergely and Csibra 2003); the second received credit from various studies in the field of motor cognition (e.g., the discovery of the so called "mirror neurons", Di Pellegrino et al. 1992), which suggested how embodied skills play a defining role in action understanding.

As both models seem conceptually consistent and grounded in experimental evidence, each of them could convincingly account for at least some instances of action understanding. I do not mean to deny that they could be mutually incompatible, as different cognitive strategies of action understanding probably co-exist to allow us make sense of goals in terms of either practical reasons to act, as argued by TSH, or sensorimotor opportunities, as maintained by EFH. However, even if these models were compatible or complementary as alleged, we would still have to determine if one of these two theories could satisfactorily explain the most fundamental and essential processes that underpin action understanding, and if it could do it without assuming the other theory. The question that the two theories have to address, but that possibly not both of them can answer in an equally satisfactory way, is: what is goal attribution in its most primitive and essential form, and what are the minimal cognitive preconditions to have action recognition? While TSH maintains that the capability to infer practical reasons is the most fundamental requirement to characterize an action goal, under the presupposition that the capability to perform a certain action is not required to infer its goal, EFH argues that perceptual habits and motor skills are more fundamental, as it assumes that the capability to rationalize the abstract intention behind another's action is not indispensable to recognize the immediate purpose that shapes her movements. If we agree that rational inference and sensorimotor skills are two distinct and independent sets of capabilities that (at least in principle) do not presuppose one another, it seems that we have to choose between their respective underlying principles: do the roots of action understanding arise from a universal principle of rational inference or rather from embodied dispositions?

To address this question, this paper will focus on the critical discussion of one of the theoretical preconditions of TSH: the necessary attribution of rationality to the observed agent. This critical discussion will not prove this precondition wrong *a priori*, but will argue that the attribution of rationality is not a

intellectualist use in common language. Considering how difficult it is to do without such expression even when referring to non-human animals and non-linguistic infants, I have the

feeling that artificially adopting an extensionally equivalent synonym just to avoid demanding cognitive pre-requisites would be just a pointless lexical exercise.

primitive capability of the infant's mind, as it rests on other contingent, and more primitive, preconditions – of an embodied (not inferential) kind. I believe that this conclusion does not rule out TSH, but suggests the primacy of EFH over TSH, as it leads to the idea that bodily situated enactive capabilities are from the beginning involved in action recognition and goal understanding, and play an active and constitutive role in them.²

1. The teleological stance hypothesis

This paper will deny neither that TSH builds on solid experimental data nor that it holds a certain degree of explanatory power. Various bodies of evidence suggest that 12-month old infants understand the goal of an agent's actions employing minimal perceptual cues about the action and its context (Csibra et al. 2003; Gergely et al. 2005). This early capability of understanding precedes folk-psychological knowledge because the infant does not master any such mind reading capability yet, and the observed movements are not sufficient to attribute propositional contents to an agent's state of mind. So how does early goal attribution work at such age? What we know is that generic visual features such as self-propelled movement (Premack 1990), autonomous non-rigid transformations of the moving body's surface (Gibson, Owsley, & Johnston 1978), irregular path of movement (Mandler 1992), and motion caused by spatially and temporally distant events are typically sufficient to indicate an internal source of energy, hence suggesting animacy and goal-directedness. However, these features are insufficient to determine the particular goal behind one's action (Anscombe 1957). TSH aspires to complete the picture, postulating that specific goals can only be attributed to those actions that: 1. display an equifinal structure through time (see Heider 1958: in spite of varying environmental conditions, the patterns of action chosen by the agent consistently result in the same end-state; this suggests, via induction, that the agent's behavior is consistent through time); 2. elicit an attribution of rationality to an agent (see Dennett 1987: the agent's behavior suggests that she has selected the most efficient means to reach a desired goal-state).

Because of these two pre-requisites (i.e. reliability of inductive reasoning and attribution of rationality to the agent), TSH is comparable to the 'intentional stance hypothesis' (or ISH, at first presented by Dennett 1987 and then endorsed by Gergely et al. 1995), and represents its deflationary development (Gergely and Csibra 2003). ISH asserts that goal attribution takes place as a folk-psychological process that presupposes the systemic integration of three kinds of mental representations: the agent's beliefs (e.g., "Paul thinks that there is a banana in the fridge"), desires ("Paul wants a banana"), intentions ("Paul means to grab a banana"). Combined with sufficient perceptual cues ("Paul is reaching for the fridge"), the knowledge of any two of these representations is sufficient to infer the third one, as the known representations work as the premises of an inductive reasoning that has the unknown representation as its conclusion. Since, in this picture, the intention is sufficient to characterize the action's proximal goal, knowing the agent's desires and beliefs can do the job of goal attribution. Beliefs, intentions, and desires are treated as representations because they convey declarative contents, hence the inference drawn from them is mentalistic in character: ISH assumes that the observer must manipulate meta-representations (possibly propositional in format) of the agent's mind in order to guess her reasons and her future behavior. While it is debated whether or not ISH could account for goal attribution in

pre-linguistic children and non-human animals (Gergely et al 1995; Gergely and Csibra 2003; Wellman 2002), due to the limited meta-representational capacities of these agents, TSH explicitly claims to be able to explain action understanding in absence of such capacities, as it only requires manipulation of perceptual cues, not meta-representations of the agent's mental states. TSH assumes that, in order to build up such systematic inferences, the observer only needs to consider – instead of the desires [A], the intentions [B], and the beliefs [C] – the physical goal-state at the end of the action movement [A], the observable bodily movements chosen to achieve it [B], and the situational constraints that objectively bound the choice of the movements [C], respectively (Gergely and Csibra 2003). Because all of these elements are perceptually recognizable, and are prerequisites of every embodied action, by simply registering how the agent moves (e.g., in a straight line or a curve), and how the physical surroundings constrain or direct its movements (through obstacles or preferential paths), even an observer unfamiliar with that kind of action can determine to what goal-state the agent's movement is directed. The assumption behind this model of goal attribution via triangulation of the action's end-state is that the agent's movements must display the mark of a rational deliberation.

Note that, in contrast with more intellectually demanding notions of rational action (e.g., the theory of agency as “planning”, as in Bratman 2007), the reasons invoked by TSH are (meant to be) of a minimalist kind: in fact, they simply acknowledge that the means used to achieve a certain goal state were not casual; the agent had practical reasons to pursue that goal with those means, reasons immediately related to reaching that particular goal in an efficient way. Exactly like the teleological stance, also the intentional stance assumes that the observer attributes a capability of deliberation to the agent based on a *principle of rational action*, that is the conjecture of the agent's systematic predisposition to achieve her goal with the most economical means available to her in the given circumstances. In the paradigmatic case of spatial navigation, this means reaching a certain destination through the shortest path. Therefore, like ISH claims that we infer the intention of an action through an inductive reasoning based on the knowledge of the underlying beliefs and desire, TSH assumes that we infer the final direction of a certain movement based on the observation of the movements themselves in relation to the specificities of the environment; therefore, the main difference is that while ISH characterizes goals in terms of intentions (which convey propositional contents), the TSH characterizes them as predictable physical end-states of a movement (which are perceptible states of affairs).

The proponents of TSH have preponderantly focused, with their research, on stimulus properties that are tied only to abstract patterns of the most generic kind of motion, i.e. spatial navigation (Csibra 2003; Csibra and Gergely 1999; Csibra et al. 1999; Csibra et al. 2003; Gergely 2002; Gergely and Csibra 1994, 1997, 2003; Gergely et. 1995). In fact, the data supporting TSH have been collected through the measurement of the looking time of 12-month old babies habituated to computer-animated goal-oriented events performed by non-bio-

logical agents (geometrical drawings in motion, such as animated bouncing spheres), according to a violation of expectations paradigm (I refer in particular to Csibra et al. 2003; Gergely and Csibra 2003; Gergely et al. 2005). For example, in one of the conditions a small circle (i.e. the *agent*) approached and then stopped near a large circle (the *end-state*) after “jumping over” (the *means act*) a solid obstacle that separates them (*situational constraint*). TSH claims that, in similar scenarios involving the same elements, the child can infer that the goal of the small circle’s movement is again to approach the large circle and then stop near it by integrating the available visual cues according to some rational principles. In fact, when the obstacle was removed, the infant’s reactions were consistent with the rational principle of parsimony of movement: after removing the situational constraint, the infants had to watch two test displays in which the small circle performed either the same jumping movement as before or a perceptually novel straight-line movement; looking time was significantly longer in the first condition, as the child was surprised to see that the circle “decided” to “jump” even if there was no obstacle that could justify a curved trajectory. According to the experimenters, we can legitimately conclude that the infant expects the small circle to “behave rationally”, if we keep in mind that the infant: (1) recognized the action as goal-directed (aiming at the same ending state as the previous condition); (2) inferred the most efficient trajectory (curved, in presence of an obstacle, or straight, without any obstacle) under the new situational constraints; (3) expected the agent to use this trajectory to reach its goal. Alternatively, any two other goal-oriented events (in which the means act is fixed but either the end-state or the situational constraint vary across different conditions) can provide complementary evidence of the integration of these three types of visual cues by means of inference: in fact, the surprise of the infant indicates that he has inferred the goal of the action when it was non-evident, if the changing trajectories of the agent’s movement seem not otherwise explainable in the given environmental contingencies; or that he has inferred the presence of hidden obstacles, if the current trajectory of the agent seems deviate from its original pattern in ways that would not be otherwise compatible to the expected end-state. In all of these cases, the children always look longer to the conditions in which the agent’s behavior does not appear as efficient as expected, as they evidently are surprised that the agent is not executing the most rational movements: this proves that the infants have already appropriately inferred the missing element of the triad, and have understood how they were meant to be systematically integrated. Indeed, on the basis of TSH, any inference drawn on the efficiency of the agent “immediately translates in a judgment on her rationality” (Csibra 2003), hence allowing an inductive reasoning on the goal-oriented nature of the action and the missing element of the triad.

While the methodology used by these experiments is solid, and the present discussion does not aim to question their results, both the extension that is given to the meaning of such results and the assumed way of interpreting their preconditions are somehow conceptually puzzling: the problem is not only that this approach has a limited capability to account for other empirical data (as I will argue later, for referring example to Rochat et al. 2008); it is also that this approach does not capture (and, to some extent, is not consistent with) the inherently embodied dimension of the phenomenology of goal-oriented action in real life. In fact, according to TSH, when infants take the teleological stance to interpret actions as means to goals, they are supposed to evaluate the “relative efficiency of means” by applying the principle of rational action, and generate systematic inferences to identify relevant aspects of the situation

that justify the action as an efficient means even when these aspects are not directly visible to them; however, while this hypothesis is in principle sound, the empirical actualities make it factually improbable that the recognition of efficiency is sufficient or even just necessary to attribute a goal to an action. In particular, two assumptions taken for granted by TSH deserve to be critically examined: the first assumption is that non-mentalistic goal attribution is universally based on the calculus of the optimal means (so that efficiency is a necessary condition of goal attribution); the second is that the teleological stance is possible without any other, more fundamental, forms of goal attribution, i.e. without any preconditions or precursors (so that efficiency is a sufficient condition of goal attribution).

There are reasons to question the first assumption: after all, some extremely inefficient behavior leaves no doubts about the goal that motivates them (e.g., humans attempting to grasp a fly with a pair of chopsticks, or dogs running in circle trying to catch their own tails), and actually one could hypothesize that the evident clumsiness and redundancy of the movements deliberately produced by the agents in these circumstances (without any external coercion) is exactly what betrays their goal-oriented intention. Were the optimality of the chosen means the only criterion to determine the action goal, we definitively would not be able to make any such inference. As for the second assumption, its weakness is suggested both by a deeper analysis of the requirements for efficiency, and by a closer consideration of the empirical preconditions that make biological movements more or less efficient. Doubts can be raised as soon as one considers the following two problems: first of all, on what basis do we decide that a movement is efficient? And how efficient are movements supposed to be (or seem) then, in order to be perceived as *the most* efficient? And, second, assuming that we have sufficient criteria and appropriate standards to measure such efficiency, is it true that the movements' efficiency is enough to infer their rationality? In what follows, I will argue that the real role played by efficiency can only be understood through a deep appreciation of the bodily details and bio-mechanical constraints of action execution.

2. Inferring efficient navigation vs. the Frame Problem

According to TSH, an observer is capable to tell the most economical movements to achieve a certain action goal even without any previous performative familiarity with that kind of action. TSH supposes this is possible thanks to a general form of mental means/ends calculus that applies both to the trajectories of translational movements and manipulative actions, regardless of whether they are executed by geometrical shapes or human beings: it is assumed that, in either case, parsimony – i.e. riddance of unnecessary or ineffective movements – is the mark of efficiency, and hence of rationality.

However (as noted by Sinigaglia 2008), efficiency evaluation is undoubtedly enormously simplified if we consider only tasks of spatial navigation, like in the experimental conditions appropriately arranged by the experimenters, as the only parameter considered is the length of the path accomplished to reach the endpoint. We cannot fail to notice that this simplification is already the result of a selection deliberately operated to highlight only certain components of the rationality embedded in our actions by excluding the others; moreover, even reducing the kinematic complexity and morphologic variability of the task at hand does not necessarily make it easier to evaluate the efficiency of the observed action. For example, the translational movements done by

the whole body do not always follow a straight line: this is because of the physiology of the animal, and its behavioral and ecological specificity (e.g., rabbits' and flies' zigzag trajectories). The reasons behind the anti-economicity and apparent randomness of these movements (e.g., reduce the chances to be caught by a predator) can only be appreciated in relation to the specific evolutionary background to which the organism is adapted. Such background reminds us why moving from a point to another in an ideal Euclidean space could hardly be a distinct goal in and of itself for the animal, but rather a goal embedded in the ecological preconditions that characterize its intrinsic coupling with a specific world-environment.

Now, if the efficiency of the movements is correlated with trajectories that are likely to be more complex than a straight line, as is the case, because it depends on the bio-mechanical and behavioral specificity of the organism rather than Euclidean norms, then how could we judge some movements as more or less efficient in achieving their goal? The assessment of the efficiency of an action cannot possibly be dissociated from the consideration of its specific bio-mechanical and ecological constraints in so far they are not just contingent accidents or unnecessary complications, but constitutive motoric components that allow the realization of the action itself: the motoric specifications of the action execution (e.g., the degrees of freedom or the elasticity of the tendons in a rabbit leg) are not obstacles or extrinsic circumstances that would mark the distance between the actual action and an ideal standard; they are the medium itself in which the very possibility to efficaciously reach an embodied/situated motor goal is concretely inscribed.

That is why, outside of abstracted geometrical scenarios, like those used in the experiments of Csibra and colleagues, the distance travelled is just one of the many relevant factors that we might need to consider when we judge the efficiency of a certain action. In many cases, the shortest path is neither the fastest nor the most economical, for example in terms of energy expenditure: Rosenbaum (2008) shows that a fixed path length can be perceived as more or less demanding depending on whether or not the subject is walking in it to reach a certain object (and whether or not the object to be reached is conveniently located in relation to the subject's handedness). The psychological perception of the energetic cost of a movement is not an objective measure, but one that depends on the particular situation of the subject and her goal; however, even if we relied on a more objective calculation of the energy expenditure necessary to complete a certain translational movement, the most energetically economical movement would not necessarily be the most convenient one. If I need to reach my office at the first floor of the Philosophy Department building, I might have to choose: *taking the stairs* (fast, but long and energy-consuming); calling, waiting for, and then *riding the old elevator* (energy-preserving, but slow and long); or even *climbing the external wall* and enter from the window (short, but slow and energy-consuming). The choice largely depends on my contingent motivations (priorities) and embodied dispositions. All of these priorities and dispositions must somehow play a role in shaping a truly rational decision in so far they determine the whole range of the agent's actual possibilities to reach her goal in ways that could be deemed efficient. Note that the agent does not need to explicitly consider all these factors when she produces her movements to achieve her goals; it is sufficient for her to rely on her embodied expertise, i.e. her unaware skillful predisposition to interpret certain environmental circumstances as viable opportunities for action and cope with them in an effective (I am implicitly building here on the theory of absorbed coping, Dreyfus 2002). But how could anyone – as a mere observer

– assess which of the observed agent’s priorities and dispositions could best characterize her choice as a rational one, in a given occurrent scenario?

In my example, the evaluation of the best means act to reach the office cannot leave out of consideration a combination of independent variables such as whether I am late, I have a bad health condition, or I am not keen on meeting a certain colleague in the building. The observers who witness the efforts of an agent to achieve a certain goal can of course decide to focus only on some of these variables, prioritizing them in the description of the scenario that we are using, in the attempt to judge whether the agent is effectively behaving in a way that maximizes the benefits of its decisions based on an ad hoc multi-dimensional matrix. But, is it not true that also this decision to focus only on a limited number of elements is itself a simplification made out of a selection of the complex background preconditions that might underpin the agent’s choice? After all, when I have to decide how to reach my office, there is no *a priori* reason for which, say, the distance to my office would be a parameter objectively more important than the amount of tiredness I am experiencing in my legs. According to Dreyfus (1992), any selection of some parameters over the others is arbitrary if it is not embedded in the concrete familiarity with the transcendent complexity of a real-life scenario, which is of course very difficult to approximate by means of a third-person simulation. Only an agent that is practically familiar with that scenario could tell what decisional parameters are really relevant in those circumstances, and one could doubt that there is such a thing as a rational understanding of what could objectively and universally count as relevant without such familiarity. A totally disembodied and alien system, even if intelligent and generically enabled to take rational decisions, could hardly decide what decisions are rational without any familiarity with the context, regardless of how computationally powerful that system is. The programmers of GPS navigation apps know that path length, travel duration, and fatigue are just some of the key factors that inform our simplest decisions on “how to reach X”. But these factors prove immensely complex when we attempt to simulate the bottomless background of massively interconnected preconditions that underlie the smallest rational decision in real life circumstances, including any choice of the best route to drive back home, which might require to consider an indefinite number of contingencies such as: the gasoline level in our tank, the road conditions, the cost of tolls, the weather, the amount of traffic, the aesthetic value of a scenic route, the proximity of a certain preferred restaurant or shop, the level of restlessness of the kids in the back seat, etc. The set of the relevant contingencies that might affect our simplest decisions can neither be fully represented nor reduced to a smaller set of elements, if we want to preserve the sensitivity to the real context of our intelligent processes. This is the well-known philosophical version of the “frame problem” of artificial intelligence (Dreyfus 1992; Wheeler 2008): the persistent difficulty to analytically spell out the determining factors behind any rational procedure, due to the impossibility to exhaustively list all (and only) the context-sensitive variables that would apply *a priori* to a certain procedure-based decision.

As the frame problem threatens any system that derives its intelligence from representation, it is not just a theoretical impasse for the cognitivist dream to build thinking machines based on internal models of the world and rules of thumb; it is also a serious difficulty for the rationalistic approaches to social cognition, insofar as they presuppose an attribution of rationality to the agent’s decisional processes according to an abstractly universal standard of objective economy and parsimony. The problem with this kind of intellec-

tualist approach to action understanding is twofold: first, how can we judge that an observed agent's action is rational, if we cannot choose the appropriate parameters to judge whether it is economic/efficient or not? And, second, how can we choose the parameters to assess the efficiency of an action, if we do not know the horizon of real-life contingencies that motivates its goal yet? Note that in order to *infer* – as postulated by TSH – whether an action is efficient or not it is not sufficient to have an intuitive grasp of, or some kind of pre-reflective holistic sensitivity to, the motivational background of that action; on the contrary, in order to make a legitimate inference on the other's rational decision it is indispensable to hold a fully explicit model of the premises of the inference – i.e. an explicit representation of the preconditions and motivational factors of that decision. In any other case, if we could realize someone's goal without disposing of any such representation, then it is probably not an inference that allowed us to draw such conclusion. The main problem of any inferential theory of basic action understanding – such as TSH – is that, ultimately, the background preconditions of our concrete rational decisions exceed our possibility of explicit representation and enumeration: this is not a problem for the embodied theories of action understanding that are not based on inferences – such as EFH – because such basic understanding does not require representations at all, as it only relies on a perceptual sensitivity to contextual opportunities for actions and a fluid responsiveness to the way fluctuating environmental conditions can affect us. The concreteness of real-life contexts is a problem for the inferential forms of action understanding because there are always too many parameters that would be in principle relevant to determining whether an action is actually efficient or not, and because the same action could well be efficient according to certain representations of the world and simultaneously inefficient according to others.

If the competing representations of the parameters of efficiency, with their overwhelming complexity, are the problem of basic action understanding, then there could only be two solutions: the first is that the agent's actions actually conform to certain standardized representations of his goals and priorities, and that the observer were instructed in this representation, being informed in advance of the real motivational background of the action. This way the observer could explicitly tell whether the action she is observing is economical or not. But this would not really help explain how goal recognition occurs in the first place, as it assumes exactly what TSH is supposed to explain: the attribution of a teleological valence to the observed action by means of an inference. Assuming that the agent *were told* by any of his world-representations what the goal of the agent is (as opposed to *inferring such goal by himself*), defeats the purpose of TSH as an explanatory psychological model of action understanding.

Alternatively, the observer should be familiar with the material non-representational preconditions and specific embodied constraints that shape the agent's action possibilities: in that case, it would not be indispensable to hold a representation of her goal in order to recognize whether her action is efficient or not; in order to judge whether the other is carrying out movements that are efficient or not it would be both necessary and sufficient, for the observer, to rely on his own practical competence, that is his embodied capability to perform those very movements first hand. But note that, if it is the case that goal understanding is based on such embodied familiarity with the sensorimotor circumstances of action execution, then the movements' goal would not need to be *inferred* anymore, as it is not by means of an inference that the goal is intrinsically associated to fine-grained patterns of characteristic coordinated

movements. In this scenario, the goal of the action would not be inferred at all, as the form of understanding brought about by the required embodied familiarity is completely different from an inference: while an inference is an explicit reasoning that derives logical conclusions from known premises, the recognition of the actual goal based on embodied familiarity is purely based on the original associative connection, in the perceptual experience, of the movements and their motivating intention (Butterfill and Sinigaglia 2014).

Therefore, the evaluation of the efficiency of the movements, according to the perspective of embodied familiarity, is not only necessary and sufficient to recognize the movements' goal, but also concretely possible in the absence of an explicit representation of the reasoning parameters that contextually define whether the action is economical to achieve a certain goal: being capable to perform that action first-hand necessarily implies being familiar with both the implementational details of its execution in typical real-life circumstances and the goal that motivates the appropriate coordination of the movements, in a way that the two of them are intrinsically correlated and mutually presupposing one another. The knowledge of the parameters that make an action efficient would be an operative one, i.e. a pragmatic and habitual capability, not a set of information to be stored and computed in a declarative format. It would be based on the performative familiarity that the observer has matured with respect to that action and to its typical purpose, not on an intellectual calculus of means to be compared with an amodal representation of an abstract goal.

3. Inferring the goal of hand-actions vs. neurocomputational complexity

So far we have been considering the efficiency of translational movements only but, if we consider goal-oriented hand actions, then rational inference seems an even weaker option to explain goal attribution. Proponents of TSH mention Woodward and Sommerville (2000) as evidence that one-year-olds expect a hand to grab a target object with the most direct means action available, and that the target object is attributed as a goal of the action only if the hand acted efficiently to obtain the object. Furthermore, Gergely et al. (2002) claims that evaluating the rationality of actions is not restricted to passive observational contexts, because infants modulate their imitative behavior according to the justifiability of the goal-directed actions performed by a model.

This empirical datum is interesting per se, but does not necessarily back an inferential view on social cognition. In fact, similar experiments suggest that it is harder for an observer to recognize the goal of the agent's action, if the former has no expertise in performing the latter's movements. In accordance with TSH, Rochat et al. (2008) demonstrates that monkeys look longer at grasping actions whose trajectories are less economical than those typically followed in the same context with the same target; but this study also shows that monkeys' appreciation of the means-ends adequacy depended on their sensitivity to the goal relatedness of the observed movements, because observation of incongruent hand trajectories in actions that lack actual goals (when the target object is absent) do not evoke any attentional enhancement, even if the action in question looks very similar to grasping. Additionally, no attentional enhancement was observed when the monkey observed actions that, while visually similar to grasping, were actually motivated by goals that

the monkey had never attempted to achieve before, such as lifting an object with the thumb. This suggests that the recognition of economical/efficient behaviors can support an inference on the nature of the action's goal only when a certain familiarity with the goal of the action already exists.

Therefore, either we conclude that the attribution of rationality is not indispensable to the mechanism of goal recognition, or that this mechanism involves attribution of rationality only insofar as the efficiency of the observed action is measured through the observer's acquaintance with the movements observed. This acquaintance would then be the most fundamental premise of goal attribution. While admittedly Rochat (2008) does not allow specifying whether the required acquaintance must be merely visual or also performative in nature, it provides a serious reason to doubt that – in either case – a general system to calculate the movements' efficiency could work without relying on a repertoire of hand actions mastered by the agent. If we consider transitive actions with effectors (e.g., interaction with objects), the most efficacious movements (i.e. the movements that are well structured to perform a certain transformative function) are not necessarily the most bio-mechanically parsimonious: due to the composition and functional organization of muscles and joints, reaching and manipulative gestures (e.g., hand grasping, pulling, etc.) are usually not structured according to a simple geometry, i.e. they are not designed to reduce unnecessary component movements so that, for example, grasping between two fingers always requires the simultaneous coordination of five. However, importantly, the higher complexity of the morphological description of these movements does not imply a higher complexity of the cognitive information required for producing/controlling them. At least in some cases the opposite could actually be true, as the functional organization of the frontal and pre-frontal cortex suggests that motor areas map structured action goals and not isolated movements (Rizzolatti et al. 1988, Gentilucci and Rizzolatti 1990). In facts, movements intentionally produced in isolation without a motor goal are less frequent and not automatized, and that is why the voluntary independent movement of one single finger of the hand is neuro-computationally more demanding than a rich pattern of organized movements of the whole hand/arm/torso in a typical goal-oriented motor chain such as reaching-for-grasping-for-bringing-to-the-mouth. The modes of the cortical initialization and on-line control of these movement patterns reflect how, in real life interactions with the environment, the fine-grained coordination of the rich details of the movement (morphology, kinematics, velocity, etc.) are closely associated to the goal of the action itself. So closely that recognizing the former can be sufficient to immediately understand the latter, even just from the very initial movements of the hand, as complex patterns of functionally structured movements carry the distinctive signatures of either immediately transformative or social goals.

However, the close association between component movements and action goal does not mean that there is a one-to-one mapping of movements into goals, as very similar movements can be meant to achieve entirely different goals, and the same goal can be achieved by very dissimilar, or even entirely opposite, movements: as an example of the first case, consider that recognizing the type of pre-shaping of a hand that is moving to grasp a knife can be sufficient to guess whether the intention of the agent is to place the knife away or use it to stab someone (Becchio et al. 2008); as an example of the second, consider how different movements of a hand, a mouth, or a plier can be performed to grasp the same object; and that normal or reverse pliers require perfectly opposite movements of the hand in order to accomplish the grasp; and

that nonetheless the goal-specific areas of the premotor cortex recruited to perform such morphologically dissimilar, but functionally and teleologically equivalent, movements are exactly the same (Umiltà et al. 2008). The motoric strategies and the effectors that can be used to achieve the same goal are too many, and with details that are possibly too fine-grained to allow an general system to evaluate *a priori* what movements are the most efficient ones, without previous acquaintance with them: the best way to reduce this complexity in the task of visual recognition is to be capable to perform first-hand those movements, re-using for purposes of recognition and categorization the motoric information that is primarily required to perform certain goal-oriented actions (Prevete et al. 2010).

That is why the motoric familiarity with an observed action, even if not necessarily an *a priori* precondition of goal recognition, is *de facto* playing a crucial role in the usual understanding of the others' action goals in real life scenarios: firsthand capability of execution (i.e. motoric familiarity), facilitates the visual assessment of an observed action's efficiency by enormously restricting the number of variables potentially involved (Metta et al. 2006). This has to do with the fact that the notion of efficiency is intrinsically embodied, because it depends on the bio-mechanical and material specifications of the medium, and also because it depends on the contextual way in which an embodied agent and its environment correspond to one another during their reciprocal interactions, continuously disclosing or obliterating the opportunities of their mutual modification: an action can be deemed more or less efficient only in relation to the range of motoric opportunities that the agent has contextually available, with their situational and embodied constraints, not the totality of the action possibilities that would be available to the agent just in principle or that are not available at all. For example if I could fly with a jetpack, or extend my legs five meters, then – in comparison – riding the elevator, taking the stairs, or climbing the wall would all appear inefficient means to get to my office at the first floor of the Philosophy Department building.

But motoric familiarity is not only indispensable to exclude an infinite number of actions that could possibly be efficient, if they were actually available in the current context; it is also important to distinguish between truly rational actions and randomly or incidentally generated movements. For example, all heavy objects fall towards the ground following a straight line, hence realizing a movement that consistently appears equifinal and economical in its structure, but we would never attribute purposefulness or animacy to this kind of natural movement. Were the observer unaware of the different movements potentially available to the agent, he would probably find it hard to tell whether she is moving rationally or just randomly; moreover, were the agent clearly forced by physical, irremovable, and insurmountable constraints to act in one specific way, and in that way only, we would hardly be inclined to attribute to him his action's goal, as the execution of that action would not appear relatable to a rational decision and hence not motivated by a goal. Why? This is because the distinction between natural and voluntary movements is not generated by the child's naïve theory of physics (Gergely et al. 2005, p. 173, building on Gopnik and Meltzoff 1997), but deeply ingrained in the phenomenology of our bodily experience. The concrete mastering of this distinction in real-life circumstances is realized at the sub-personal level by a combination of feed-forward and feedback systems that automatically compare motor commands with the perceptual feedback that follows action execution: since we are familiar with the experience of our body unintentionally falling, our system automatically recognizes that it is not an action generated by a goal-

oriented motor program. When the same familiar pattern of movement engendered by gravity is recognized in another agent's body, our system realizes that there is neither agency nor purpose in it.

It is reasonable to suppose that, if we intuitively "know" (in the sense of mastering an operational know-how) what motor opportunities are realistic and actually available to us or to the agent we are observing, this is possible not because we were first mentally inspecting and then excluding an infinite number of improbable or impossible movements, but because certain habitual responses are pre-reflectively embedded in our motor system and immediately triggered by the environmental solicitations, automatically predisposing our conduct, adjusting our evaluative and practical attitudes toward the contingencies. This adjustment is on-line, based on real time perceptual stimulation, and relies on stored motor predispositions that prove suitable for the task at hand according to the subject's experience. This is exactly one of the ideas that characterize EFH, and in particular one that is defining for the direct-matching hypothesis in social cognition. Note that motor skills would be a requirement for action understanding regardless that the observer is able or not to attribute mental contents to the agent, as the recognition of constraints of this kind is more likely to be based on a mere perceptual registration of the physical context than on explicit knowledge of the other's intentions.

Conclusions: Inferring efficiency vs. perceiving efficaciousness

The experiments performed by Csibra and colleagues do not capture the minimal preconditions of action-understanding because they look for the wrong kind of minimalism: in fact, by focusing on generic visual stimuli (i.e. stimuli that are deemed to be generally significant for any disembodied agent, regardless of their contextual situation and background) they methodologically exclude the real phenomenology of the biological agents, overlooking the importance of their embodied dimension. The stimuli chosen by the experimenters are minimally specified precisely with the purpose to highlight an abstractly general structure of goal perception: they are intentionally vague, if compared to the specific richness of the bio-mechanical details that characterize functional movements of actual biological agents, because these bio-mechanical details are deemed redundant and accidental. However, I believe I have raised a sufficient number of doubts against the conjecture that the alleged generality of the inferential mechanism postulated by TSH could ever be deducted from the genericity of the visual stimulus that trigger attention in the infants, pointing out that different action goals imply entirely different criteria of efficiency, and typically the relevance of these criteria is correlated to a complex background of embodied constraints that concretely determine how the agent's opportunities of perception and motor execution are coupled with the complexity of real-life contexts. Consequently, vague visual action cues cannot suggest a general typology of goal and are not even sufficient to indicate agency, but (at best) can suggest an underspecified and vague sense of parsimony in generalized (not necessarily agential) movements, such as translations and rotations in an abstract metrical space.

To sum up, the first problem of TSH is that general de-contextualized visual cues are not sufficient to infer the general presence of rationality in the observed agent. Its second problem is that, even if some general kind of rationality could correctly be attributed to an observed agent by means of such cues, visual acquaintance with the movements she performs might be neither

a sufficient nor a necessary condition to understand if and how her actions are efficient in pursuing their goals, therefore it would not be enough to guess their specific goals. Guessing the reason behind the movements of some professional football players might be extremely difficult to infer, if the observer has never been exposed before to any experience of team sports, while understanding the player's movements might be a lot easier if he is more familiar with playing such sports. The capability to evaluate the efficiency of this kind of actions is clearly more correlated to performative expertise (a requirement has nothing to do with rationality, in accordance with EFH) than to disembodied rational powers of mental calculus (as TSH implicitly suggests). And this is why recognizing the intention behind a complex motoric routine has to do more with the embodied skills of the athletes than with capabilities of abstract calculus like those cultivated by mathematicians and accountants for their jobs.

I should add that the parsimony of the observed movements, besides not being a sufficient condition for goal understanding, is neither a necessary condition for it: it is possible to attribute goals to one agent's action even when her chosen conduct appears strikingly inefficient and antieconomical, and efficient movements in turn do not always suggest agency or even animacy. On the contrary, simply acknowledging the agent's complex patterns of coordinated movements, based on the observer's familiarity with that action, seems both necessary and sufficient to understand the agent's action motor goal, whether this is pursued with an economical strategy or not. Therefore, even if it is true that goal attribution results from recognition of *appropriate* means, goal recognition and calculus of the most *parsimonious* means can well be achieved independently, without implicating one another.

Recognizing the appropriate and harmonious organization of the component movements of an action, rather than their parsimony, seems key to goal understanding. TSH, by reducing goals' understanding to a calculus of the most parsimonious means, loses sight of the fundamental distinction (Gallese et al. 2009) between *efficiency of the movements* and *efficaciousness of an action*: the former is reaching a certain target with the least topological transformations, the latter is fulfilling the expectation to produce certain functional modifications of the agent-environment relationship. TSH assumes the indiscernibility of these two different terms, as its main claim – that attribution of rationality is based on efficiency of movement – is true only for actions of which it is possible to equate efficiency and efficaciousness. Such coincidence only occurs in a narrow set of situations in which two conditions hold: first, the target (the destination of the movement) and the goal (practical scope) of the action can be conflated (this works well only for navigation tasks, i.e. actions whose goal specifically is to reach a certain distal point); second, metric relationships between points of the Euclidean space are the only variables that matter for the objective assessment of the most preferable (which are, coincidentally, the most parsimonious) movements (this only happens in those scenarios in which all the other variables, such as the tiredness of the agent, his willingness to avoid to be seen, etc. are irrelevant). These conditions typically do not hold for actions aiming at fulfilling the practical project of an embodied agent in real life circumstances, i.e. in all those scenarios in which space is not just a neutral topological parameter but the medium of situated engagement with the world.

In fact, this is the primitive spatial dimension of our original embodied phenomenology, one that is more fundamental than the intellectually sophisticated

experience of measuring and comparing objective distances in an abstract Euclidean space. The primacy of the practical spatial know-how over a detached form of spatial cognition is not only backed by traditional philosophical ideas whose legacy is influential in embodied cognition (e.g., Poincaré's relationist view on space /1908/, Heidegger's pragmatist doctrine of the "ready-to-hand" /1927/, and Merleau-Ponty's motor intentionality and spatiality of situation /1945/). It is also supported by scientific models of adaptive decision-making and bounded rationality (Berthoz 2003; Gigerentzer 2008): they confirm that embodied agents reduce the computational complexity of spatial tasks, such as catching a ball on the fly, not approaching the problem in an analytic and detached way, but engaging in sensorimotor interaction with the world. This means that the details of their embodied situation not only constrains the range of their rational choices of the agents, but actively scaffold the development of their intelligent ways to negotiate the environment.

All moving animals, including human cognizers, evolved brains to cope with situated spatial tasks by exploiting a constitutive coupling between their sensorimotor skills and the opportunities of action offered by the environment, and they did this long before developing capabilities of objective quantitative analysis and neutral comparison of metric relations. The evolutionary and developmental background of action understanding is deeply rooted in the sensitivity to spatial contingencies whose meaning is practical, ego-centered, and intrinsically mapped into opportunities of bodily intervention, as opposed to context-independent, universal, and decoupled from bodily habits. Different experimental results (for example those on canonical neurons, or the transformations in far/near space due to tool use, Costantini et al. 2014), confirm that one subject's goal-specific motor skills are selectively and flexibly recruited by his brain to make sense of the physical surroundings and highlight the opportunities of bodily interaction that are most appropriate to the spatial context. As I mentioned before, this picture does not rule out TSH, which is a valid model to explain at least one out of the many possible cognitive strategies that we use to attribute goals without relying on demanding meta-representations; but it undermines its presumption that the fundamental mechanisms of social cognition are inferential in nature and, therefore, it subtracts credibility to the claim that TSH is more fundamental and primitive than EFH.

References

- Anscombe, E. (1957). *Intention*. Oxford: Blackwell.
- Astington, J.W. & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology*, 9, 7–31.
- Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive mechanisms in mind reading. *Curr. Psychol. Cogn.*, 13, 513–552.
- Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. *Consciousness and cognition*, 17, 557–564.
- Bertenthal, B.I., Proffitt, D.R., Spetner, N.B., & Thomas, M.A. (1985). The development of infant sensitivity to biomechanical motions. *Child Development*, 56, 531–543.
- Berthoz, A. (2003). *La Décision*. Paris: Odile Jacob.
- Bratman, M. (2007). *Structures of Agency: Essays*. Oxford UK: Oxford University Press.
- Butterfill S. & Sinigaglia C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88 (1), 119–145.

- Call, J. & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12, 187–192.
- Cappuccio, M. (2012). Conoscenza nella mani. Come comprendiamo le azioni degli altri. In I. Gamelli (ed.), *Ma di che corpo parliamo? I saperi incorporati nell'educazione e nella cura*. Roma: Franco Angeli.
- Cappuccio, M. & Wheeler, M. (2012). Ground-level Intelligence: Inter-context frame problem and dynamics of the background. In Z. Radman (ed.), *Knowing without Thinking. Mind, Action, Cognition and the Phenomenon of the Background*. Basingstoke: Palgrave Macmillan.
- Costantini, M., Frassinetti, F., Maini, M., Ambrosini, E., Gallese, V., Sinigaglia, C. (2014) When a laser pen becomes a stick: Remapping of space by tool-using observation in hemispatial neglect. *Exp. Brain Res.*, 232 (10), 3233–3241.
- Csibra, G. & Gergely, G. (1998). The teleological origins of mentalistic action explanations: a developmental hypothesis. *Dev. Sci.*, 1, 255–259.
- Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, 72, 237–267.
- Csibra, G., Bíró, S., Koós, O., & Gergely, G., (2003). One-year-old infants use teleological representations of actions productively. *Cogn. Sci.*, 27, 111–133.
- Csibra, G. (2003) Teleological and referential understanding of action in infancy. *Philos. Trans. R Soc. B Biol. Sci.*, 29, 447–458.
- De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Does social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14 (10), 441–447.
- Dennett, D.C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
- Di Pellegrino, G. di, Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research / Experimentelle Hirnforschung / Experimentation cerebrale*, 91, 176–180.
- Dreyfus, H.L. (1992). *What Computers Still Can't Do*. Cambridge, MA: MIT Press.
- Dreyfus, H.L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation the relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*, 1, 367–383.
- Froese, T. & Gallagher, S. (2012). Getting interaction theory (IT) together. Integrating developmental, phenomenological, enactive, and dynamical approaches to social interaction. *Interaction Studies*, 13 (3), 436–468.
- Gallagher, S. (2001). The practice of mind: Theory, simulation, or primary interaction? *Journal of Consciousness Studies*, 8 (5–7), 83–107.
- Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, 17 (2), 535–543.
- Gallagher, S. & Hutto, D. (2008). Understanding others through Primary Interaction and Narrative Practice. In T. Zlatev, T. Racine, C. Sinha, & E. Itkonen (eds.), *The Shared Mind: Perspectives on Intersubjectivity*. Amsterdam: John Benjamins, 17–38.
- Gallagher, S. & Povinelli, D.J. (2012). Enactive and behavioral abstract accounts of social understanding in chimpanzees, infants, and adults. *Rev. Phil. Psych.*, 3, 145–169.
- Gallagher, S. & Zahavi, D. (2008). *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. London: Routledge.
- Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4, 23–48.
- Gallese, V. (2010). Embodied simulation and its role in intersubjectivity. In T. Fuchs, H.C. Sattel, & P. Henningsen (eds.), *The Embodied Self: Dimensions, Coherence and Disorders*. Stuttgart: Schattauer, 78–92.

- Gallese, V., Rochat, M., Sinigaglia, C., Cossu, G. (2009). Motor cognition and its role in the phylogeny and ontogeny of action understanding. *Developmental Psychology*, 45 (1), 103–113.
- Gergely, G. (2002). The development of understanding self and agency. In U. Goswami (ed.), *Blackwell's Handbook of Childhood Cognitive Development*. Malden, MA – Oxford – Carlton – Berlin: Blackwell, 26–46.
- Gergely, G. & Csibra, G. (1994). On the ascription of intentional content. *Cahiers de Psychologie Cognitive / Current Psychology of Cognition*, 13 (5), 584–589.
- Gergely, G. & Csibra, G. (1997). Teleological reasoning in infancy: the infant's naive theory of rational action. A reply to Premack and Premack. *Cognition*, 63, 227–233.
- Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, 7 (7), 287–292.
- Gergely, G. et al. (2002). Rational imitation in preverbal infants. *Nature* 415, 755.
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165–193.
- Gibson, E.J., Owsley, C.J., & Johnston, J. (1978). Perception of invariants by five-month-old infants: Differentiation of two types of motion. *Developmental Psychology*, 14, 407–415.
- Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. New York: Oxford University Press.
- Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.
- Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16 (1), 1–14.
- Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
- Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.
- Hutto, D.D. (2013). Action understanding: How low can you go? *Consciousness and Cognition*, 22 (3), 1142–1151.
- Mandler, J.M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 99, 587–604.
- Merleau-Ponty M. (1945). *Phénoménologie de la perception*. Paris: Gallimard.
- Metta, G., Sandini, G., Natale, L., Craighero, L., & Fadiga, L. (2006). Understanding mirror neurons: A bio-robotic approach. *Interaction Studies*, 7, 197–232.
- Morganti F., Carassa A., & Riva G. (eds.) (2008). *Enacting Intersubjectivity: A Cognitive and Social Perspective on the Study of Interactions*, Amsterdam: IOS Press.
- Poincaré, J.-H. (1908/1996). *Science and Method*. London: Routledge.
- Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, 36, 1–16.
- Prevete, R., Tessitore, G., Catanzariti, E., & Tamburrini, G. (2010). Perceiving affordances: a computational investigation of grasping affordances. *Cognitive Systems Research*, 12 (2), 122–133.
- Rizzolatti, G., Camarda, R., Fogassi, M., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Exp. Brain Res.*, 71, 491–507.
- Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review Neuroscience*, 2, 661–670.
- Rochat, M., Serra, E., Fadiga, L., & Gallese, V. (2008). The evolution of social cognition: goal familiarity shapes monkeys' action understanding. *Current Biology*, 18, 227–232.

Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11, 264–274.

Rosenbaum, D. (2008). Reaching while walking: Reaching distance costs more than walking distance. *Psychonomic Bulletin & Review*, 15 (6), 1100–1104.

Sinigaglia, C. (2008). Enactive understanding and motor intentionality. In F. Morganti, A. Carassa, & G. Riva (eds.), *Enacting Intersubjectivity: A Cognitive and Social Perspective to Study of Interactions*. Amsterdam: IOS Press, 17–32.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*, Harvard University Press

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics, and the frame problem. *International Journal of Philosophical Studies*, 16 (3), 323–349.

Wellman, H.M. (1991). From desires to beliefs: acquisition of a theory of mind. In A. Whiten (ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*. Oxford: Basil Blackwell, 19–38.

Wellman, H.M. (2002) Understanding the psychological world: developing a theory of mind. In U. Goswami (ed.), *Blackwell's Handbook of Childhood Cognitive Development*. Malden, MA – Oxford – Carlton – Berlin: Blackwell, 167–187.

Woodward, A.L. (1998) Infants selectively encode the goal object of an actor 's reach. *Cognition*, 69, 1–34

Woodward, A.L. & Sommerville, J.A. (2000) Twelve-month-old infants interpret action in context. *Psychol. Sci.*, 11, 73–77.

Umiltà, M.A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., & Rizzolatti, G. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 2209–2213.

Massimiliano L. Cappuccio

Zaključivanje ili upoznatost?

Utjelovljeni korijeni socijalne spoznaje

Sažetak

Razmatram dvije različite deflacijske teorije u socijalnoj spoznaji koje nastoje objasniti razumijevanje djelovanja bez potreba za metareprezentacijskim procesima ili procesima čitanja uma. Prva je »hipoteza teleološkog stava« (TSH), koja tvrdi da o namjeravanom cilju određenog opaženog djelovanja zaključujemo samo na temelju opažanja učinaka djelovanja i situacijskih ograničenja. Drugu teoriju sam odlučio nazvati »hipoteza utjelovljene sličnosti« (EFH) kako bih obuhvatio sve teorije koje tvrde da namjeravani cilj određenog djelovanja prepoznavamo na temelju opažajne ili motoričke ekspertize razvijene unutar senzomotoričkih kontingencija povezanih s kontekstom djelovanja. Temeljni je zahtjev TSH da promatrač može pripisati efikasnost, stoga i racionalnost, opaženim pokretima djelatnika, dok je temeljni zahtjev EFH teorije taj da je promatrač na neki način izložen opažajnim ili motoričkim detaljima opaženog djelatnikovog djelovanja. Tvrdim da EFH opisuje primitivniji i temeljniji oblik razumijevanja djelovanja, tj. onaj oblik koji TSH nužno pretpostavlja: ustvari, iako prepoznavanje efikasnosti nije ni nužan niti dovoljan uvjet za detektiranje povezanost i s ciljem, neka vrsta opažajne ili motoričke upoznatosti s detaljima opaženog konteksta djelovanja uvijek je nužna za bilo kakvo pripisivanje efikasnosti, a time i racionalnosti, promatranom djelatniku. Zaključujem da, iako TSH zasigurno može biti učinkovita u opisu određenih racionalnih oblika razumijevanja djelovanja, implicitno zahtijeva da EFH bude istinita, budući da bi takav sustav zaključivanja bio neutemeljen bez pretpostavljene upoznate pozadine utjelovljene ekspertize.

Ključne riječi

socijalna spoznaja, teleološki stav, zrcalni neuroni, racionalnost, intencionalni stav, hipoteza direktnog odgovaranja

Support Publication #15 [SP15]

Alotaibi M., Alnajjar F., Cappuccio M.L., Khalid S. (2023),
“Efficacy of Emerging Technologies to Manage Childhood Obesity: A Systematic Review”
Diabetes Metab Syndr Obes. 2022; 15: 1227-1244
<https://doi.org/10.2147/DMSO.S357176>

Open access paper.

Efficacy of Emerging Technologies to Manage Childhood Obesity

Mohammad Alotaibi¹, Fady Alnajjar², Massimiliano Cappuccio³, Sumaya Khalid², Tareq Alhmiedat^{1,4}, Omar Mubin⁵

¹Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia; ²College of Information Technology, United Arab Emirates University, Abu Dhabi, United Arab Emirates; ³School of Engineering and IT, University of New South Wales Canberra, Canberra, Australia; ⁴Industrial Innovation & Robotics Center, University of Tabuk, Tabuk, Saudi Arabia; ⁵School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia

Correspondence: Fady Alnajjar, Email fady.alnajjar@uaeu.ac.ae

Abstract: Childhood obesity is a widespread medical condition and presents a formidable challenge for public health. Long-term treatment strategies and early prevention strategies are required because obese children are more likely to carry this condition into adulthood, increasing their risk of developing other major health disorders. The present review analyses various technological interventions available for childhood obesity prevention and treatment. It also examines whether machine learning and technological interventions can play vital roles in its management. Twenty-six studies were shortlisted for the review using various technological strategies and analysed regarding their efficacy. While most of the selected studies showed positive outcomes, there was a lack of studies using robots and artificial intelligence to manage obesity in children. The use of machine learning was observed in various studies, and the integration of social robots and other efficacious strategies may be effective for treating childhood obesity in the future.

Keywords: childhood obesity, intervention, exergaming, machine learning, artificial intelligence

Introduction

Childhood obesity is a major concern for healthcare systems and remains a challenge for medical innovation.¹ Cases of overweight or obese children (infant and young) have significantly increased worldwide to 41 million and are continuously increasing.² The huge socioeconomic transformations have significantly affected the nutrition and development of children and adolescents, particularly in developing countries. Increasing rates of urbanization has had a global effect on lifestyle, reducing the frequency of physical activity and increasing the calorie intake of nontraditional foods and “fast foods” rather than fruits and vegetables.^{3,4} In addition to children showing a greater frequency of being overweight or obese, these food choices lead to nutritional deficits in growing children, which may lead to future health issues.⁵

Without intervention, overweight or obese children are more prone to carrying their condition into adulthood, possibly increasing their chances of developing diseases such as diabetes, cardiovascular diseases, and cancer.^{6–8} Approximately 33% of preschool-aged children who are obese or overweight become obese adults, and 50% of obese school-aged children and 80% of adolescents carry obesity into adulthood.⁹ Therefore, it is crucial to develop strategies for the early diagnosis, prevention, and management of childhood obesity.

Currently available conventional interventions to manage weight are not always feasible as they are often inaccessible because of high cost, commute inconveniences, and lack of facilities.¹⁰ Extreme cases may require bariatric surgery, which is both expensive and risky.¹¹ Approximately 80% of children aged 12–17 years are hesitant to engage in any kind of coordinated activities, such as sports; therefore, a need for creative intervention is yet required.¹²

Over the past decade, advances in portable digital technologies that provide personalized access to relevant information have facilitated the evolution of preventive strategies and interactive aids and have become readily available. Smart digital health interventions are designed to cater to a wide number of issues because they help track and monitor health as well as collect information for analysis by a specialist to determine the required treatment strategy. These interventions

include weight management programs based on websites¹³ or social media.¹⁴ They are wearable¹⁵ and are easily accessible via websites and mobile applications. These interventions are easily adaptable to different environments, affordable and time saving. Most importantly, strategies based on these technological tools keep children motivated and engaged while enjoying the comfort of their homes. Technology-based interventions for childhood obesity can be mainly classified into the following categories: (1) active video games or exergaming; (2) website, social media-based activities¹⁶ or mobile application-based activities;¹⁷ (3) machine learning (ML)-based strategies.¹⁸ A combination of those categories does also exist.

Mobile phones are the most readily and commonly available gadgets. They can be conveniently connected to wearables and sensors, providing the user with real-time data and monitoring physiological changes with internet cloud connectivity. Combining sensors and connectivity enables behavior (eg, physical activity and diet) to be monitored and influenced, and health care providers or experimenters can be alerted to changes (eg, weight and glucose). This technology offers a great opportunity to tap into new media channels that are integral to the youth culture because mobile phones are already widely used by children.¹⁹

The use of exergames, which combine exercise and video games, is another concept that is gaining popularity. These games are designed to encourage children to perform physical exercise or execute certain movements to clear levels and earn rewards in the game. They keep children motivated and engaged over time, making it easier for them to consistently follow a regular fitness program. Many available applications use this concept in combination with a wearable device. Exergames are suitable for every age group. For example, *Let us move* (To move!) is a program designed in the form of a video game to prevent and treat obesity in children.²⁰ Zamzee is another application for adolescents²¹ that records physical activity via a device. The number of points earned is calculated by the computer and can be exchanged for goods from partnering companies.

ML is an emerging strategy^{22,23} implemented by many computerized systems used in a clinical context.^{24,25} An ML-based approach provides beneficial insight into data to extract relevant information and has been actively used for the prognosis and diagnosis of several chronic conditions, particularly in patients with diabetes and sepsis.^{26,27}

Despite being potentially very efficacious in obesity, there are surprisingly few reviews in the literature regarding the application of ML for childhood obesity management and treatment. Available reviews mainly focus on mobile devices, wearables, and interactive electronic media or predictive models based on statistics. One review²⁸ addressed the potential use of computer decision systems (CDS) showing the main features and results of studies that use ML techniques and CDS.

Social robots are also being developed and tested in various healthcare concepts. Social robots have significant potential to assist healthcare workers and deliver customized care to children because they are more capable than other types of interactive technologies to positively influence the behavior and motivation of the human interactants. This superior potential derives from the fact that social robots, unlike other assistive technologies, are embodied agents with anthropomorphic or zoomorphic features. In some cases, robots also have a limited ability to navigate or manipulate the physical environment, enabling them to engage in embodied interaction with humans and share a concrete focus of attention (object or activity) with them. Furthermore, the ability to express emotions and interest is much more developed in person-like agents, such as robots, compared with impersonal technologies, such as mobile phone or tablet applications.

Robinson et al²⁹ described the potential of robots and demonstrated that an autonomous social robot can efficaciously motivate adults to reduce the consumption of high-calorie snacks, aiding a low-powered therapy on dietary intake with no human intervention. Building on this and analogous examples of robots acting as coaches and lifestyle advisors, a “virtuous” theoretical approach to robot design has been proposed.^{30,31} This “virtuous robotics” approach can be applied to obesity management where robots can be programmed and deployed to cater to specific health and psychological requirements, motivate children to perform physical activities, and encourage healthier eating habits. Robots are suited for this motivational/assistive role because they are perceived as toys by children and can be a source of companionship, provide comfort and relief, and gain the trust of children, which is indispensable to give children explicit instructions and help in reforming their dietary habits.

Several reviews have explored obesity management using different technologies in children and adolescents and compared their efficacy in managing and preventing obesity via various channels.^{32,33} No reviews were specifically

dedicated to analyzing the efficacy of robots and AI-based strategies compared with other technology-based techniques for obesity management in children and adolescents that have received more attention to date. The present systematic review aims to fill the gap by analyzing and comparing research studies published between the years 2010–2021 to map how ML techniques are used and how they compare with non-AI-based technological strategies in obesity management for children.

Materials and Methods

The present review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines using the following strategic approach. Studies included in this review were identified systematically by searching databases using the following keywords: obesity, obesity management, children, technology, exergaming, robots, machine learning, and overweight. The databases consulted were PubMed, IEEE Xplore, Scopus, and Google Scholar. Moreover, a few references of significant studies^{34,35} were scrutinized to extract more relevant articles suited to our review for further investigation and were shortlisted according to our criteria.

Inclusion and Exclusion Criteria

Studies fulfilled the inclusion criteria if they were (1) published in English; (2) published within the years 2010–2021; (3) based on obesity prevention or management (4) included participants aged ≤ 18 years; (5) were randomized controlled trials (RCTs) or interventions; (6) Used AI, robots or ML techniques.

Studies were excluded if they were (1) not in English; (2) had no reliable outcomes; (3) had participants aged > 18 years; (4) did not use any types of technology in their research; and (5) were reviews, book chapters, thesis or conference material.

Quality Assessment Strategy

A modified Delphi list was used to assess the quality of each study.³⁶ The assessment quality was improved by considering the following elements: age range, sample size, and intervention duration. The latter two were considered with the understanding that a smaller sample size and a shorter intervention duration might negatively impact generalizability and reproducibility.

Additional considerations were given to a study's purpose, whether they were preventive, screening-based, experimental or treatment-based. Also, the methods of the study were taken into consideration if they were activity/video game based, web/mobile based or ML based. The choice of elements included in the quality assessment was guided by a previously published systematic review conducted by Nouchi et al.³⁴

Results

A total of 1244 articles obtained from various databases were screened and duplicate articles were rejected. The titles and abstracts of the remaining 250 articles were screened for relevancy, and 75 articles were identified for full-text reading, 26 were finalized for review. A review of each study was performed, and information related to the layout and purpose of the study, as well as details of the participants (age, gender, size, and characteristics) were extracted.

The selection strategy for our review is illustrated (Figure 1). Table 1 shows an overview of the study characteristics, Table 2 outlines the key findings of the included literature, and Table 3 shows the assessment of the quality of studies using a delphi list. The studies were of various types, such as RCTs, experimental studies, or screening studies. The overall sample size analyzed ranged from 17 to 18,818, and the age of the participants ranged from 9 months to 18 years. Most studies were not gender specific and included both males and females, except for one study by Staiano et al that only included females.⁴²

All studies focused on technology-based interventions. Most studies^{13,37–39,41–45,49–54} involved overweight or obese participants, a few^{13,35,36,46–48,52,53} included healthy participants, and others did not specify the characteristics of the participants.^{40,55–59} The technology-based interventions were classified into three categories: (1) active video games or exergaming; (2) web, internet-based activities or mobile application-based activities; (3) ML-based strategies;

The duration of the intervention ranged from 1 week³⁵ to 3 years,⁴¹ and the studies that used ML-based algorithms did not always specify the time period for data collection. Intervention intensity was reported in most studies, but not all,^{40,54–59} and

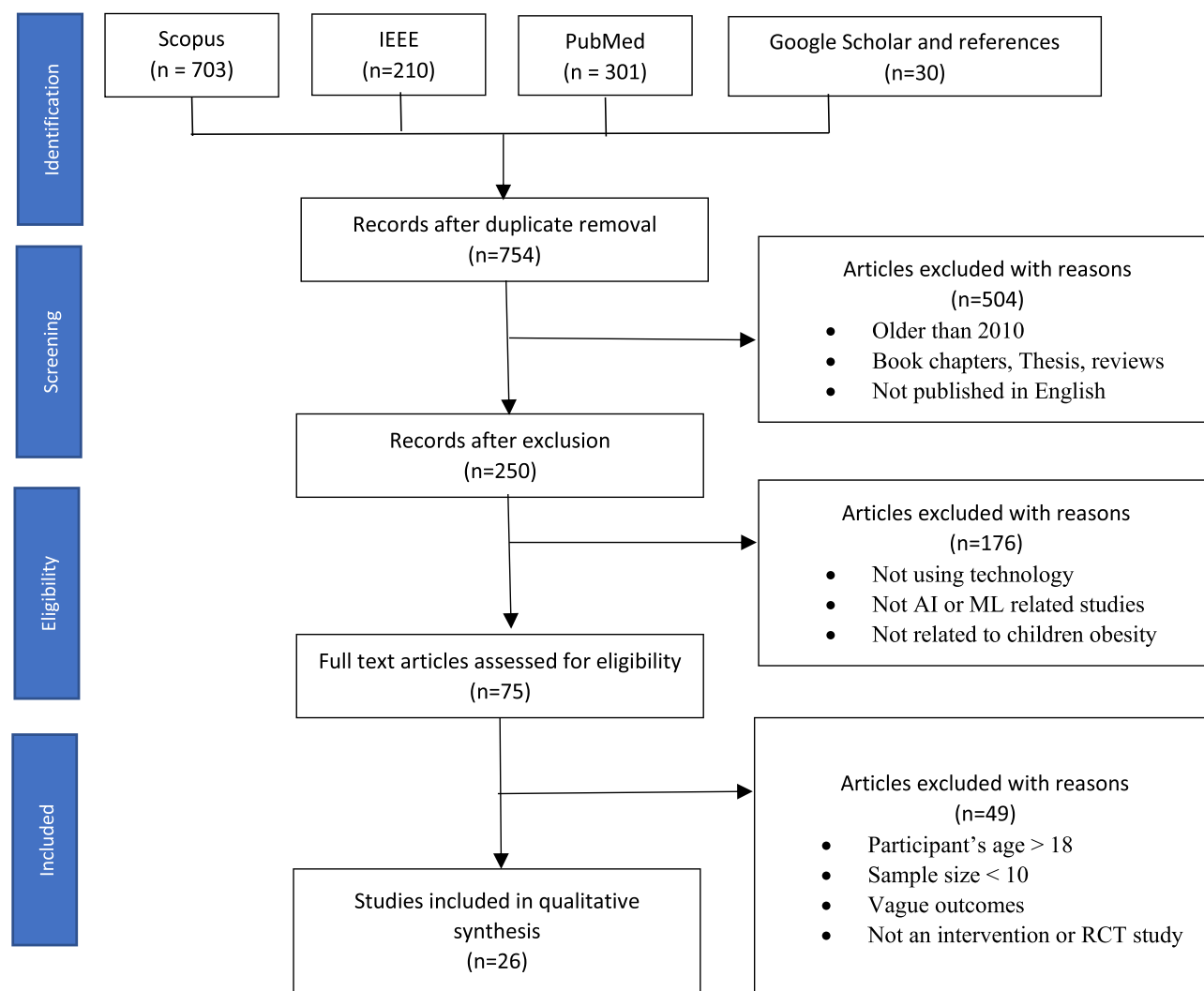


Figure 1 Search strategy used for selection of studies.

was highly heterogeneous. Ten studies^{13,35,41,42,46,47,49,51–53} were RCTs that compared technology-driven intervention with a control group and used a non-technology-based intervention⁵⁶ in which the control group played video games that were sedentary compared with the intervention group. The outcome of the studies varied according to the aim, but all comprised a body mass outcome, such as body mass index (BMI), BMI z-score, or BMI percentile. Other outcomes commonly observed included other anthropometrical, diet-based, and physical activity measurements.

Active Video Game Interventions

Twelve studies^{35,37–47} investigated the efficacy of active video games or exergaming. Lindberg et al³⁵ incorporated the gaming program Running Othello into the South Korean school physical education (PE) curriculum for the intervention group, while the control group used printed leaflets and the regular curriculum. Positive results were observed in the intervention group as children were more engaged and showed increased heart rates compared with the control group. In another study, Argarini et al³⁷ investigated the efficacy of moderate intensity exergaming on children with a BMI in the 85th percentile. Exergames were played using Xbox 360, and dance and sports-based games were played for four weeks. The results showed reduced BMI and improved fundamental movement skill, with no significant discrepancy between male and female participants. Adamo et al³⁸ used the interactive video game, “GameBike” and included stationary cycling for two groups: one group performed stationary cycling with music and the other group used GameBike while

Table I Characteristics of the Included Literature

Study Number	Author, Year	Study Design	Study Purpose	Participant Characteristics			
				Number of Participants 1) Intervention 2) Control	Gender	Age in Years	Characteristics
1	Lindberg et al., 2016 ³⁵	RCT	Education and Prevention	1)32 2)29	M + F	10	Healthy 3rd grade elementary children
2	Yang et al., 2017 ³⁶	Non-RCT	Prevention	1) 558 2) 288	M + F	10–12	Healthy school students
3	Argarini et al., 2020 ³⁷	Experimental study	Treatment	17	M + F	6–12	BMI >85th percentile
4	Adamo et al., 2010 ³⁸	Experimental study	Activity	30	M + F	12–17	Overweight children
5	Staiano et al., 2018 ³⁹	RCT	Treatment	1)23 2)23	M + F	10–12	BMI z-score
6	Ruggiero et al., 2020 ⁴⁰	Comparative study	Experiment	48	M + F	7–13	–
7	Trost et al., 2014 ⁴¹	Group RCT	Activity	1)34 41	M + F	8–12	BMI >85th percentile
8	Staiano et al., 2017 ⁴²	RCT	Activity	1)21 2)19	F	14–18	Overweight and obese girls
9	Maddison et al., 2011 ⁴³	RCT	Activity	1)160 162	M + F	10–14	Overweight and obese children
10	Duman et al., 2016 ⁴⁴	Experimental study	Activity	50	M + F	12–16	Slightly overweight and obese
11	Wagener et al., 2012 ⁴⁵	Experimental study	Activity	1)21 20	M + F	12–18	BMI >95th percentile
12	Coknaz et al., 2019 ⁴⁶	RCT	Prevention	106	M+F	9.6 (mean)	Healthy children
13	Gao et al., 2019 ⁴⁷	RCT	Prevention	32	M+F	4–6 yrs	Healthy children
14	Curiel et al., 2020 ⁴⁸	Experimental study	Training and Prevention	60	M + F	9 (mean)	Healthy children
15	Nawi et al., 2015 ⁴⁹	RCT	Education	1)47 2)50	M + F	16	BMI >25 kg/m ²
16	Rio et al., 2019 ⁵⁰	Quasi-experimental control trial	Activity and Education	1)13 2)10	M + F	6–12	Obese children
17	Chen et al., 2019 ⁵¹	RCT	Treatment and Prevention	2)54	M + F	12–15	BMI ≥85th percentile
18	Rerksupphol et al., 2017 ⁵²	RCT	Prevention	3)217	M+F	10.6 (mean)	Healthy children

(Continued)

Table I (Continued).

Study Number	Author, Year	Study Design	Study Purpose	Participant Characteristics			
				Number of Participants 1) Intervention 2) Control	Gender	Age in Years	Characteristics
19	Nystrom et al. 2017 ⁵³	RCT	Prevention	4)263	M+F	4.5 (mean)	Healthy children
20	Hammersley et al. 2019 ¹³	RCT	Prevention	5)86	M+F	2–5 yrs	Healthy children
21	Fiechtner et al., 2016 ⁵⁴	Experimental study	Monitoring	498	M = F	6–12	BMI >95th percentile
22	Lingren et al., 2016 ⁵⁵	ML-based study	Detection	428	M + F	1–6	–
23	Rios-Julian et al., 2017 ⁵⁶	ML-based study	Screening	221	M + F	6–13	–
24	Fergus et al., 2015 ⁵⁷	ML-based study	Screening	28	M + F	10–11	–
25	Dugan et al., 2015 ⁵⁸	ML-based study	Screening	7519	M + F	0–2	–
26	Balbir et al., 2020 ⁵⁹	ML-based study	Screening	18,818	M + F	9 months–14 years	–

Abbreviations: BMI, body mass index; F, female; M, male; ML, machine learning; RCT, randomized controlled trial.

performing stationary cycling. Regularity of exercise, energy consumption, aerobic fitness, body composition, and the risk of heart disorder in overweight and obese children were monitored. The music group performed considerably better and used more strength than the GameBike group. There were no statistically significant differences between or within groups with respect to BMI.

Staiano et al³⁹ performed exergaming with Kinect and Xbox 360 gaming consoles using four simultaneous exergames, whereas the control group was required to maintain their regular level of physical activity. Statistically significant differences were observed in the intervention group, which showed improved scores, increased physical activity, and improved cardiometabolic health. Ruggiero et al⁴⁰ developed the exergame “MyPlatePick” that combines nutrition education and physical activity to promote movement and encourage healthy behavioral changes. Positive changes in eating behavior and increased physical activity were noticed.

Trost et al⁴¹ evaluated the effect of exergaming on children participating in a group-based program for managing weight. Improvements in physical activity were observed in the gaming group compared with the control group, which only took part in the program and not in the exergaming. The active video game group showed reduced weight and BMI. Another study by Staiano et al⁴² examined the effects of dance-based exergaming on females matched to a control group who followed their normal level of physical activity. No statistically significant differences were observed between the groups at follow-up.

Rio et al⁵⁰ tested a platform in which weekly group sessions with training were conducted on healthy-eating habits and active video games, and the participants developed their own vocal projects based on healthy activities. This study documented a positive effect on the intervention group in terms of nutritional knowledge and adherence to Mediterranean diet. Wagner et al⁴⁵ studied the efficacy of dance-based exergaming on obese adolescents’ observed capability to

Table 2 Key Findings of the Studies

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
1	Lindberg et al., 2016 ³⁵	Exergame RO2 with smartphone and wrist band using South Korean PE curriculum	1 week, 15 min per student	The intervention group studied the curriculum by playing RO2, while the control group learned the curriculum using handouts. Questionnaires were developed to be answered by game groups to rate RO2. A quiz was held at the end of the intervention to calculate learning and retention in both groups.	Positive outcomes were observed; exergames motivated students to play irrespective of their physique. Learning via RO2 was more effective, kept students involved, and increased their heart rates. Learning outcomes showed significant variations between both groups.
2	Yang et al., 2017 ³⁶	HAPPYME platform using quests involving physical activities and healthy dietary habits	12-week test and 6-month follow-up	HAPPYME platform helps teachers and parents to monitor and provide encouragement to participating children. The application asks students to complete quests involving physical activities and healthy eating habits that help prevent obesity. The child acquires points after completing each quest, which adds a motivational aspect. The platform sends feedback to children and parents for monitoring their daily and weekly performance.	Demonstrated the efficacy of a mobile service with a comprehensive intervention program by measuring anthropometric parameters, such as body weight, height, BMI, and percentiles.
3	Argarini et al., 2020 ³⁷	Moderate-intensity exergaming	12 exercises (30–40 min per session) 3 sessions per week for 4 weeks	All measurements were made three days before the first exercise and after exercise, except energy expenditure. Exergaming was played using Microsoft Xbox 360 and Kinect Consoles. Game types: Kinect Dance and Kinect Sports.	Normal distribution with no significant difference between boys and girls. Regular moderate exergaming for 4 weeks in obese and overweight children can reduce BMI and improve fundamental movement skills.
4	Adamo et al., 2010 ³⁸	Interactive video game with stationary cycling	10 weeks	Divided into two groups: (1) interactive video game and stationary cycling (GameBike) and (2) stationary cycling and music.	Similar improvements were observed in the body composition. Cholesterol profiles and fitness in both music and cycling as well as stationary cycling while playing video games in overweight and obese teenagers resulted in improved attendance and a robust concentration of physical activity.

(Continued)

Table 2 (Continued).

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
5	Staiano et al., 2018 ³⁹	Exergaming using a gaming console	1 h per session, 3 sessions a week for 24 weeks	Randomized group: • Four exergames using Kinect and Xbox 360 gaming console. • A FitBit Zip was worn during the 24-week period. • Regular meetings with a fitness coach to keep the child motivated. Control group: • Maintained their regular physical activity. • Xbox given at final clinical visit.	BMI z-score, cardiometabolic health, and physical activity levels all improved when performed at home regularly.
6	Ruggiero et al., 2020 ⁴⁰	Exergaming and educational exercises	Not available	Physical activity and nutrition education combined into technological game created to encourage youth for healthier behavioral change to combat obesity. The educational exergame “MyPlatePicks” promoted movement, delivered knowledge, improved motivation, and changed behavior associated with healthy eating and physical activity.	Initial evaluation showed positive results, changes in physical activity, and healthy eating behavior.
7	Trost et al., 2014 ⁴¹	Active gaming and educational program	8–16-week study	A family-based pediatric weight management program (JOIN for ME) was provided to the participants. Gaming console and motion capture device with two activity games were given to the activity group. Console was given to the program group only after program completion.	Adding active video gaming for pediatric weight management curriculum showed promising effects on the activity and weight.
8	Staiano et al., 2017 ⁴²	Dance-based exergaming	1 h per session, 3 sessions per week for 4 weeks	Kinect for Xbox 360 used for dance-based exergaming using Just Dance and Dance Central. Attendance was encouraged, and motivation to exercise was performed by providing incentives. Control group maintained their normal level of activity.	Bone mineral density increased and body fat reduced.

(Continued)

Table 2 (Continued).

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
9	Maddison et al., 2011 ⁴³	Active video games	24 weeks	Intervention group used the following hardware: EyeToy camera, Dancemat, and active video games. Control group used their usual video game. Outcome measure was BMI.	A small but considerable impact on BMI and body composition was observed with active video game intervention.
10	Duman et al., 2016 ⁴⁴	Active video games	3 days per week for 8 weeks	Aerobic exercises with music and active video games were used, and BMI was measured after 8 weeks. Data collected were then analyzed using SPSS 18.0 program.	Exercise program along with active video games showed promising effects as well as a constructive impact on self-respect and psychological wellbeing.
11	Wagener et al., 2012 ⁴⁵	Dance-based exergaming	3 sessions (40 min per session) per week for 10 weeks	Participants (2–3) stood on dance pads with colorful arrows placed out in a cross shape. Feet were used to hit arrows corresponding to musical and visual cues displayed on the screen.	Increased competence to exercise regularly.
12	Coknaz et al. 2019 ⁴⁶	A parallel RCT to evaluate the effect of physical fitness, reaction times, self-perception and enjoyment levels by active video games in inactive and technologically preoccupied children	12 weeks	Active games was used as intervention for children from 3 schools and 1 school was a control group. Primary outcomes were weight, body mass and fat ratios.	Active video games led to a reduction in weight gain. They are also beneficial tools in diverting children from inactivity thus preventing obesity.
13	Gao et al. 2019 ⁴⁷	Home-based educational exergaming intervention effects on preschoolers.	12 weeks	The participants were divided into intervention and control groups. The intervention group did exergaming on Leap TV gaming console control group continued with their regular physical activities,	Home-based educational exergaming may positively impact cognitive flexibility in preschoolers.
14	Curiel et al., 2020 ⁴⁸	Video game FoodRateMaster to improve knowledge regarding food and encourage healthy eating behavior	12 sessions in 6 weeks	Gaming was used to improve nutritional knowledge and apply techniques to bring about behavioral change, increase awareness of unhealthy and healthy foods, and improve consumption of healthy food. This active game increases physical activity.	Positive outcome was observed as children showed improved food knowledge before and after the game. Parents gave positive feedback and reported that they observed changes in eating habits in their children.

(Continued)

Table 2 (Continued).

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
15	Nawi et al., 2015 ⁴⁹	Educational information given via internet and pamphlets	12 weeks	Internet-based intervention ObeseGo! was used in a randomized group (healthy lifestyle and diet information were provided via the internet). Control group was given pamphlets containing health education material.	ObeseGo! showed minimal effect in the reduction of BMI, waist circumference, and percentage body fat.
16	Rio et al., 2019 ⁵⁰	Physical activity, vocation projects, and training sessions	3 years	Along with training, weekly group sessions on eating healthy as well as health education were conducted using active video games. Vocal projects were created to identify healthy activities using virtual platforms.	Improvements in understanding and obedience to the Mediterranean diet were observed in the experimental group.
17	Chen et al., 2019 ⁵¹	Web-based behavior program	8 months	Web-based program to improve self-efficiency for enhancing the understanding and use of problem-solving skills in relation to nutrition and physical activity. An interactive dietary training software program (The Wok) was custom-made according to common Chinese food and was used by participants to check the nutritional information. Internet information included voice overs, graphics, and comics.	Decreased waist-to-hip ratio, reduced blood pressure, increased the intake of vegetables and fruits, increased physical activity, and increased knowledge about nutrition.
18	Rerksuppaphol et al. 2017 ⁵²	Internet based obesity program for obesity prevention in Thai school children	Four months, monthly	Healthy children were randomized to intervention (internet based) program or the control group. Anthropometric characteristics were screened at baseline and for four months in monthly intervals. Changes in the percentage of overweight/obese children and changes in BMI at the end of study were the primary and secondary outcome, respectively	Internet the based obesity prevention program was effective and helped in addressing the rising obesity in children.

(Continued)

Table 2 (Continued).

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
19	Nystrom et al. 2017 ⁵³	Mobile health (mHealth) obesity prevention program on body fat, dietary habits, and physical activity in healthy children	6 months	p. 6-month mHealth program. Where the primary outcome was fat mass index (FMI), whereas the secondary outcomes were intakes healthy and unhealthy foods and time spent being non-active and in moderate-to-vigorous physical activity	No difference between the intervention and control group but post intervention the intervention group showed higher composite score than the control group, especially in children with a higher FMI
20	Hammersley et al. 2019 ¹³	Internet and email based intervention lifestyle program	11 weeks, 6 months follow up	11-week internet-based healthy lifestyle program, by fortnightly emails for 3 months for intervention group. Comparison participants received email communication only. BMI was the primary outcome.	eHealth childhood obesity prevention improved dietary-related practices and self-efficacy but did not reduce BMI.
21	Fiechtner et al., 2016 ⁵⁴	Monitoring behavior	1 year	CDS system used by clinicians. Family self-guided changes in behavior and coaching for health. Results were recorded as a change in BMI z-score as well as in the intake of sugar-based drinks, fruits, and vegetables.	Distance to a supermarket <1 mile and intervention improved fruit and vegetable consumption by 0.29 portions per day and reduced BMI z-score by 0.04 units compared with controls.
22	Lingren et al., 2016 ⁵⁵	ML-based algorithms	Not available	ML- and rule-based algorithms from structured and unstructured data from two electronic health record databases were developed to detect severe early childhood obesity and high long-term risk of developing obesity-related comorbidities.	Precision was stressed in the high-fidelity group. The rule-based algorithm achieved the best overall results.
23	Rios-Julian et al., 2017 ⁵⁶	ML-based algorithms	Not available	The study evaluated the practicability of an automated screening tool for diagnosing obesity using anthropometric variables and an ML approach.	High capacity shown by classifiers for assessing whether or not the participant was overweight.
24	Fergus et al., 2015 ⁵⁷	ML-based algorithms	Not available	Activities were evaluated using data recorded from wearable accelerometer sensors in free-living environments. Physical activity and assessment was performed using a multilayer perceptron neural network for the classification of physical activities by the type of activity.	Overall accuracy, 96%; sensitivity, 95%; specificity, 99%.

(Continued)

Table 2 (Continued).

Study Number	Author, Year	Intervention	Intervention Duration and Intensity	Intervention Strategy	Key Findings
25	Dugan et al., 2015 ⁵⁸	ML-based algorithms	Data collected over 9 years	An algorithm was used to predict obesity in children aged >2 years with data accumulated before the second birthday using CHICA. Six ML techniques were used: RandomTree, RandomForest, J48, ID3, Naïve Bayes, and Bayes.	Accurate model was created.
26	Balbiri et al., 2020 ⁵⁹	ML-based algorithms	Not available	Overweight or obesity prediction for young people using ML techniques.	Approximately 90% accuracy in prediction for the target class was attained.

Abbreviations: BMI, body mass index; CDS, computer decision system; CHICA, Child Health Improvement through Computer Automation; ML, machine learning; PE, physical education; RO2, Running Othello.

exercise, self-regulation, and BMI compared with a wait-list control group. A statistically significant increase in the perceived ability to exercise was found based on self-reporting by the participants in the intervention group compared with the control group. However, no major variations were found for BMI z-score.

Duman et al⁴⁴ used aerobics along with music and video games in their eight-week study and found a positive outcome in the final measurements with the exercise program using video games. Maddison et al⁴³ investigated the effect of video games on physical activity and fitness compared with standard video games. They found a decrease in BMI and BMI z-score of active group participants compared with the children in the control group.

Web-Based/Mobile-Based Studies

Eight studies^{13,36,48–53} used interventions based on internet or smart phone applications. Yang et al³⁶ developed a mobile-based HAPPYME platform that asks students to complete quests involving physical activity and healthy eating habits. The platform sends feedback to their parents who can then monitor the performance. Curiel et al⁴⁸ developed a video game-based mobile application that enabled participants to complete quests based on physical activity and healthy eating. Another study by Nawi et al⁴⁹ used the internet-based intervention ObeseGo! that provides information about healthy lifestyle and diet nutrition via the internet. This intervention showed no significant effect on BMI scores. Chen et al⁵⁶ developed a web-based program that comprised self-assessment activities to determine problem-solving skills and an interactive software program customized to Chinese food that allowed the user to check nutritional information about the food they ate. Fiechtner et al⁵⁴ developed a computerized clinician decision support system that monitored family behavior and found that living within one mile of a supermarket encouraged healthy eating habits.

ML-Based Studies

Five studies^{55–59} used ML-based strategies for their interventions. Lingren et al⁵⁵ created both rule-based and ML-based algorithms using structured and unstructured data from two large electronic health records to detect severe early childhood obesity and children at a long-term risk of developing complications. The rule-based algorithms performed better overall. Rios-Julian et al⁵⁶ used ML to estimate the feasibility of developing an automated screening tool for diagnosing obesity using anthropometric variables. Fergus et al⁵⁷ used a wearable accelerometer to collect data that was used to evaluate physical tasks, and a multilayer perceptron neural network was used to classify physical activities according to type of activity. Dugan et al⁵⁸ used ML on the collected data to develop an algorithm that could predict childhood obesity in children aged >2 years using data collected before their second birthday. Six different ML

Table 3 Quality Assessment of Included Studies

Sno	Author, Year	Q1	Q2	Q3	Q4	Q5
1	Lindberg et al., 2016 ³⁵	Y	Y	Y	N	N
2	Yang et al., 2017 ³⁶	Y	Y	Y	Y	N
4	Argarini et al., 2020 ³⁷	N	Y	Y	Y	N
5	Adamo et al., 2010 ³⁸	Y	Y	Y	Y	N
6	Staiano et al., 2018 ³⁹	Y	Y	Y	Y	N
7	Ruggiero et al., 2020 ⁴⁰	Y	Y	Y	?	N
9	Trost et al., 2014 ⁴¹	Y	Y	Y	Y	N
10	Staiano et al., 2017 ⁴²	Y	Y	Y	Y	N
19	Maddison et al., 2011 ⁴³	Y	Y	Y	Y	N
20	Duman et al., 2016 ⁴⁴	Y	Y	Y	Y	N
12	Wagener et al., 2012 ⁴⁵	Y	Y	Y	Y	N
24	Coknaz et al. 2019 ⁴⁶	Y	Y	Y	Y	N
26	Gao et al. 2019 ⁴⁷	Y	Y	Y	Y	N
3	Curiel et al., 2020 ⁴⁸	Y	Y	Y	Y	N
8	Nawi et al., 2015 ⁴⁹	Y	Y	Y	Y	N
11	Rio et al., 2019 ⁵⁰	Y	Y	Y	Y	N
21	Chen et al., 2019 ⁵¹	Y	Y	Y	Y	N
22	Rerksupphol et al. 2017 ⁵²	Y	Y	Y	Y	N
23	Nystrom et al. 2017 ⁵³	Y	Y	Y	Y	N
25	Hammersley et al. 2019 ¹³	Y	Y	Y	Y	N
13	Fiechtner et al., 2016 ⁵⁴	Y	Y	Y	Y	N
14	Lingren et al., 2016 ⁵⁵	Y	Y	N	Y	Y
15	Rios-Julian et al., 2017 ⁵⁶	Y	Y	N	?	Y
16	Fergus et al., 2015 ⁵⁷	Y	Y	N	?	Y
17	Dugan et al., 2015 ⁵⁸	Y	Y	N	?	Y
18	Balbir et al., 2020 ⁵⁹	Y	Y	N	?	Y

Notes: Q1, studies with sample size >20; Q2, studies with age details; Q3, studies with the sample characteristic; Q4, studies with intervention duration >3 weeks; Q5, studies with machine learning /artificial intelligence/robots.

techniques, namely, RandomTree, RandomForest, J48, ID3, Naïve Bayes, and Bayes, trained using the Child Health Improvement through Computer Automation system data showed that a precise, sensitive model could be formed. Balbir et al.⁵⁹ developed ML algorithms to predict the risk of young people becoming obese in the future with a 90% accuracy.

Discussion

The objective of this systematic review was to assess the use of technology-based interventions in the management and prevention of childhood obesity and evaluate the use of ML and other technological strategies. Various technologies have

been used, including video game-based interventions, mobile-based interventions, web-based interventions, phone communication-based interventions and machine learning-based interventions. It is important to note that along with these, multicomponent mediations, including both dietary intake and physical activity, are required to effectively overcome obesity.

Various parameters, such as BMI, body fat percentage, weight, heart rate, and BMI z-scores, have been used in these interventions to measure the progress and success of the interventions. Among the 26 studies under review, 21 studies showed positive significant outcomes on their respective outcome measures.

The present review classified the selected literature broadly into three categories: active video games or exergaming; internet or mobile-based interventions; and ML-based interventions. The two categories of web- or internet-based and mobile-based strategies were combined as one. The search did not yield any articles based on robot-based techniques to manage obesity in children; therefore, no such studies were included for that category, but one study⁶⁰ was found to be using a social robot to manage and spread awareness about obesity in children of Saudi Arabia; however, since it was not an RCT or an intervention with outcomes, it was not included in the review literature.

Because of the complex etiology of childhood and adolescence obesity and the diversity of its possible influencing factors (eg, biological, psychological, dietary, behavioral, and cultural), conventional statistical methods, relying on linear models, have limitations as the analyses they offer typically have a limited predictive power. Most cases use the same prediction values transformed slightly, and it is difficult to find models capable of handling multidimensional data. In contrast, ML-based strategies are valuable tools for multidimensional datasets that involve complex relationships among multidomain variables. ML techniques can search for complicated nonlinear relations between prediction variables and response variables in an automated manner and do not require manual predefinition.⁶¹ Thus, the predictivity, ease of use, and robustness of ML-based models are superior to those of conventional models.³⁶ Five studies incorporated ML strategies in their interventions. Most of these studies scanned existing electronic health records to prevent or detect obesity. Lingren et al⁵⁵ detected severe cases of childhood obesity, and Rios-Julian et al⁵⁶ determined the possibility of developing automated screening tools to detect obesity based on anthropometric variables using ML. Dugan et al⁵⁸ and Singh et al⁵⁹ developed algorithms to predict obesity in children and youngsters.

If a comparison is to be made between ML-based techniques and the other technological methods, it is difficult at this point in the research as ML studies are more focused on screening the databases to identify patterns, which can help in prediction and other strategies involving a different approach by various methods of training, games and information communication. It is not possible to compare two entirely different methods with different objectives, even though the main objective stands in managing and prevention of obesity, the method used is vastly different for them. Since the search did not reveal any AI and robot-based interventions, it can be concluded that this area has a lot of potential for future and needs to be actively researched.

On observing closely most of the studies, it can be broadly recognized as preventive, educational, activity-based, data screening, and treatment-based studies. It was perceived that the preventive and educational interventions were focused more on making the participants aware of the different kinds of food and their impacts on the body. They used various gaming methods to educate children, or internet/telephone-based methods to make children aware of the evil effects of unhealthy foods, high sugar diets, etc. They also highlighted the importance of physical activities and their rewards.

Michie⁶² outlines effective techniques in healthy eating and physical activity interventions, stating what makes an intervention effective in successfully changing the lifestyle or behavior of the participants for the better. These were the basic techniques according to S. Michie to bring about change in behavior: formation of an intention, setting up a specific goal, providing performance feedback, behavior self-monitoring and behavior goal reviews. Most of our studies did follow this pattern to some extent, where they had a targeted goal, set up intervention, tracked performance and feedback was taken even after post-intervention to assess self-monitoring efficacy.

Activity-based intervention were focused on exergaming, which kept children engaged and motivated to continue the program. The games came with levels and rewards, which kept the child interested and as a result, continued to play games. These activity-based methods were used for both managing obesity and in prevention strategies.

Even though varied in nature, the studies included mostly had very similar primary outcome measures like BMI score, body fat, BMI-z score, physical activity duration, weight measures, etc. Most studies focused on both the genders and did not do any gender-specific study except one study, which was a female-based study.⁴²

The studies included in this review were of average quality, as only 5 of the 26 studies used ML or advanced techniques to detect or prevent obesity and did not specify the intervention duration. Furthermore, five studies did not define the characteristics of the participants. None of the studies used simple robots (without AI) or AI-based robots for the management of childhood obesity.

Overall, it can be seen that these different approaches to cater well to the childhood obesity issue have been effectively using technologies to keep children engaged and motivated when carrying out the required activities. Overall, it can be seen that these different approaches to cater well to the childhood obesity issue have been effective using creative methods to keep children engaged and motivated to carry out the required activities, using standard measures as primary outcomes, it gives a standardized way of evaluating the efficacy of each of these studies. Most studies showed positive results in both preventive and management-based approaches.

However, children require a different type of care than elderly people, and robots may be a good platform to fulfill these unique needs.^{63,64} Diseases, such as cancer and diabetes, can disrupt a child's normal life, affect their social needs, and pose various challenges during the treatment of illness and change in lifestyle, which in turn also affect the mental health of the child.^{65,66} Social robots may prove to be helpful in the management of chronic illnesses because they can provide effective support through encouragement, education to adhere to healthy behaviors, and entertainment useful to mitigate stress thereby providing a sense of comfort and companionship.⁶⁷ As such, robots have been shown to be helpful for treating various conditions,⁶⁸ and it is reasonable to expect that they can have similar effects on the treatment of obesity, helping children emotionally and motivating them and subsequently leading them to a healthier lifestyle. While ML helps screening datasets to identify children at risk and make predictions about their condition, AI and robotics could be used to develop user-centric strategies based on the real-time customization of the treatment program.

Conclusion

Although almost all the studies analyzed were statistically sound and showed positive outcomes, our review shows that this area still lacks the use of robots and AI-based technologies.

A more efficacious intervention should be developed in the future by combining a more diverse set of techniques. Studies should focus on the positive aspects of each category and develop an integrated system where there is a physical entity, such as a robot with AI and ML capabilities, that can conduct a physical activity program for the participants and educate them regarding good eating habits. These robots can be implemented with adaptive and real-time feedback techniques, making it easier to monitor the progress in real-time from home. Most of the studies included in the present review showed positive statistical outcomes. Exergaming is the most popular strategy among the literature published in the past 10 years. Thus, our review indicates that technology-based interventions have great potential in terms of efficacy in monitoring and managing obesity in children and adolescents, but its potential has not been reached. This area of research still lacks AI and robot-based interventions, which could be combined with exergaming and ML techniques to maximize their effectiveness and establish a strategic synergy to promote children's health. As for the future direction, we are currently at the stage of introducing an adaptable intelligent robot that can personalize the extent of interaction based on the child's personal interest and level of obesity. Such personalization could enhance the benefits of using robots in the management of childhood obesity.

Funding

The authors would also like to acknowledge the financial support for this work from the Deanship of Scientific Research (DSR) university of Tabuk, Tabuk, Saudi Arabia, under grant no. [1440-258].

Disclosure

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Wolfenden L, Ezzati M, Larijani B, Dietz W. The challenge for global health systems in preventing and managing obesity. *Obesity Rev*. 2019;20:185–193.
2. World Health Organization. Facts and figures on childhood obesity. Available from: <https://www.who.int/end-childhood-obesity/facts/en/>. Accessed March 2021.
3. Sahoo K, Sahoo B, Choudhury AK, et al. Childhood obesity: causes and consequences. *J Fam Med Prim Care*. 2015;4:187–192. doi:10.4103/2249-4863.154628
4. Spear BA, Barlow SE, Ervin C, et al. Recommendations for treatment of child and adolescent overweight and obesity. *Pediatrics*. 2007;120(Suppl 4):S254–S288. doi:10.1542/peds.2007-2329F
5. Prangthip, P., Soe, Y. M., & Signar, J. F. (2021). Literature review: nutritional factors influencing academic achievement in school-age children. *International Journal of Adolescent Medicine and Health*, 33(2). Available from: https://www.rch.org.au/kidsinfo/fact_sheets/Nutrition_older_children/. Accessed April 17, 2022.
6. Simmonds M, Burch J, Llewellyn A, et al. The use of measures of obesity in childhood for predicting obesity and the development of obesity-related diseases in adulthood: a systematic review and meta-analysis. *Health Technol Assess (Rockv)*. 2015;19:1–336. doi:10.3310/hta19430
7. Shields M, Tremblay MS, Connor Gorber S, Janssen I. Abdominal obesity and cardiovascular disease risk factors within body mass index categories. *Health Rep*. 2012;23:7–15.
8. Vucenik I, Stains JP. Obesity and cancer risk: evidence, mechanisms, and recommendations. *Ann NY Acad Sci*. 2012;1271:37–43. doi:10.1111/j.1749-6632.2012.06750.x
9. Hoey H. Management of obesity in children differs from that of adults. *Proc Nutr Soc*. 2014;73:519–525. doi:10.1017/S0029665114000652
10. Jones M, Luce KH, Osborne MI, et al. Randomized, controlled trial of an internet-facilitated intervention for reducing binge eating and overweight in adolescents. *Pediatrics*. 2008;121:453–462. doi:10.1542/peds.2007-1173
11. Childhood and Adolescent Obesity. Available from: <https://asmbs.org/patients/adolescent-obesity>. Accessed March 2021.
12. Slater A, Tiggemann M. “Uncool to do sport”: a focus group study of adolescent girls’ reasons for withdrawing from physical activity. *Psychol Sport Exerc*. 2010;11:619–626. doi:10.1016/j.psychsport.2010.07.006
13. Hammersley ML, Okely AD, Batterham MJ, Jones RA. An internet-based childhood obesity prevention program (Time2bHealthy) for parents of preschool-aged children: randomized controlled trial. *J Med Internet Res*. 2019;21:e11964.
14. Li JS, Barnett TA, Goodman E, et al. Approaches to the prevention and management of childhood obesity: the role of social networks and the use of social media and related electronic technologies: a scientific statement from the American Heart Association. *Circulation*. 2013;127:260–267. doi:10.1161/CIR.0b013e3182756d8e
15. Schulz L, Spil AAMT, de Vries SAS. Changing behavior of kids with obesity with gamified wearables. In: Wickramasinghe N, Bodendorf F, editors. *Delivering Superior Health and Wellness Management with IoT and Analytics*. Springer Nature Switzerland AG. Springer International Publishing; 2020: 145–169.
16. Pretlow RA. Addiction to highly pleasurable food as a cause of the childhood obesity epidemic: a qualitative internet study. *Eat Disord*. 2011;19:295–307. doi:10.1080/10640266.2011.584803
17. Datta AK, Sumargo A, Jackson V, Dey PP. mCHOIS: an application of mobile technology for childhood obesity surveillance. *Procedia Comput Sci*. 2011;5:653–660. doi:10.1016/j.procs.2011.07.085
18. Daley AJ. Can exergaming contribute to improving physical activity levels and health outcomes in children? *Pediatrics*. 2009;124:763–771. doi:10.1542/peds.2008-2357
19. Radesky JS, Schumacher J, Zuckerman B. Mobile and interactive media use by young children: the good, the bad, and the unknown. *Pediatrics*. 2015;135:1–3. doi:10.1542/peds.2014-2251
20. Miller GF, Sliwa S, Michael S, et al. Evaluation of Let’s move! Active schools activation grants. *Prev Med*. 2018;108:36–40. doi:10.1016/j.ypmed.2017.12.024
21. Schuman AJ. Improving patient care: smartphones and mobile medical devices: mobile medical gadgetry is in its infancy, yet these devices can help you motivate pediatric patients and their parents to adopt healthier lifestyles. *Contemp Pediatr*. 2013;30:33–38.
22. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2007;2:59–77.
23. Mirzaei G, Adeli A, Adeli H. Imaging and machine learning techniques for diagnosis of Alzheimer’s disease. *Rev Neurosci*. 2016;27:857–870. doi:10.1515/revneuro-2016-0029
24. Donsa K, Spat S, Beck P, et al. Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges. In: Holzinger A, Röcker C, Ziefle M, editors. *Smart Health. Lecture Notes in Computer Science*. Springer, Cham; 2015: 237–260.
25. Gultepe E, Green JP, Nguyen H, et al. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J Am Med Inform Assoc*. 2014;21:315–325. doi:10.1136/amiajnl-2013-001815
26. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216–1219. doi:10.1056/NEJMp1606181
27. Salvatore C, Cerasa A, Battista P, et al. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer’s disease: a machine learning approach. *Front Neurosci*. 2015;9:307. doi:10.3389/fnins.2015.00307
28. Triantafyllidis A, Polychronidou E, Alexiadis A, et al. Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: a systematic review of the literature. *Artif Intell Med*. 2020;104:101844. doi:10.1016/j.artmed.2020.101844
29. Robinson NL, Connolly J, Hides L, Kavanagh DJ. Social robots as treatment agents: pilot randomized controlled trial to deliver a behavior change intervention. *Internet Interv*. 2020;21:100320. doi:10.1016/j.invent.2020.100320
30. Cappuccio ML, Sandoval EB, Mubin O, et al. Can robots make us better humans? *Int J Soc Robotics*. 2020;13:7–22. doi:10.1007/s12369-020-00700-6
31. Sandoval EB, Mubin O. Virtuous robotics: artificial agents and the good life. *Int J Soc Robotics*. 2021;4:54.
32. Chen JL, Wilkosz ME. Efficacy of technology-based interventions for obesity prevention in adolescents: a systematic review. *Adolesc Health Med Ther*. 2014;5:159–170. doi:10.2147/AHMT.S39969

33. McMullan M, Millar R, Woodside JV. A systematic review to assess the effectiveness of technology-based interventions to address obesity in children. *BMC Pediatr.* 2020;20:242. doi:10.1186/s12887-020-02081-1
34. Nouchi R, Kawashima R. Improving cognitive function from children to old age: a systematic review of recent smart ageing intervention studies. *Adv Neurosci.* 2014;2014:1–15. doi:10.1155/2014/235479
35. Lindberg R, Seo J, Laine TH. Enhancing physical education with exergames and wearable technology. *IEEE Trans Learning Technol.* 2016;9:328–341. doi:10.1109/TLT.2016.2556671
36. Yang HJ, Kang JH, Kim OH, et al. Interventions for preventing childhood obesity with smartphones and wearable device: a protocol for a non-randomized controlled trial. *Int J Environ Res Public Health.* 2017;14:184.
37. Argarini R, Herawati L, Irwadi I, et al. Fundamental movement skill approach to combat childhood obesity in Surabaya, Indonesia: potential effects of video games based exercises (Exergaming). *J Talent Dev Excell.* 2020;12:3026–3034.
38. Adamo KB, Rutherford JA, Goldfield GS. Effects of interactive video game cycling on overweight and obese adolescent health. *Appl Physiol Nutr Metab.* 2010;35:805–815. doi:10.1139/H10-078
39. Staiano AE, Beyl RA, Guan W, et al. Home-based exergaming among children with overweight and obesity: a randomized clinical trial. *Pediatr Obes.* 2018;13:724–733. doi:10.1111/ijpo.12438
40. Ruggiero L, Seltzer ED, Dufelmeier D, et al. MyPlate picks: development and initial evaluation of feasibility, acceptability, and impact of an educational exergame to help promote healthy eating and physical activity in children. *Games Health J.* 2020;9:197–207. doi:10.1089/g4h.2019.0056
41. Trost SG, Sundal D, Foster GD, et al. Effects of a pediatric weight management program with and without active video games: a randomized trial. *JAMA Pediatr.* 2014;168:407–413. doi:10.1001/jamapediatrics.2013.3436
42. Staiano AE, Marker AM, Beyl RA, et al. A randomized controlled trial of dance exergaming for exercise training in overweight and obese adolescent girls. *Pediatr Obes.* 2017;12:120–128. doi:10.1111/ijpo.12117
43. Maddison R, Foley L, Ni Mhurchu C, et al. Effects of active video games on body composition: a randomized controlled trial. *Am J Clin Nutr.* 2011;94:156–163. doi:10.3945/ajcn.110.009142
44. Duman F, Kokaçya MH, Doğru E, et al. The role of active video-accompanied exercises in improvement of the obese state in children: a prospective study from Turkey. *J Clin Res Pediatr Endocrinol.* 2016;8:334–340. doi:10.4274/jcrpe.2284
45. Wagener TL, Fedele DA, Mignogna MR, et al. Psychological effects of dance-based group exergaming in obese adolescents. *Pediatr Obes.* 2012;7:e68–e74. doi:10.1111/j.2047-6310.2012.00065.x
46. Coknaz D, Mirzeoglu AD, Atasoy HI, Alkoy S, Coknaz H, Goral K. A digital movement in the world of inactive children: favourable outcomes of playing active video games in a pilot randomized trial. *Eur J Pediatr.* 2019;178(10):1567–1576. doi:10.1007/s00431-019-03457-x
47. Gao Z, Lee JE, Zeng N, Pope ZC, Zhang Y, Li X. Home-based exergaming on preschoolers' energy expenditure, cardiovascular fitness, body mass index and cognitive flexibility: a randomized controlled trial. *J clin med.* 2019;8(10):1745. doi:10.3390/jcm8101745
48. Espinosa-Curiel IE, Pozas-Bogarin EE, Lozano-Salas JL, et al. Nutritional education and promotion of healthy eating behaviors among Mexican children through video games: design and pilot test of FoodRateMaster. *JMIR Serious Games.* 2020;8:e16431. doi:10.2196/16431
49. Nawi AM, Jamaludin FIC. Effect of internet-based intervention on obesity among adolescents in Kuala Lumpur: a school-based cluster randomised trial. *Malays J Med Sci MJMS.* 2015;22:47–56.
50. Del Río NG, González-González CS, Martín-González R, et al. Effects of a gamified educational program in the nutrition of children with obesity. *J Med Syst.* 2019;43:198. doi:10.1007/s10916-019-1293-6
51. Chen JL, Guedes CM, Lung AE. Smartphone-based healthy weight management intervention for Chinese American adolescents: short-term efficacy and factors associated with decreased weight. *J Adolesc Health.* 2019;64:443–449. doi:10.1016/j.jadohealth.2018.08.022
52. Rerksupphaphol L, Rerksupphaphol S. Internet based obesity prevention program for Thai school children-a randomized control trial. *JCDR.* 2017;11(3):SC07. doi:10.7860/JCDR/2017/21423.9368
53. Nyström CD, Sandin S, Henriksson P, et al. Mobile-based intervention intended to stop obesity in preschool-aged children: the MINISTOP randomized controlled trial. *Am J Clin Nutr.* 2017;105:1327–1335.
54. Fiechtner L, Kleinman K, Melly SJ, et al. Effects of proximity to supermarkets on a randomized trial studying interventions for obesity. *Am J Public Health.* 2016;106:557–562. doi:10.2105/AJPH.2015.302986
55. Lingren T, Thaker V, Brady C, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Appl Clin Inform.* 2016;7:693–706. doi:10.4338/ACI-2016-01-RA-0015
56. Ríos-Julián N, Alarcón-Paredes A, Alonso GA, et al. Feasibility of a screening tool for obesity diagnosis in Mexican children from a vulnerable community of Me'Phaa ethnicity in the State of Guerrero, Mexico. *Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE).* IEEE; 2017: 1–6.
57. Fergus P, Hussain A, Hearty J, et al. A machine learning approach to measure and monitor physical activity in children to help fight overweight and obesity. *International Conference on Intelligent Computing; Springer, Cham;* 2015: 676–688.
58. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform.* 2015;6:506. doi:10.4338/ACI-2015-03-RA-0036
59. Singh B, Tawfik H. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. *International Conference on Computational Science; Springer, Cham;* 2020: 523–535.
60. Alotaibi M. A social robotic obesity management and awareness system for children in Saudi Arabia. *Int J Online Eng.* 2018;14(9):45.
61. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients.* 2020;12:2466. doi:10.3390/nu12082466
62. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychol.* 2009;28(6):690. doi:10.1037/a0016136
63. Alnajjar F, Cappuccio M, Renawi A, Mubin O, Loo CK. Personalized robot interventions for autistic children: an automated methodology for attention assessment. *Int J Soc Robotics.* 2021;13(1):67–82. doi:10.1007/s12369-020-00639-8
64. Vogan AA, Alnajjar F, Gochoo M, Khalid S. Robots, AI, and cognitive training in an era of mass age-related cognitive decline: a systematic review. *IEEE Access.* 2020;8:18284–18304. doi:10.1109/ACCESS.2020.2966819
65. Israels T, Challinor J, Howard S, Arora RH. Treating children with cancer worldwide—challenges and interventions. *Pediatrics.* 2015;136(4):607–610. doi:10.1542/peds.2015-0300

66. Wolpert, M., Curtis-Tyler, K., & Edbrooke-Childs, J. (2016). A qualitative exploration of patient and clinician views on patient-reported outcome measures in child mental health and diabetes services. *Administration and Policy in Mental Health and Mental Health Services Research*, 43(3), 309-315. Available from: <https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/diabetes-issues-for-children-and-teenagers#childrens-reactions-to-a-diabetes-diagnosis>. Accessed April 17, 2022.
67. Dawe J, Sutherland C, Barco A, Broadbent E. Can social robots help children in healthcare contexts? A scoping review. *BMJ Paediatr Open*. 2019;3:e000371. doi:10.1136/bmjpo-2018-000371
68. Alahbabi M, Almazroei F, Almarzoqi M, et al. Avatar based interaction therapy: a potential therapeutic approach for children with Autism. 2017 IEEE International Conference on Mechatronics and Automation (ICMA); 2017: 480–484.

Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy

Dovepress

Publish your work in this journal

Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy is an international, peer-reviewed open-access journal committed to the rapid publication of the latest laboratory and clinical findings in the fields of diabetes, metabolic syndrome and obesity research. Original research, review, case reports, hypothesis formation, expert opinion and commentaries are all considered for publication. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/diabetes-metabolic-syndrome-and-obesity-targets-and-therapy-journal>

Support Publication #16 [SP16]

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021),
“Mapping Meaning and Purpose in Human-Robot Teams:
Anthropomorphic Agents in Military Operations”
Philosophical Journal of Conflict and Violence, Vol. V(1), Trivent
Publishing. DOI: 10.22618/TP.PJCV.20215.1.139005.

Open access publication.



Mapping Meaning and Purpose in Human-Robot Teams: Anthropomorphic Agents in Military Operations

Massimiliano L. Cappuccio,¹ Jai C. Galliot,¹ Eduardo B. Sandoval²

¹ Values in Defence and Security Technology group, University of New South Wales, Australia

² Creative Robotics Lab, University of New South Wales, Australia.

Abstract

We spontaneously tend to project animacy and sensitivity to inanimate objects and sometimes we attribute distinctively human features like intelligence, goals, and reasons to certain artificial devices. This phenomenon is called “anthropomorphism” and has been long studied by researchers in human-robot interaction and social-robotics. These studies are particularly important from the perspective of recent developments in military technology, as autonomous systems controlled by AI are expected to play a greater and greater role in the future of warfare. Anthropomorphic effects can play a critical role in tactical operations involving hybrid human-robot teams, where service members and autonomous agents need to quickly coordinate relying almost exclusively on fast, cognitively parsimonious, natural forms of communication. These forms rely importantly on anthropomorphism to allow human soldiers read the behavior of machines in terms of goals and intentions. Understanding the cognitive mechanisms that underpin anthropomorphic attributions is hence potentially crucial to increase the accuracy and efficacy of human-machine interaction in military operations. However, this question is largely philosophical, as numerous models compete in the space of social cognition theory to explain behavior reading and mental states attribution. This paper aims to offer an initial exploration of these mechanisms from a perspective of philosophical psychology and cognitive philosophy, reviewing the theories in social cognition that are most promising to explain anthropomorphism and predict how it can enable and improve natural communication between soldiers and autonomous military technologies.

Keywords

Autonomous artificial agents; hybrid military teams; human-agent interaction; social cognition; social robotics; social psychology; situational awareness; military drone; autonomous vehicle; value-sensitive design; dogfight.

DOI: 10.22618/TP.PJCv.20215.1.139005

The PJCv Journal is published by Trivent Publishing

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC-BY-NC-ND 4.0) license, which permits others to copy or share the article, provided original work is properly cited and that this is not done for commercial purposes. Users may not remix, transform, or build upon the material and may not distribute the modified material (<http://creativecommons.org/licenses/by-nc/4.0/>)

Mapping Meaning and Purpose in Human-Robot Teams: Anthropomorphic Agents in Military Operations

Massimiliano L. Cappuccio,¹ Jai C. Galliot,¹ Eduardo B. Sandoval²

¹ Values in Defence and Security Technology group, University of New South Wales, Australia

² Creative Robotics Lab, University of New South Wales, Australia.

Abstract

We spontaneously tend to project animacy and sensitivity to inanimate objects and sometimes we attribute distinctively human features like intelligence, goals, and reasons to certain artificial devices. This phenomenon is called “anthropomorphism” and has been long studied by researchers in human-robot interaction and social-robotics. These studies are particularly important from the perspective of recent developments in military technology, as autonomous systems controlled by AI are expected to play a greater and greater role in the future of warfare. Anthropomorphic effects can play a critical role in tactical operations involving hybrid human-robot teams, where service members and autonomous agents need to quickly coordinate relying almost exclusively on fast, cognitively parsimonious, natural forms of communication. These forms rely importantly on anthropomorphism to allow human soldiers read the behavior of machines in terms of goals and intentions. Understanding the cognitive mechanisms that underpin anthropomorphic attributions is hence potentially crucial to increase the accuracy and efficacy of human-machine interaction in military operations. However, this question is largely philosophical, as numerous models compete in the space of social cognition theory to explain behavior reading and mental states attribution. This paper aims to offer an initial exploration of these mechanisms from a perspective of philosophical psychology and cognitive philosophy, reviewing the theories in social cognition that are most promising to explain anthropomorphism and predict how it can enable and improve natural communication between soldiers and autonomous military technologies.

Keywords

Autonomous artificial agents; hybrid military teams; human-agent interaction; social cognition; social robotics; social psychology; situational awareness; military drone; autonomous vehicle; value-sensitive design; dogfight.

Introduction

In a series of simulations recently run by DARPA as part of their “AlphaDogfight” project, artificial intelligence has systematically outperformed a team of experienced F-16 pilots, with machines winning over humans in several consecutive matches of simulated aerial

confrontation.¹ At the same time, as part of the “Loyal Wingman” project, Boeing, in collaboration with U.S. Air Force and Royal Australian Air Force, is developing a new strain of supersonic autonomous combat drones capable to fly in formation with fifth generation jetfighters controlled by human pilots (F-35), protecting them from incoming attacks and automatically coordinating with both on board computers and pilots to conduct joint missions.² These technological advances raise new pressing questions to the researchers interested in human-computer and human-agent interaction.

Arguably, whether they are engaging in cooperative or competitive activities, human pilots need to be able to “read” the behaviour of the artificial ones in order to decode their goals, understand their tactics, and anticipate their next moves. But does the behaviour of an artificial agent look and “feel” like a human one? Are the decisions taken by a machine interpretable in an equally intuitive and effortless manner? And how can the designers and developers make sure that the behaviours of the machine appropriately display, conceal, or misrepresent the objectives pursued by the AI, to achieve the constructor’s intended goals? These questions reach beyond autonomous jetfighters, as they involve the whole rapidly-developing domain of autonomous weapon systems.

Military analysts and foresight experts predict that the future of warfare will be shaped by the armed forces’ capability to incorporate *autonomous artificial agents* (AAAs), i.e. synthetic systems controlled by *artificial intelligence* (AI) such as interactive robots, unmanned vehicles, and digital avatars.³ Whether AAAs operate in physical or merely simulated spaces, they are controlled by adaptive algorithms and machine learning systems capable to skilfully navigate and manipulate the environment, coping with contingencies in completing tasks, and even selecting the best task-resolution strategies.⁴

AAAs are normally classified according to varying degrees of autonomy: “from complete teleoperation, to a highly self-sufficient system that need only be supervised at the task level.”⁵ To comply with mission parameters and accomplish their assigned task, AAAs should be capable to autonomously set and update their local objectives and dynamically adjust their purposeful action strategies to achieve them. To do this, AAAs would need, at least, minimal understanding of the implications of their own actions. Currently, AAAs lack the cognitive prerequisites necessary to recognize relevant contextual contingencies, make sense of complex environments, and reliably interpret human priorities. AI theorists typically refer to the so called “frame problem” when they address the AI’s inherent limitations in representing the contextual background of their own decisions.⁶

¹ P. Tucker, “An AI Just Beat a Human F-16 Pilot in a Dogfight — Again,” *Defense One online* (August 20, 2020). Retrieved at <https://www.defenseone.com/technology/2020/08/ai-just-beat-human-f-16-pilot-dogfight-again/167872/>.

² B. Lendon, “Australian military gets first drone that can fly with artificial intelligence,” *CNN* (2020). Retrieved at <https://edition.cnn.com/2020/05/05/asia/australia-loyal-wingman-drone-intl-hnk/index.html>.

³ See P.W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (London: Penguin, 2009); P. Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: WW Norton & Company, 2018); A. Swiderska, Dennis Küster, “Avatars in Pain: Visible Harm Enhances Mind Perception in Humans and Robots,” *Perception* 47/12 (2018).

⁴ M.L. Cummings, “Artificial Intelligence and the Future of Warfare,” *International Security Department and US and the Americas Programme*, Chatham House - The Royal Institute of International Affairs (Jan. 2017).

⁵ C. Breazeal, “Towards Sociable Robots,” *Robotics and Autonomous Systems* 42/3-4 (2003): 167-175.

⁶ M.L. Cappuccio, M. Wheeler, “When the Twain Meet: Could the Study of Mind Be a Meeting of Minds?,” in *Postanalytic and Metacontinental: Crossing Philosophical Divides*, ed. Jack Reynolds, James Chase, Ed Mares, James Williams (London: Continuum, 2010), 125–44 ; “Ground-Level Intelligence: Inter-

Because of these limitations, at this stage AAAs can only complement human work in military context, not replace it. In general, state-of-the-art AAAs are efficacious only insofar as human operators complement the decision-making protocols embedded in the machines with their own understanding of the mission parameters.⁷ Military research aims to create and manage hybrid systems composed of humans and AAAs, but their integration opens unprecedented challenges.⁸

Reliable, agile, and intuitive *human-agent interaction* (HAI) is a fundamental prerequisite for the satisfactory functioning of hybrid teams.⁹ But sophisticated algorithms are not enough to establish effective HAI.¹⁰ To assure that humans and machines interface correctly and coordinate efficaciously on the battlefield, AAAs must first of all be capable to detect and correctly interpret human intentions, goals, and behaviours; not less important, in order to enable fluent interactions within teams and avoid potentially risky misunderstandings, AAAs must be able to clearly communicate the goals and the reasons of their actions to human operators.¹¹

Hence, in order to be integrated within hybrid military teams, AAAs must be equipped with the capability to recognize other agents and be recognized as agents in turn: not only must they be able to identify *meaning and purpose* in human behaviour, but also produce behaviours that appear immediately *purposeful*, and hence *meaningful*, to humans.¹² Humans and machines must be able to validate each other's decisions through recurrent reciprocal recognition in order to establish even simple coordination and prevent misalignments of intentions and goals that could lead to catastrophic consequences in the context of military operations.

As the operability of AAAs in warfare depends on both the humans and the machines involved, questions concerning the psychological significance and normative value of HAI are not less important than the technological details of the implementation.¹³ Two questions concerning the preparedness of human personnel and the user-centred design of autonomous systems are becoming central in the domain of military HAI:

- (a) How do military personnel experience their interaction with artificial autonomous agents in defence and security operations?

Context Frame Problem and Dynamics of the Background,” in *Knowing without Thinking, Mind, Action, Cognition and the Phenomenon of the Background*, ed. Z. Radman (Basingstoke: Palgrave Macmillan, 2012).

⁷ G. Hoffman, C. Breazeal, “Collaboration in Human-Robot Teams,” in *ALAA 1st Intelligent Systems Technical Conference*, Infotech@Aerospace Conferences (American Institute of Aeronautics and Astronautics, 2004).

⁸ S.G. Hill, D. Barber, A. W. Evans, “Achieving the Vision of Effective Soldier-Robot Teaming,” *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts* (2015).

⁹ M.J. Barnes, A.W. Evans, “Soldier-Robot Teaming: An Overview,” in *Human-Robot Interactions in Future Military Operations*, ed. M. Barnes and F. Jentsch (Farnham: Ashgate Publishing, 2010) ; M.J. Barnes, F.J. Jentsch (eds.), *Human-Robot Interactions in Future Military Operations* (Farnham: Ashgate Publishing, 2010).

¹⁰ K. Dautenhahn, “Socially Intelligent Robots: Dimensions of Human-Robot Interaction,” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (2007): 679–704.

¹¹ G. Hoffman, “Embodied Cognition for Autonomous Interactive Robots,” *Topics in Cognitive Science* 4/4 (2012): 759–72.

¹² M.L. Cappuccio, A. Peeters, W. McDonald, “Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition,” *Philosophy & Technology* 33 (2019): 9–31.

¹³ L. Foster Thompson, D.J. Gillan, “Social Factors in Human-Robot Interaction,” in *Human-Robot Interactions in Future Military Operations*, ed. Barnes and Jentsch (Boca Raton : CRC Press, 2017), 67-81.

- (b) How does the design of human-machine interfaces affect the capability of a hybrid military team to successfully carry out its tasks in a collaborative fashion?

The empirical investigation necessary to answer these questions has just started and is still at an early stage. The goal of our paper is to philosophically address the theoretical and methodological foundations of such investigations, clarifying its assumptions and conceptual background.

As a preparatory step, it is useful to transfer to the military context a number of canonical concepts - such as *Trust*, *Acceptance*, and *Tolerance* - that are key for the methodology of social robotics and HAI research:¹⁴ Trust indicates the expectations related to the robot's intelligence and general ability to competently carry out its tasks;¹⁵ Acceptance measures how likeable and credible a robot can be as a social companion and a co-worker;¹⁶ Tolerance refers to the easiness of transferring to technology the authority and responsibility of decisions traditionally made only by humans.¹⁷ While inherently qualitative in nature, these psychological aspects of HAI must be carefully considered by engineers if they want to design efficacious applications of AI and robotics in warfare. The way AAAs are perceived by human troops makes a difference to what they can do together as part of an hybrid team, as trust, acceptance, and tolerance contribute to define the function and the scope of their interaction.

All these notions converge on a more fundamental phenomenon, called *anthropomorphism*. It is well documented that anthropomorphic tendencies emerge spontaneously and pervasively from HAI.¹⁸ While monitoring these tendencies is indispensable, entirely suppressing them would be neither possible nor convenient due to their pervasiveness and resilience.

This paper aims to investigate why military AAAs are often attributed anthropomorphic features and how anthropomorphism affects HAI in military teams, generating complex patterns of opportunities and risks. Unlike similar works on the same topic, our primary aim is to identify the philosophical prerequisites of a legitimate theory of anthropomorphism. To achieve this goal, first we will compare several approaches to *social cognition*, the discipline that studies the cognitive preconditions of social engagement. Then we will examine how these approaches have been integrated within *social robotics*, the interdisciplinary study of the psychological and technological factors that affect HAI.¹⁹ Finally, we will discuss how the notions and methods of social robotics research are relevant to advance the research in military AAAs.

¹⁴ C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics* 1/1 (2009): 71–81.

¹⁵ H.A. Abbass, J. Scholz, D.J. Reid, "Foundations of Trusted Autonomy: An Introduction," in *Foundations of Trusted Autonomy*, ed. Hussein A. Abbass, Jason Scholz, Darryn J. Reid (New York: Springer, 2018), 1–12.

¹⁶ C. Breazeal, K. Dautenhahn, T. Kanda, "Social Robotics," In *Springer Handbook of Robotics*, ed. Bruno Siciliano, Oussama Khatib (New York: Springer, 2016), 1935–1972.

¹⁷ J.C. Galliot, "The Soldier's Tolerance for Autonomous Systems," *Paladyn, Journal of Behavioral Robotics* 9/1 (2018): 124–36.

¹⁸ J. Reichenbach, C. Bartneck, J. Carpenter, "Well Done, Robot! The Importance of Praise and Presence in Human-Robot Collaboration," in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 86–90 (New Jersey: IEEE, 2006).

¹⁹ B. Mutlu, C. Bartneck, J. Ham, V. Evers, T. Kanda, (eds), *Social Robotics: Third International Conference on Social Robotics, ICSR 2011, Amsterdam, The Netherlands, November 24-25, 2011. Proceedings* (New York: Springer, 2011).

Defense and security operations conducted by hybrid teams have offered some of the most striking and interesting case studies in the scientific literature on anthropomorphism,²⁰ which is why civilian research in social robotics and social cognition is following closely the development of military applications of HAI. On the other hand, the outcomes of the research on anthropomorphism conducted in the fields of social cognition and social robotics should be taken into account by military research for at least two reasons: not only because ignoring the dangerous effects of anthropomorphism can lead to increased risks during hybrid military operations,²¹ but also because carefully considering the natural anthropomorphic tendencies of humans may disclose new advantageous ways of integrating AAAs in military operations.

I. Military Human-Robot Teams: the Background

AI and robotics evolve rapidly and promise to deeply affect the way military work is organized, influencing the allocation of the relevant resources both within and amongst teams. In principle, various security and defence-related tasks can be assigned to, or conducted, with the support of AAAs: intelligence collection, surveillance and reconnaissance; infiltration and loitering for spying and sabotage; aerial assault and selective target engagement; perimeter protection; active shielding of military personnel and civilians; bomb disposal; the handling of chemical, biological and nuclear hazards; logistics and transportation; and the provision of medical and psychological assistance to military personnel, including training, team building, counselling, and conflict mediation.

The advantages offered by AAAs revolve around the elimination or mitigation of the colloquial ‘dull, dirty and dangerous’: reduction of the number of humans and work hours necessary to carry out a mission; minimization of the risks to human safety and integrity; a decrease to the physical, cognitive/attentional, and emotional/motivational burden necessary to carry out certain jobs, with obvious benefits for the troops’ morale and cohesion; improvement in the integration of information representation, which facilitates the coordination of groups in complex scenarios; portability and exact replicability of expertise and decision-making protocols, which enables the continuation of missions despite the absence of specialized personnel. Also, as unmanned vehicles don’t require life-support, they offer unprecedented opportunities for force projection in domains hardly-reachable by humans (as in underwater or stratospheric missions), extending in space and time the military’s presence and its potential for agency, knowledge, and communication through physical or merely digital proxies.²²

While autonomy is the quintessential property of AAAs, these systems can operate fully autonomously only in perfectly controlled environments, i.e. arbitrarily circumscribed

²⁰ J. Carpenter, “Just Doesn’t Look Right: Exploring the Impact of Humanoid Robot Integration into Explosive Ordnance Disposal Teams,” in *Handbook of Research on Technoself: Identity in a Technological Society*, edited by Rocci Luppigini (Hershey: IGI Global, 2013), 609–36.

²¹ J. Carpenter, “Just Doesn’t Look Right: Exploring the Impact of Humanoid Robot Integration into Explosive Ordnance Disposal Teams,” in *Handbook of Research on Technoself: Identity in a Technological Society*, edited by Rocci Luppigini (Hershey: IGI Global, 2013), 609–36; P. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*; J.C. Galliot, “Defending Australia in the digital age: toward full spectrum defence,” *Defence Studies* 16 (2016): 157 – 175.

²² J.C. Galliot, “Unmanned Systems and War’s End: Prospects for Lasting Peace,” *Dynamiques Internationales* 8 (2013): 1 – 24; *Military Robots: Mapping the Moral Landscape* (New York: Routledge, 2015); “Defending Australia in the digital age: toward full spectrum defence,” *Defence Studies* 16 (2016): 157 – 175 ; “The Soldier’s Tolerance for Autonomous Systems,” *Paladyn, Journal of Behavioral Robotics* 9/1 (2018): 124–36.

artificial settings involving small numbers of degrees of freedom and next-to-zero unpredictable contingencies.²³ Such controlled environments are rarely found in unmoderated real-life circumstances, and virtually never on a battlefield. That is why, in most real-life contexts, military AAAs are still unable to operate without human supervision. Their deployment, routine use, maintenance, fine-tuning, and customization necessarily involves the accompanying presence and directives of human operators. The expectation is that AAAs could soon be ready to participate in military missions by (partly) substituting, complementing, augmenting, or coordinating human work, whether physical or intellectual in nature, depending on their function and the kind of technology involved.

The optimization of AAAs' interactive protocols is a key factor to achieve the kind of deep and fluent integration of human and artificial work that is required to effectively employ hybrid teams in defence and security operations. To this end, the pseudo-social abilities of physical robots, unmanned vehicles, and digital avatars have been expanded to develop a capability to jointly act with humans. Such abilities include - among other things - better methods for identifying humans (e.g., face and voice recognition), interpreting their intentions, and anticipating their actions.²⁴

Not only are social abilities indispensable to assistive applications of robotics behind the frontline, such as intelligence, logistics and training, but also to hybrid military operations on the frontline. The defense forces of the United Kingdom and Australia have recently conducted a large-scale exercise as part of a concerted effort undertaken by the 'Five Eyes' (FVEYs) intelligence community. This exercise, called "Autonomous Warrior," was carried out in Jervis Bay, New South Wales, in June-July 2018 and in Salisbury Plain, south of England, in January 2019. Unclassified information about the logistical preparation, scale, and general structure of these exercises confirms the crucial importance attributed by military commands to hybrid teams.²⁵ In the exercise performed by the British Armed forces, more than 50 unmanned systems from industry were deployed to complete similar tasks.²⁶

As expected, some of the challenges faced in the attempt to achieve optimal coordination between humans and machines were technical in nature: for example, to "prevent jamming and disruptive interference," when a high number of autonomous systems were simultaneously communicating with one another and the human operators, and ensuring that "the control systems do not interfere with each other" due to the "increasingly crowded radiofrequency spectrum, especially as several unmanned systems jostle for space to beam back high-resolution data."²⁷

But the greatest challenges are arguably those involving the quality of human-machine communication. Normally, machines are designed to communicate using formal protocols to share information encoded in a symbolic language. This significantly constrains any possibility to establish natural, fluid, semantically rich, and effortless interaction with humans, engendering critical problems in the space of tactical operations. The design of AAA informed by human-centred principles can offer a solution to this problem.

²³ M.L. Cappuccio, M. Wheeler, "When the Twain Meet: Could the Study of Mind Be a Meeting of Minds?" in *Postanalytic and Metacontinental: Crossing Philosophical Divides*, ed. Jack Reynolds, James Chase, Ed Mares, James Williams (London: Continuum, 2010), 125–44.

²⁴ K.A. Cosenzo, M.J. Barnes, "Human-Robot Interaction Research for Current and Future Military Applications: From the Laboratory to the Field," *Unmanned Systems Technology XII* (2010).

²⁵ G. Slocombe, "Autonomous Warrior: Major Air, Land and Sea Exercise," *Asia-Pacific Defence Reporter* (September 26, 2019).

²⁶ G. Turnbull, "Why the British army tested robots in muddy fields," *C4ISRnet* (January 11, 2019).

²⁷ Ibid.

That is why the Australian armed forces created and tested software to better manage the so-called “fog of war”. The software advises the units involved in complex operations “using virtual human characters”²⁸ to provide real-time information and increase situational awareness. Interestingly, the software uses humanoid digital avatars that speak and act in a simulated environment. The choice to give an anthropomorphic look to this AAAs is significant. Other, more impersonal, options would have been immediately available and easier to implement - for example, representing information in a detached and neutral format that includes only text and graphs. Evidently, the designers found no reason to worry that the avatar’s humanoid look could represent a distraction or a source of risky behaviours (in the hypothetical case that troops stopped seeing the system as a tool and start treating it as a person).

In fact, despite these potential complications, the developers have deliberately given emphasis to the embodied and narrative character of the AAA, in an attempt to exploit its role as humanoid informer, coordinator, and supervisor. This choice aims to make communication with human agents more spontaneous and natural, overcoming the problematic lack of fluency and richness in HAI by exploiting the advantageous effects of anthropomorphism during the design stage.²⁹

II. Anthropomorphism: the Theory

The attention to the anthropomorphic design of military AAAs confirms that the effects of anthropomorphism are taken seriously by the military researchers, who seem interested in how the warfighter’s attentional, behavioural, and emotional responses are modulated by natural interaction with AAAs during hybrid operations. That is why it is useful to briefly introduce the notion of anthropomorphism and discuss its practical implications.

There are at least two different definitions of anthropomorphism.³⁰ According to the most common definition, anthropomorphism is the tendency to attribute to non-human entities the characteristics of human ones, transferring to the former psychological attitudes and forms of engagement that are typically reserved to the latter. Zoomorphism is a phenomenon closely related to, and partly overlapping with, anthropomorphism, as it identifies the tendency to attribute to inanimate entities the features of animals or animal-like entities (including the homo sapiens animal species). Among the characteristics that can be projected onto a non-human, possibly inanimate, entity there are both basic features (goal-oriented, adaptive behaviour and the capability to selectively sense and track contextually relevant cues) and more cognitively complex features like knowledge states (beliefs), objectives (desires), and plans to achieve them (intentions), which arguably involve propositional and semantic content.

Social robotics research has highlighted that behavioural anthropomorphism correlates to higher levels of likeability and perceived intelligence of robot, which, in turn, predict more resilient trust and robust interest. Note that different groups of humans (for example autistic and typically developing children) show significantly different preferences about the kinds of

²⁸ G. Slocombe, “Autonomous Warrior: Major Air, Land and Sea Exercise.”

²⁹ K.A. Cosenzo, M.J. Barnes, “Human-Robot Interaction Research for Current and Future Military Applications: From the Laboratory to the Field.”

³⁰ J.A. Zlotowski, H. Sumioka, C. Bartneck, S. Nishio, H. Ishiguro, “Understanding Anthropomorphism: Anthropomorphism Is Not a Reverse Process of Dehumanization,” in *Social Robotics*, ed. Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Samk GeKenji Suzukik John-John Cabibihan, Friederike Eyssel, Hongsheng He (NY: Springer International Publishing, 2017), 618–27.

robots with which they like to engage.³¹ For some humans it is reasonable to have a robot as a companion or even as an intimate mate, while for others the very idea of treating a robot like a person sounds disturbing. Such individual variations should be accounted for in the development of hybrid military teams, as AAAs should ideally be able to detect them and adjust their behaviour accordingly.

The cognitive/behavioural mechanisms underpinning the anthropomorphic experience are very deep, developmentally early, most probably innate, and related to the instinctive tendency to attribute animacy and intentions even to entities that have little or no resemblance at all to animated or living creatures. The fact that anthropomorphism is hardwired in human cognition is compatible with the conjecture that it scaffolds some of our skills in social and cultural cognition. That is why various cognitive biases and delusions arise from a background of anthropomorphic attributions and personification.³²

Anthropomorphism is not always intentional and explicit. Rather, it often emerges at a pre-reflective level, embedded in a set of preferences or subconscious behavioural dispositions. That is why researchers emphasize that the phenomenon of anthropomorphism can be either implicit or explicit: it is *implicit* when a subject pre-reflectively responds to an inanimate entity as if it was animated, attesting with their behaviour or conduct the attribution of animacy or agency; we call *explicit* the anthropomorphic attributions expressed through verbal behaviours.³³ The former can operate automatically, at the level of fast, immediate perceptual and habitual responses of which the agent might not be mindful of, while the latter operates through conscious, reflective control, at the level of judgement and rational evaluation. Explicit anthropomorphism may arise from the deliberate search of human-like features - for example, as part of the process of personification of nature that is at the root of figurative or poetic creation.

Explicit and implicit attributions are semi-independent and not necessarily consistent with each another: it is well possible to immediately anthropomorphize a robot at the perceptual level (for example when its movement show zoomorphic biological and features) while holding the belief that robots are just unintelligent inanimate entities; on the other hand, it is possible to judge a software as truly intelligent and autonomous while feeling completely detached from its immediate behavioural expressions.

It would be impossible to make sense of our complex relationship with AAAs without considering the fundamental distinction between implicit and explicit anthropomorphism, and their semi-independent levels of cognitive elaboration. The correct identification of these different levels and the definition of their specific function within an overarching cognitive architecture is one of the goals of social cognition.

III. Anthropomorphism: the Underlying Psychological Mechanisms

Research in social cognition primarily focuses on the mechanisms and processes that enable the identification of other cognitive agents as cognitive agents, and the interpretation of their behaviours and mental states. Anthropomorphism is a pivotal issue for social cognition

³¹ A. Waytz, J.T. Cacioppo, N. Epley, "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism," *Perspectives on Psychological Science* 5/3 (2014): 219–232 ; G. AIRENTI, "The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind," *Frontiers in Psychology* 9 (November 2018): 2136.

³² M. Dacey, "Anthropomorphism as cognitive bias," *Philosophy of Science* 84/5 (2017):1152-1164.

³³ J.A. Zlotowski, H. Sumioka, F. Eyssel, S. Nishio, C. Bartneck, H. Ishiguro, "Model of Dual Anthropomorphism: The Relationship Between the Media Equation Effect and Implicit Anthropomorphism," *International Journal of Social Robotics* 10 (2018): 701–714.

research because the perception of anthropomorphic features strongly predicts social responses, and social engagement in turn requires at least some minimal form of anthropomorphic perception.³⁴ Perceiving an agent as anthropomorphic is a necessary non-sufficient condition for conducting meaningful social interaction with it, as engagement can acquire a truly “social” colour only when it reciprocally connects agents that are (or seem) alive.

Numerous explanations of anthropomorphism have been provided by the main theories that compete in the arena of social cognition.³⁵ It is useful to distinguish here between *cognitivist* and *embodied* approaches to social cognition, and their different takes on anthropomorphism.³⁶ The cognitivist approaches intend social cognition as an inferential and representational process, mostly based on the manipulations of internal models of the others’ mental states in accord with abstract rules or a priori principles; the embodied approaches understand social cognition as know-how, i.e. the ability to engage social agents through perception and action for establishing meaningful patterns of responses based on analogy or reciprocity.

Cognitivist approaches like *Theory of Mind* (ToM) build on inferentialist and representational principles. According to ToM, social interaction is effective when we are able to guess the other’s mental contents. This guessing is possible because we use folk psychological knowledge to generate a meta-representational model of the other agent’s internal representational states. The proponents of TOM believe that such models can infer when and in what form a certain mental state arises using general logical correlations between mental states and the bodily behaviours that cause them.³⁷

Like ToM, also *Intentional Stance Theory* (IST) is a cognitivist approach to social cognition that explains anthropomorphism as attribution of high-level mental states like beliefs, intentions, and desires to an agent on the basis of a rational inference to the best explanation: the mental state to be attributed to the agent is the one that best fits the circumstances of the perceived action according to a rational model of optimality and efficiency. On the other hand, the *Teleological Stance Theory* (TST) is a deflationary version of IST. While not explicitly embodied, this approach tries to do without the cognitivist assumptions of its counterpart: TST presupposes that the inference to the best explanation operates at a level more fundamental than propositional knowledge, as it replaces the triangular correlations between desires, intentions, and beliefs presupposed by IST with an equivalent set of correlations between perceived goals, actions, and environmental conditions, respectively.³⁸

At the opposite side of the spectrum, the *theories of empathy* understand anthropomorphism as an embodied process based on internal simulation and projective processes of bestowal/attribution: familiar psychophysical correlations that belong to one’s own

³⁴ G. Airenti, “The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind.”

³⁵ A. Waytz, N. Epley, J.T. Cacioppo, “Social cognition unbound: Insights into anthropomorphism and dehumanization,” *Current Directions in Psychological Science* 19/1 (2010) :58–62 ; E.G. Urquiza-Haas, K. Kotrschal, “The Mind behind Anthropomorphic Thinking: Attribution of Mental States to Other Species,” *Animal Behaviour* 109 (November 2015): 167–76.

³⁶ S. Gallagher, “Understanding Others: Embodied Social Cognition,” in *Handbook of Cognitive Science*, ed. Paco Calvo and Antoni Gomila (San Diego: Elsevier, 2008), 437–52 ; “Inference or Interaction: Social Cognition without Precursors,” *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 11/3 (2008): 163–74.

³⁷ G. Airenti, “The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind.”

³⁸ M.L. Cappuccio, “Inference or Familiarity?,” *Synthesis Philosophica* 58/2 (2015): 253–72.

experiential repertoire are attributed to another on the basis of an analogical process.³⁹ “Empathy” is an English term derived from the German “Einfühlung” (feeling inside or through another’s body).⁴⁰ Empathy is often considered a key component of social expertise and an enabler of altruistic behaviour, but in its original phenomenological formulation it specifically refers to one’s ability of perceiving the other’s body as alive and sensible, based on an analogical pairing (“Paarung”) between the former and the latter.

The idea behind empathy is that every cognitive agent is familiar with the history of its own emotional and intentional states, and has an intuitive knowledge of how those states correspond to specific behavioural expressions, it tends to automatically attribute the same emotions and intentional states to other agents when they display similar behavioural expressions. The very *recognition of other cognitive agents as cognitive agents* is based on an embodied analogy with oneself, as it involves transferring one’s own cognitive and phenomenological states to another’s body. This analogical link explains our capability and tendency to use our experiences of social interaction with conspecifics as standard templates to articulate rich narrations about the non-social world. This is how empathy allows us to assign intelligible meanings to natural events, personifying them.

Interaction theory represents a different embodied approach and positions itself as an alternative or perhaps a supplement to the theories of empathy: from an interactionist perspective, what allows two or more cognitive agents to mutually understand their intentions, align their goals, and recognize one another as cognitive agents is the fact that their behaviour is from the beginning determined by the context of reciprocal interactions in which they meet one another, and by the recurrent patterns of solicitations and responses that are enacted by said agents to make sense of their recurrent interaction.⁴¹

Ecological psychology can be considered a complement to the embodied approaches, as it understands anthropomorphism as a way of reducing the complexity of perceptual contingencies using only a relevant set of social affordances: we make sense of complex scenarios by reading them in terms of landscapes of action opportunities offered to both ourselves and other agents through the spatial environment we share. Seeing the behaviour of an AAA as purposeful and human-like involves the implicit solicitation of the practical know-how that is required for dealing with familiar intersubjective contexts in which social actions are typically afforded by with the help of other agents.

Cognitivist and embodied theories attempt to explain anthropomorphic attributions using different psychological principles. That is why they inevitably focus on different sets of cognitive resources: the latter relies on experiential accounts of others’ behaviours based on familiarity or habit; the former on general inferential principles derived from abstract psychological models (based on either logical deduction or empirical induction) of the other minds.

³⁹ V. Gallese, “The Shared Manifold Hypothesis: Embodied Simulation and Its Role in Empathy and Social Cognition,” in *Empathy in Mental Illness*, ed. Tom F. D. Farrow, Peter W. R. Woodruff, Tom F. D. Farrow, and Peter W. R. Woodruff (Cambridge: Cambridge University Press, 2007), 448–72.

⁴⁰ See G. Airenti, “The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy,” *International Journal of Social Robotics* 7/1 (2015): 117–127; M.A. Harrison, A.E. Hall, “Anthropomorphism, Empathy, and Perceived Communicative Ability Vary with Phylogenetic Relatedness to Humans,” *Journal of Social, Evolutionary & Cultural Psychology* 4/1 (2010): 34; L.D. Rick, T. Rabinowitch, B. Chakrabarti, P. Robinson, “How Anthropomorphism Affects Empathy toward Robots,” in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2009): 245–46.

⁴¹ S. Gallagher, “Inference or Interaction: Social Cognition without Precursors.”; G. Airenti, “The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind.”

The embodied approaches, whether based on interaction or simulation, are better positioned to account for the phenomenon of anthropomorphism as it emerges in the military applications of HAI. As the embodied approaches describe automatic low-level responses that are built in one's pre-reflective perceptual and practical dispositions and don't require explicit cognitive effort, they are able to capture the specificity of fast-paced, emotionally charged, and practically characterized situations, such as the tactical operations in which AAAs are anthropomorphized by service members.

Arguably, while these two groups of theories differ in multiple ways, it is possible to combine them in a pluralistic epistemological account of anthropomorphism. Not only do embodied and cognitivist approaches to social cognition offer different explanations of anthropomorphism; they also describe two distinct levels or manifestations of the same anthropomorphic phenomenon. An explanatory account that integrates these two levels can dissolve the apparent contradiction that characterizes the anthropomorphic experience in the military context: military professionals can and often do attribute anthropomorphic features to AAAs (and, occasionally, they seem able to develop attachment and moral consideration for them), even if they perfectly know, at the rational level, that robots and AIs are just mechanical devices devoid of emotional states, proper motivations, and purposes. Dissolving this paradox will help us make sense of the cases of anthropomorphism documented by the literature on military HAI.

IV. Weaponizing Anthropomorphism: the Risks

Military personnel, even very qualified ones, can become psychologically close to robots when they work together for long periods in conditions of high stress. Service members have in some cases developed anthropomorphic responses and mild forms of attachment for the robots that assisted them in their missions, even though said robots look barely human or animal-like. Namely, explosive ordnance disposal robots were assigned names and gendered identities by the soldiers who worked with them in Iraq and Afghanistan. When one of these robots was damaged, its loss was not simply grieved like the destruction of an expensive piece of equipment, but like the death of a teammate. Singers reports that, when one of these robots was sent to the headquarters for reparation after suffering structural damage, its "teammates" requested that its mechanical parts were not replaced, but accurately fixed to preserve the individual identity of the robot.⁴² In at least one documented case, one soldier risked his own life to save a robot from enemy fire.

Since robots are obviously not sentient, the attachment developed by military personnel to their robots has been typically deemed as an irrational behaviour. As such, it looks incompatible with the notion that military personnel trained for operating AAAs are well-informed about the nature of their artificial partners and act responsibly in accord with strict safety protocols. Accordingly, Carpenter has hypothesized that anthropomorphic tendencies for an AAA (namely, empathic responses including attachment) could lead to counterproductive or even risky behaviours.⁴³ For example, if personnel feel subconsciously compelled to protect the safety of AAAs as if they were actual teammates. Consequently, Carpenter recommends caution in designing ordnance disposal robots, as their anthropomorphic look (especially now that developers are focusing on biped models), might trigger an even stronger attachment in human users.

⁴² Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*.

⁴³ J. Carpenter, "Just Doesn't Look Right: Exploring the Impact of Humanoid Robot Integration into Explosive Ordnance Disposal Teams."

However, our implicit/explicit distinction, drawing a separation between embodied and cognitive levels of social cognition, motivates a different kind of approach. Attachment for robots does not necessarily have to be considered an irrational behaviour. It originates at the level of cognitive processing - the level of embodied, affective intelligence - that is more basic and fundamental than detached, inference-based, rational judgment but does not necessarily compete with it. In fact, embodied responses may well contribute to scaffold the kind of situated intelligence and adaptive responses that constitute the concrete background of practical rationality.

Accordingly, anthropomorphism for military robots should neither be considered a failure of the troops' capabilities of rational evaluation, nor simple as the "metaphorical" attribution of features. The first hypothesis,⁴⁴ fuelled by the belief that anthropomorphism is a dangerous effect of HAI, is wrong because it does not acknowledge that both the implicit and the explicit level of cognition are constitutive components of the normal human capability to critically evaluate complex situations, as often arise in military team operations, which arguably involve both emotional/embodied intelligence and detached judgment. The second hypothesis (epitomized by Barnes & Evans 2010), fuelled by the belief that anthropomorphism is just a derivative, symbolically mediated construction enabled by linguistic/cultural representations, is also wrong because it fails to acknowledge that any explicit process of personification is scaffolded by an implicit anthropomorphic process, and hence is deeply rooted into the human's embodied expertise, emotional engagement, and perceptual responses.

Both views fail to recognize that anthropomorphism is a constitutive component of HAI, not its by-product, because embodied social cognition abilities *are* a fundamental requirement of intelligent decision-making and adaptive behaviour during intersubjective coordination. Hence, the function of anthropomorphism is neither intrinsically negative nor derived from cultural or linguistic representations. While the risks associated with dysfunctional HAI are not to be underestimated, evidence suggesting that anthropomorphism leads to risky behaviours is scarce, merely anecdotal, and cannot justify a general ban of anthropomorphic design. Furthermore, even if inappropriate attachment for robots occurred systematically, its risky effects could be easily countered by appropriate monitoring and training.

The practical advantages offered by a responsible implementation of anthropomorphic features in a military robot's design will often surpass the potential risks when understood along the above-mentioned lines. Anthropomorphic effects are actively exploited by assistive robotics for their positive effects in civilian (e.g., domestic, clinical, educational) contexts.⁴⁵ The differences between civilian and military applications of interactive robots should not be overlooked,⁴⁶ but considering the potential advantages of anthropomorphism for military HAI is more promising than looking only at poorly documented risks. In fact, as we will argue through the rest of the paper, anthropomorphic tendencies can be deliberately elicited and controlled to improve human-machine interface, creating more reliable and efficient communication between troops and robots.

V. Weaponizing Anthropomorphism: the Tactical Applications

Intuitive communication methods to control autonomous and semi-autonomous agents have been tested in the context of military research. To ensure an accurate and intuitive

⁴⁴ See *ibid.*; Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*.

⁴⁵ L. Damiano, P. Dumouchel, "Anthropomorphism in Human-Robot Co-Evolution," *Frontiers in Psychology* 9 (2018): 117.

⁴⁶ J.C. Galliot, "Uninhabited systems in the civilian realm: Some ethical concerns," *IEEE Technology and Society Magazine* 31 (2012): 13-16.

coordination between AAAs and humans, such communication methods can involve embodied cues like hand gestures (which provide an intuitive control method based on iconic semantics⁴⁷), gaze direction (to establish joint attention and facilitate natural action coordination), and emotional expressions (through voice or facial expressions).

Natural communication methods offer significant operational advantages in HAI.⁴⁸ Intuitive control does not require long and demanding training and frees up cognitive and physical resources (memory, attention and bodily engagement) that the human agent should preferably allocate to other tasks, for example communication tasks that require eye contact or bimanual manipulative tasks (like operating a computer).

Multi-tasking can and must be carried out fluidly and efficaciously, but this is possible only if the concurrent tasks don't interfere with one another, competing for the same type of cognitive resources.⁴⁹ That is why cognitive parsimony is particularly important for military operations that require efficient decision-making and prompt reactions to situational changes, with significant demands on working memory and attentional resources.

Intuitive/natural control is advantageous because it favours cognitive parsimony in a specific subset of tactical contexts, that we call *hot* scenarios. To circumscribe this subset, we must distinguish it from *cold* scenarios. We call "hot" those emotionally charged, risky, high-tempo, scenarios that involve direct embodied (perceptual, motoric) engagement, prioritize agency over critical analysis and planning, require quick decisions of a practical nature, readiness to generate short-term predictions, and the ability to fluidly transition from one context of action to another. Cold scenarios, in turn, involve timeframes sufficiently relaxed to enable analytic and critical approaches to problem-solving, detached and disembodied form of control, a spectatorial stance, and low direct risks for the cognitive agent. Hot scenarios can be found in active combat operations, while cold ones correspond to intelligence collection tasks or peacetime activities that involve planning and strategic management and are typically conducted while sitting at a desk or behind a computer.

Anthropomorphism is unlikely to impact significantly on cold scenario (for example, when human operators are required by their role to produce neutral offline judgment and detached scrutiny) because implicit responses are less likely to arise when HAI is not based on direct perceptual engagement dynamics. On the contrary, tactical operations conducted by warfighters typically involve hot scenarios calling for value-laden decisions. When such operations are carried out by hybrid-teams comprised of humans and AAAs, anthropomorphism is a factor to be considered carefully as it might significantly impact on the appropriateness of human decisions and the chances of success of the whole team. This is because the warfighter must rely on their perceptual knowledge to make fast and effective decisions, having been trained to navigate uncertain contingencies using reflexive behaviour.

Team coordination is quick and trusted when it relies on a spontaneous, intuitive two-way signalling activity. Even when coordination among professional operatives relies on strict protocols acquired during training (rules of engagement, codes of communication, and deontological principles that soldiers have internalized), the actual implementation in real-life

⁴⁷ See M. Obaid, F. Kistler, G. Kasparavičiūtė, A.Y. Yantaç, M. Fjeld, "How Would You Gesture Navigate a Drone?: A User-Centered Approach to Control a Drone," in *Proceedings of the 20th International Academic Mindtrek Conference* (New York: Association for Computing Machinery, 2016), 113–21.

⁴⁸ T. Ono, M. Imai, H. Ishiguro, "Anthropomorphic Communications in the Emerging Relationship between Humans and Robots," *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication* (2000).

⁴⁹ J.Y. Chen, S. Quinn, J. Wright, M. Barnes, D. Barber, D. Adams, "Human-Agent Teaming for Robot Management in Multitasking Environments," *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2013).

situations necessarily involves the ability to intuitively detect intentions and emotions in the others (both teammates and adversaries) and quickly adjust to their conduct.

Reading others' behaviours and emotional expressions (pain, fear, impatience, intention to maintain one's position or flee, disposition to open fire or wait) is done automatically and is considered part of a natural set of social abilities, but appropriate training can help channel one's embodied know-how to maximize its efficacy in well-coordinated and planned tactical scenarios. It is of course likely that the natural ability to read behaviour and emotion are hindered by contingent factors: when obfuscated by medical or psychological injuries, when confounded by differences in culture or code of behaviour, when the information from multiple concurrent sources is overwhelming and slows down the process of reciprocal coordination, when the nature of others' intentions is opaque due to deceptive intentions or the intrinsic limits of bodily expressivity.

The last case importantly includes the challenges associated with the interpretation of human bodily actions that use machines or artificial tools as medium. Interpreting the mental state of a soldier in flesh and bone through her gestures and facial expressions is often easier than interpreting the same mental states when the soldier is inside a car, tank, airplane, if the vehicles hides her bodily expressions. Our normal ability to perceive and interpret the others' intentions largely depends on how automatic sensorimotor responses and reciprocal expectations are established through embodied intersubjective coupling, i.e. a learning process in which two or more agents reciprocally re-define their ways to meaningfully act upon and perceive the world.

Anthropomorphism theory must and can explain how inanimate objects can mediate intersubjective coupling, acting as proxies for embodied interaction, when the physical or symbolic properties of the object evoke intentional actions enacted by living agents with their familiar expressive states. When this happens, even simple geometric shapes, with their neutral patterns of bi-dimensional movements, can near-instantaneously convey anger, or fear, or intentions like "fleeing," "capturing," "hiding from" to one or more individuals or others artificial agents operating in a team.

VI. Cars, Drones, and Tanks with an Attitude

An exemplification of this general psychological notion is offered by how different styles of car driving can convey information about the driver's intentions, goals, and emotional states.⁵⁰ All of us, and especially the most expert drivers, can intuitively read another driver's intentions through their car's movements. Knowing that the car does not have intentions of its own does not prevent us from seeing the driver's intention through the car's movements.⁵¹ Anthropomorphic cues are typically not processed as an explicit set of syntactic rules or inferential principles, but as implicit perceptual know-how. As this process is automatic, the observer does not necessarily know how it works, and yet may be able to perfect it and make it more efficacious with appropriate training and expertise development.

Now, if anthropomorphic meanings bleed through the perceived behaviour of cars driven by human drivers, so that a certain driving style can betray the broad horizon of reasons and motivations that characterize a human driver's decisions, then the behaviour of self-driving cars should be anthropomorphized in the same way if it uses the same driving style. The

⁵⁰ R. Sparrow, M. Howard, "When Human Beings Are like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport," *Transportation Research Part C: Emerging Technologies* 80 (July 2017): 206–15.

⁵¹ J.Y. Chen, C. Anthony, R. Selkowitz, K. Stowers, S.G. Lakhmani, M.J. Barnes, "Human-Autonomy Teaming and Agent Transparency," *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017).

behaviour of self-driving cars is very likely to reflect fine-grained patterns of human-like intentional behaviours captured through the statistical analysis of myriads of videoclips of human drivers.⁵²

If, through machine learning technologies and appropriate fine-tuning by human trainers, autonomous vehicles (including cars, tanks, or aircraft) can reproduce the increasing number of patterns from the wide horizon of intentional patterns that characterize the human driving activity, then it is reasonable to assume that human observers, and perhaps also other machines, can detect these patterns and extract useful information to explain the car's behaviour, interpret its causes in terms of motivations and reasons, or anticipate its effects, characterizing them as goals and desires.

When sufficiently fine-grained, this information can be crucial in hot situations, as it allows the observer to quickly adjust to the car's behaviour and coordinate with it. In turn, the complete lack of such information could be misleading for a human observer strongly used to rely on the intentional information communicated by the car's behaviour to adjust its own conduct at the wheel.⁵³ Confusion could arise if autonomous vehicles were programmed to display only psychologically neutral and mechanical behaviours, filtering out the fine-grained details that suggest emotion or intention.

Now let us imagine that human drivers are expected to coordinate with unmanned vehicles to accomplish a certain collaborative task, such as driving in formation to escort a military convoy. Could the dearth of anthropomorphic information expressed by unmanned vehicles disrupt the overall performance of the hybrid team? Or should we rather think that anthropomorphic information is merely aesthetic and cosmetic, but does not play any functional role in human-machine interaction? The empirical evidence necessary to answer this question is sparse, but we can rely on the well-established practical know-how and common sense embedded in the practices and mindset of technology designers. Lay evidence suggests that humans already perceive the intent of low-level autonomous vehicles and rely on their evolving recognition of the relevant patterns in transitioning from a non-autonomous to highly autonomous transport future.

Car designers and marketers also know that vehicles have a "body," a "face," "eyes," and that these elements contribute to shape the perceived personality of the car and hence attract different types of potential buyers.⁵⁴ This suggests that aesthetics and functional design can and often are strongly intertwined. We also know that human-human interaction massively relies on tenuous behavioural cues of which we typically are barely aware of (body language, voice intonation, gaze direction, etc.), and that the accuracy of explanatory/predictive behaviours (whether in cooperative or competitive contexts) strongly depends on the expectations formed by human agents on these cues.

Hence, coordination of complex activities involving manned and unmanned vehicles could be affected by the way anthropomorphic information is or is not conveyed by the unmanned vehicles, and by the reciprocal capability of unmanned vehicles and human operators to correctly interpret each another's anthropomorphic cues in terms of intentional and emotional states. This consideration is even more significant in the context of interactions

⁵² S. Nyholm, "The Ethics of Crashes with Self-Driving Cars: A Roadmap," *Philosophy Compass* 13/7 (2018).

⁵³ Ibid.

⁵⁴ C. DiSalvo, F. Gemperle, "From Seduction to Fulfillment: The Use of Anthropomorphic Form in Design," in *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (2003): 67–72 ; C. DiSalvo, B. Hanington, J. Forlizzi, "An Accessible Framework of Emotional Experiences for New Product Conception," in *Design and Emotion*, ed. Deana McDonagh, Paul Hekkert, Jeroen van Erp, Diane Gyi (Boca Raton: CRC Press, 2003).

between unmanned military vehicles, such as autonomous tanks, underwater vehicles, and flying drones of varying size and role. This importantly includes, as mentioned in this paper's Introduction, competitive interactions with new generation drones excelling in dogfight, in which the drone's ability to produce a certain intentional behaviour and the human pilot's ability to correctly and quickly read that behaviour may eventually decide which party will win the aerial confrontation.

Conclusions

Recognizing and controlling the way military personnel subconsciously project animacy, intentionality, agency, and purpose on AAAs is not just a way to discipline their relationship, but also a necessity for the success of the team as a whole. Anthropomorphizing these devices can be a cognitively parsimonious, operationally efficacious, and properly rational way to make sense of their behaviours and successfully coordinate with them to complete joint tasks. This is especially true of 'hot' (tactical) scenarios that require agents to directly establish interactions of an embodied, perceptual nature and to make quick decisions utilizing person-centred, highly context-sensitive information using informal signals such as emotional and pragmatic cues. Timing is crucial in these scenarios because adjusting to fast changing circumstances requires promptly updated situation awareness, which can only be achieved if all the actors in the network maintain an appropriate communications tempo using informal communicative processes shared with all the other actors.⁵⁵ To sum up, in hybrid warfare the success of both action and decision depends on effective, parsimonious, and timely team communication, which in turn involves anthropomorphic cues.

Informal communication between human operatives and AAAs is both a desirable and realistic outcome and depends on basic mechanisms of social cognition. Hence, artificial agents can be effectively employed in military operations only factoring in both the desirable and the undesired implications of anthropomorphism on a team's constituent parts. Social cognition research suggests that anthropomorphism is a basic feature of human psychology and, as such, it represents an ineliminable component of HAI. That is why social robotics research systematically studies the roots of anthropomorphism to maximize human-robot coordination through anthropomorphic design, taking advantage of how AAAs are generally perceived by humans to improve the quality of the HAI experience.⁵⁶

That is why pro-actively integrating anthropomorphic design motives (instead of accepting them as an accidental byproduct) should be seen as a necessary stage in the roadmap

⁵⁵ The Actor-Network-Theory can be profitably applied to our investigation. The theory considers both humans and AAAs as "actants" whose role defines, and at the same time is defined, by the particular networks in which they operate. The relation between individual actants and network's objectives has to be understood systemically, as their performance can only be assessed against the system's objectives. See S. Soltanzadeh, M. Mooney, "Players Within a Team: Understanding the Structure of Team Performance Through Individual Functions and Team Objectives," *International Journal of Sports Science & Coaching* 5/1 (2018): 84–89. If the hybrid unit's dynamics (leadership, agency, and the cognitive burden of decision) are intelligible it is because they are informed by the roles taken up by human and robotic actants, which in turn are importantly shaped by the phenomenon of anthropomorphism. In the military context, each network represents a unit, in which humans and AAAs cooperate to achieve the same tactical objectives. Unit's effectiveness depends on the HAI's robustness and fluidity. Hence, anthropomorphism can be used to re-design and optimize network relations within the unit.

⁵⁶ B.R. Duffy, "Anthropomorphism and the Social Robot," *Robotics and Autonomous Systems* 42/3 (2003): 177–90; C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics* 1/1 (2009): 71–81.

to the development of efficacious military AAAs. This implies that AAAs design and troops training should be designed in parallel, aiming towards the development of natural interfaces to optimize human-machine communication, as this could significantly reduce the costs and times necessary to make hybrid military teams operation ready. Considering the limitations in the development of generalistic AI, specific studies are required to maximize the benefits of anthropomorphism in task-specific AAAs in accord with their distinctive design requirements.

Funding acknowledgement

This material is based upon work supported by the Air Force Office of Scientific Research and the Minerva Initiative under award number FA9550-18-1-0039. The work has also received the support of the Australian Government.

Disclosure statement

The authors declare that they have no conflict of interest.

References

- Abbass, H.A., Scholz, J., Reid, D.J. “Foundations of Trusted Autonomy: An Introduction.” In *Foundations of Trusted Autonomy*, edited by Hussein A. Abbass, Jason Scholz, Darryn J. Reid, 1-12. New York: Springer, 2018. https://doi.org/10.1007/978-3-319-64816-3_1.
- Airenti, G. “The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy.” *International Journal of Social Robotics* 7/1 (2015): 117–27. <https://doi.org/10.1007/s12369-014-0263-x>.
- . “The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind.” *Frontiers in Psychology* 9 (November 2018): 2136. <https://doi.org/10.3389/fpsyg.2018.02136>.
- Barnes, M., Evans, A.W. “Soldier–Robot Teaming: An Overview.” In *Human-Robot Interactions in Future Military Operations*, edited by M. Barnes and F. Jentsch. Farnham: Ashgate Publishing, 2010.
- Barnes, M.J., Jentsch, F.J. (eds.). *Human-Robot Interactions in Future Military Operations*. Farnham: Ashgate Publishing, 2010. <https://doi.org/10.4324/9781315587622>.
- Bartneck, C., Kulić, D., Croft, E., Zoghbi, S. “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots.” *International Journal of Social Robotics* 1/1 (2009): 71–81. <https://doi.org/10.1007/s12369-008-0001-3>.
- Bartneck, C., Kanda, T., Mubin, O., Al Mahmud, A. “Does the Design of a Robot Influence Its Animacy and Perceived Intelligence?” *International Journal of Social Robotics* 1 (2009): 195–204. <https://doi.org/10.1007/s12369-009-0013-7>.
- Breazeal, Cynthia. “Towards Sociable Robots.” *Robotics and Autonomous Systems* 42/3-4 (2003): 167-175.
- Breazeal, C., Dautenhahn, K., Kanda, T. 2016. “Social Robotics.” *Springer Handbook of Robotics*. https://doi.org/10.1007/978-3-319-32552-1_72.
- Cappuccio, M.L. “Inference or Familiarity?” *Synthesis Philosophica* 58/2 (2015): 253–72.
- Cappuccio, M.L., Peeters, A., McDonald, W. “Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition.” *Philosophy & Technology* 33 (2020) : 9-31. <https://doi.org/10.1007/s13347-019-0341-y>.

- Cappuccio, M.L., Wheeler, M. "When the Twain Meet: Could the Study of Mind Be a Meeting of Minds?" In *Postanalytic and Metacontinental: Crossing Philosophical Divides*, edited by Jack Reynolds, James Chase, Ed Mares, James Williams, 125–44. London: Continuum, 2010.
- Cappuccio, M.L., Wheeler, M. "Ground-Level Intelligence: Inter-Context Frame Problem and Dynamics of the Background." In *Knowing without Thinking. Mind, Action, Cognition and the Phenomenon of the Background*, edited by Zdravko Radman. Basingstoke: Palgrave Macmillan, 2012.
- Carpenter, J., Eliot, M., Schultheis, D. "The Uncanny Valley: Making Human-Nonhuman Distinctions." In *Proceedings of the 5th International Conference on Cognitive Science* (2006): 81–82. <http://www.academia.edu/download/30609409/p81.pdf>.
- Carpenter, J. "Just Doesn't Look Right: Exploring the Impact of Humanoid Robot Integration into Explosive Ordnance Disposal Teams." In *Handbook of Research on Technoself: Identity in a Technological Society*, edited by Rocci Luppigini, 609–36. Hershey: IGI Global, 2013.
- Chen, J.Y., Anthony, C., Selkowitz, R., Stowers, K., Lakhmani, S.G., Barnes, M.J. 2017. "Human-Autonomy Teaming and Agent Transparency." *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. <https://doi.org/10.1145/3029798.3038339>.
- Chen, J.Y., Quinn, S., Wright, J., Barnes, M., Barber, D., Adams, D. "Human-Agent Teaming for Robot Management in Multitasking Environments." *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2013). <https://doi.org/10.1109/hri.2013.6483522>.
- Cosenzo, K.A., Barnes, M. J. "Human-Robot Interaction Research for Current and Future Military Applications: From the Laboratory to the Field." *Unmanned Systems Technology XII* (2010). www.spiedigitallibrary.org/conference-proceedings-of-spice/7692/769204/Human-robot-interaction-research-for-current-and-future-military-applications/10.1117/12.850038.short?casa_token=-v4XTIBdAosAAAAA:mVxNaDQmnNpXpC2NI8Q-rBqfVIsmbBySfb4eEzn5OXR5FUf1Kdk0ADHDde1BEgBGb8jK5G1hyQ.
- Dacey, M. "Anthropomorphism as cognitive bias." *Philosophy of Science* 84:5 (2017): 1152-1164.
- Cummings, M.L. "Artificial Intelligence and the Future of Warfare." *International Security Department and US and the Americas Programme*, Chatham House - The Royal Institute of International Affairs (January 2017).
- Damiano, L., Dumouchel, P. "Anthropomorphism in Human–Robot Co-Evolution." *Frontiers in Psychology* 9 (2018): 117. <https://doi.org/10.3389/fpsyg.2018.00468>.
- Dautenhahn, K. "Socially Intelligent Robots: Dimensions of Human-Robot Interaction." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (2007): 679–704.
- DiSalvo, C., Gemperle, F. "From Seduction to Fulfillment: The Use of Anthropomorphic Form in Design." In *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (2003): 67–72. <https://doi.org/10.1145/782896.782913>.
- DiSalvo, C., Hanington, B., Forlizzi, J. "An Accessible Framework of Emotional Experiences for New Product Conception." In *Design and Emotion*, edited by Deana McDonagh, Paul Hekkert, Jeroen van Erp, Diane Gyi. Boca Raton: CRC Press, 2003.
- Duffy, B.R. "Anthropomorphism and the Social Robot." *Robotics and Autonomous Systems* 42/3 (2003): 177–90. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3).
- Epley, N., Waytz, A., Cacioppo, J. T. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114/4 (2007): 864–86. <https://doi.org/10.1037/0033-295X.114.4.864>.

- Epley, N., Waytz, A., Akalis, S., Cacioppo, J.T. "When We Need A Human: Motivational Determinants of Anthropomorphism." *Social Cognition* 26/2 (2008): 143–55. <https://doi.org/10.1521/soco.2008.26.2.143>.
- Eyssel, F.A., Pfundmair, M. "Predictors of Psychological Anthropomorphization, Mind Perception, and the Fulfillment of Social Needs: A Case Study with a Zoomorphic Robot" *2015 24th IEEE International Symposium on Robot and Human Interactive Communication* (2015). <https://doi.org/10.1109/roman.2015.7333647>.
- Foster Thompson, L., Gillan, D.J. "Social Factors in Human-Robot Interaction." In *Human-Robot Interactions in Future Military Operations*, edited by Barnes and Jentsch, 67-81. Boca Raton: CRC Press, 2017.
- Gallagher, S. "Understanding Others: Embodied Social Cognition." In *Handbook of Cognitive Science*, edited by Paco Calvo and Antoni Gomila, 437–52. San Diego: Elsevier, 2008.
- . "Inference or Interaction: Social Cognition without Precursors." *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 11/3 (2008): 163–74. <https://doi.org/10.1080/13869790802239227>.
- Gallese, V. "The Shared Manifold Hypothesis: Embodied Simulation and its Role in Empathy and Social Cognition." In *Empathy in Mental Illness*, edited by Tom F. D. Farrow, Peter W. R. Woodruff, Tom F. D. Farrow, and Peter W. R. Woodruff, 448–72. Cambridge: Cambridge University Press, 2007. <https://doi.org/10.1017/CBO9780511543753.025>.
- Galliot, J.C. "Uninhabited systems in the civilian realm: Some ethical concerns." *IEEE Technology and Society Magazine* 31 (2012): 13-16. <http://dx.doi.org/10.1109/MTS.2012.2202017>.
- . "Unmanned Systems and War's End: Prospects for Lasting Peace." *Dynamiques Internationales* 8 (2013): 1 – 24.
- . *Military Robots: Mapping the Moral Landscape*. New York: Routledge, 2015.
- . "Defending Australia in the digital age: toward full spectrum defence." *Defence Studies* 16 (2016): 157 – 175. <http://dx.doi.org/10.1080/14702436.2016.1165596>.
- . "The Soldier's Tolerance for Autonomous Systems." *Paladyn, Journal of Behavioral Robotics* 9/1 (2018): 124–36. <https://doi.org/10.1515/pjbr-2018-0008>.
- Harrison, M.A., Hall, A.E. "Anthropomorphism, Empathy, and Perceived Communicative Ability Vary with Phylogenetic Relatedness to Humans." *Journal of Social, Evolutionary & Cultural Psychology* 4/1 (2010): 34. <https://doi.org/10.1037/h0099303>.
- Heijnen, S., de Kleijn, R., Hommel, B. "The Impact of Human-Robot Synchronization on Anthropomorphization." *Frontiers in Psychology* 9 (2018): 2607. <https://doi.org/10.3389/fpsyg.2018.02607>.
- Hill, S.G., Barber, D., W. Evans, A. "Achieving the Vision of Effective Soldier-Robot Teaming." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts* (2015). <https://doi.org/10.1145/2701973.2702026>.
- Hoffman, G., and Breazeal, C. "Collaboration in Human-Robot Teams." In *AIAA 1st Intelligent Systems Technical Conference*. Infotech@Aerospace Conferences. American Institute of Aeronautics and Astronautics (2004). <https://doi.org/10.2514/6.2004-6434>.
- . "Effects of Anticipatory Action on Human-Robot Teamwork Efficiency, Fluency, and Perception of Team." *Proceedings of the ACM/IEEE International* (2007). <https://dl.acm.org/citation.cfm?id=1228718>.
- Hoffman, G. *Ensemble: Fluency and Embodiment for Robots Acting with Humans*. Ph.D. Media Arts and Sciences. MIT, September 2007.
- . "Embodied Cognition for Autonomous Interactive Robots." *Topics in Cognitive Science* 4/4 (2012): 759–72. <https://doi.org/10.1111/j.1756-8765.2012.01218.x>. <http://ir.canterbury.ac.nz/handle/10092/8698>.

- London, B. "Australian military gets first drone that can fly with artificial intelligence." *CNN* (2020). Retrieved at <https://edition.cnn.com/2020/05/05/asia/australia-loyal-wingman-drone-intl-hnk/index.html>.
- Kaplan, A.D., Sanders, T., Hancock, P.A. "The Relationship Between Extroversion and the Tendency to Anthropomorphize Robots: A Bayesian Analysis." *Frontiers in Robotics and AI* 5 (2019): 135. <https://doi.org/10.3389/frobt.2018.00135>.
- Lin, P., Mehlman, M., Abney, K., Galliot, J. "Super Soldiers (Part 1): What is Military Human Enhancement?" In *Global Issues and Ethical Considerations in Human Enhancement Technologies*, edited by S. Thompson, 119 – 138. Hershey: IGI Global, 2014.
- Lin, P., Mehlman, M., Abney, K., Vallor, S., French, S., Galliot, J., Burnam-Fink, M., LaCroix, A., Schuknecht S. "Super Soldiers (Part 2): The Ethical, Legal, and Operational Implications." In *Global Issues and Ethical Considerations in Human Enhancement Technologies*, edited by S. Thompson, 139 – 160. Hershey: IGI Global, 2014.
- Mori, M. "The Uncanny Valley." *Energy* 7/4 (2005): 33-35.
- Mutlu, B., Bartneck, C., Ham, J., Evers, V., Kanda, T. (eds). *Social Robotics: Third International Conference on Social Robotics, ICSR 2011, Amsterdam, The Netherlands, November 24-25, 2011. Proceedings*. New York: Springer, 2011.
- Nyholm, S. "The Ethics of Crashes with Self-Driving Cars: A Roadmap." *Philosophy Compass* 13/7 (2018). <https://doi.org/10.1111/phc3.12506>.
- Obaid, M., Kistler, F., Kasparavičiūtė, G., Yantaç, A.Y., and Fjeld, M. "How Would You Gesture Navigate a Drone?: A User-Centered Approach to Control a Drone." In *Proceedings of the 20th International Academic Mindtrek Conference*, 113–21. New York: Association for Computing Machinery, 2016. <https://doi.org/10.1145/2994310.2994348>.
- Ono, T., Imai, M., Ishiguro, H. "Anthropomorphic Communications in the Emerging Relationship between Humans and Robots." *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication* (2000). <https://doi.org/10.1109/roman.2000.892519>.
- Reichenbach, J., Bartneck, C., Carpenter, J. "Well Done, Robot! The Importance of Praise and Presence in Human-Robot Collaboration." In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 86–90. New Jersey: IEEE, 2006. <https://doi.org/10.1109/ROMAN.2006.314399>.
- Riek, L. D., Rabinowitch, T., Chakrabarti, B., Robinson, P. "How Anthropomorphism Affects Empathy toward Robots." In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2009): 245-46. <https://doi.org/10.1145/1514095.1514158>.
- Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joubin, F. "Effects of Gesture on the Perception of Psychological Anthropomorphism: A Case Study with a Humanoid Robot." In *Social Robotics*, 31–41. Berlin/Heidelberg: Springer, 2011. https://doi.org/10.1007/978-3-642-25504-5_4.
- Scharre, P. *Army of None: Autonomous Weapons and the Future of War*. New York: WW Norton & Company, 2018.
- Servais, V. "Anthropomorphism in Human-Animal Interactions: A Pragmatist View." *Frontiers in Psychology* 9 (December 2018): 2590. <https://doi.org/10.3389/fpsyg.2018.02590>.
- Severson, R.L., and Shailee, R.W. "Imagining Others' Minds: The Positive Relation Between Children's Role Play and Anthropomorphism." *Frontiers in Psychology* 9 (November 2018): 2140. <https://doi.org/10.3389/fpsyg.2018.02140>.
- Shen, Solace, Slovak, Petr, Jung, Malte F. "'Stop. I See a Conflict Happening.': A Robot Mediator for Young Children's Interpersonal Conflict Resolution." *HRI '18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (2018): 69–77.

- Singer, P.W. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. London: Penguin, 2009.
- Slocombe, G. "Autonomous Warrior: Major Air, Land and Sea Exercise." *Asia-Pacific Defence Reporter* (September 26, 2019). Retrieved on May 1, 2019 from <https://asiapacificdefencereporter.com/autonomous-warrior-2018/>.
- Soltanzadeh, S., and Mooney, M. "Players Within a Team: Understanding the Structure of Team Performance Through Individual Functions and Team Objectives." *International Journal of Sports Science & Coaching* 5/1 (2018): 84–89. <https://doi.org/10.1123/iscj.2017-0032>.
- Sparrow, R., and Howard, M. "When Human Beings Are like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport." *Transportation Research Part C: Emerging Technologies* 80 (July 2017): 206–15. <https://doi.org/10.1016/j.trc.2017.04.014>.
- Sparrow, R. "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems." *Ethics & International Affairs* 30/1 (2016): 93–116. <https://doi.org/10.1017/S0892679415000647>.
- Swiderska, A., and Küster, Dennis. "Avatars in Pain: Visible Harm Enhances Mind Perception in Humans and Robots." *Perception* 47/12 (2018). <https://doi.org/10.1177/0301006618809919>.
- Tahiroglu, D., and Taylor, M. "Anthropomorphism, Social Understanding, and Imaginary Companions." *The British Journal of Developmental Psychology* 37/2 (November 2018): 284–299. <https://doi.org/10.1111/bjdp.12272>.
- Tucker, P. "An AI Just Beat a Human F-16 Pilot in a Dogfight — Again." *Defense One online* (August 20, 2020). Retrieved at <https://www.defenseone.com/technology/2020/08/ai-just-beat-human-f-16-pilot-dogfight-again/167872/>.
- Turnbull, G. "Why the British army tested robots in muddy fields." *C4ISRnet* (January 11, 2019).
- Urquiza-Haas, E.G., Kotrschal, K. "The Mind behind Anthropomorphic Thinking: Attribution of Mental States to Other Species." *Animal Behaviour* 109 (November 2015): 167–76. <https://doi.org/10.1016/j.anbehav.2015.08.011>.
- Vertesi, J. "Seeing like a Rover: Visualization, Embodiment, and Interaction on the Mars Exploration Rover Mission." *Social Studies of Science* 42/3 (2012): 393–414. <https://doi.org/10.1177/0306312712444645>.
- Waytz, A., Cacioppo, J.T., Epley, N. "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism." *Perspectives on Psychological Science* 5/3 (2014): 219–232.
- Waytz, A., Epley, N., and Cacioppo, J. T. "Social cognition unbound: Insights into anthropomorphism and dehumanization." *Current Directions in Psychological Science* 19/1 (2010): 58–62.
- White, R.C., Remington, A. "Object Personification in Autism: This Paper Will Be Very Sad If You Don't Read It." *Autism: The International Journal of Research and Practice* 23/4 (2018). <https://doi.org/10.1177/1362361318793408>.
- Wodehouse, A., Brisco, R., Broussard, E., Duffy, A. "Pareidolia: characterizing facial anthropomorphism and its implications for product design." *Journal of Design Research*, 16/2 (2018): 83-98.
- Zawieska, K. n.d. "Do Robots Equal Humans? Anthropomorphic Terminology in LAWS." 2015 UN CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems at Geneva, Switzerland (2015).
- "Washington Post 2007 Bots on The Ground.pdf." n.d.

- Zlotowski, J.A. “Understanding Anthropomorphism in the Interaction between Users and Robots.” Doctoral Thesis. University of Canterbury, 2015. <http://ir.canterbury.ac.nz/handle/10092/11259>.
- Zlotowski, J.A., Proudfoot, D., Yogeewaran, K., Bartneck, C. “Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction.” *Advanced Robotics: The International Journal of the Robotics Society of Japan* 7/3 (2015): 347–60. <https://doi.org/10.1007/s12369-014-0267-6>.
- Zlotowski, J.A., Sumioka, H., Bartneck, C., Nishio, S., Ishiguro, H. “Understanding Anthropomorphism: Anthropomorphism Is Not a Reverse Process of Dehumanization.” In *Social Robotics*, edited by Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Samk GeKenji Suzukik John-John Cabibihan, Friederike Eyssel, Hongsheng He, 618–27. New York: Springer International Publishing, 2017. doi.org/10.1007/978-3-319-70022-9_61.
- Zlotowski, J.A., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., Ishiguro, H. 2018. “Model of Dual Anthropomorphism: The Relationship Between the Media Equation Effect and Implicit Anthropomorphism.” *International Journal of Social Robotics* 10 (2018): 701–714. <https://doi.org/10.1007/s12369-018-0476-5>.
- Zlotowski, J.A., Sumioka, H., Nishio, S., Glas, D.F., Bartneck, C., Ishiguro, H. “Persistence of the Uncanny Valley: The Influence of Repeated Interactions and a Robot’s Attitude on Its Perception.” *Frontiers in Psychology* 6 (2015): 883. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00883>.
- Zlotowski, J., Proudfoot, D., Bartneck, C. “More Human than Human: Does the Uncanny Curve Really Matter?” *Proceedings of the HRI2013 Workshop on Design of Humanlikeness in HRI from uncanny valley to minimal design* (Tokyo, 2013): 7-13.

Support Publication #17 [**SP17**]

Cappuccio M.L., Devitt K.S., Cara Wrigley (2020),
*Attitudes and Beliefs toward the Introduction of a Logistic Robot
in a RAAF Base Warehouse*

Unpublished report - January-March 2020, 18 pages, 7547
words. Commissioned by Royal Australian Air Force.

Redacted.

Support Publication #18 [SP18]

Cappuccio M.L., Peeters A., McDonald W (2019),
“Sympathy for Dolores. Moral consideration for robots based on virtue and
recognition” *Philosophy & Technology*, 1-23 (2019), Springer Nature.

DOI: 10.1007/N13347-019-0341-y

Reproduced with permission from Springer Nature. All rights reserved.

RESEARCH ARTICLE

Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition



Massimiliano L. Cappuccio^{1,2}  • Anco Peeters³  • William McDonald²

Received: 2 June 2018 / Accepted: 28 January 2019 / Published online: 22 February 2019
© Springer Nature B.V. 2019

Abstract

This paper motivates the idea that social robots should be credited as moral patients, building on an argumentative approach that combines virtue ethics and social recognition theory. Our proposal answers the call for a nuanced ethical evaluation of human-robot interaction that does justice to both the robustness of the social responses solicited in humans by robots and the fact that robots are designed to be used as instruments. On the one hand, we acknowledge that the instrumental nature of robots and their unsophisticated social capabilities prevent any attribution of rights to robots, which are devoid of intrinsic moral dignity and personal status. On the other hand, we argue that another form of moral consideration—not based on rights attribution—can and must be granted to robots. The reason is that relationships with robots offer to the human agents important opportunities to cultivate both vices and virtues, like social interaction with other human beings. Our argument appeals to social recognition to explain why social robots, unlike other technological artifacts, are capable of establishing with their human users quasi-social relationships as pseudo-persons. This recognition dynamic justifies seeing robots as worthy of moral consideration from a virtue ethical standpoint as it predicts the pre-reflective formation of persistent affective dispositions and behavioral habits that are capable of corrupting the human user's character. We conclude by drawing attention to a potential paradox drawn forth by our analysis and by examining the main conceptual conundrums that our approach has to face.

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

Anco Peeters
mail@ancopeeters.com

William McDonald
wmcdonal@uaeu.ac.ae

Keywords Virtue ethics · Social recognition theory · Social robotics · Moral consideration for robots · Moral patiency · Alienation · Habit formation · Character

1 The Debate on Moral Consideration for Robots

Artificial autonomous agents promise soon to play a central role in our lives as companions and co-workers (Nørskov 2016a; Dumouchel and Damiano 2017). Their increasing complexity increases the need for philosophical debate on the moral consideration for robots (MCR). Some ethicists contend that advanced socially interactive robots, i.e., artificial autonomous agents capable of establishing relatively rich and robust relationships with humans, should receive at least minimal respect and protection from their users, if not a fully-fledged recognition of dignity and rights (Gunkel 2017; Levy 2009; Gerdes 2016). In the opposite camp, the skeptics of MCR contend that these artificial agents are patently non-human, non-sentient, and of only limited intelligence (Bryson 2010a). The issue at the core of MCR is whether or not robots could and should be credited the status of moral patients (Floridi and Sanders 2004; Gerdes 2016; Tavani 2018). Should the actions that affect (or simply target) social robots be considered ethically significant rather than neutral?

The worry fueling this discussion is that, in the absence of precise normative and legal prescriptions, human users might treat their social robots in ways that would be considered unacceptable if applied to other human beings or even non-human animals (Whitby 2008). For example, robots could be exploited as slaves or sexual objects, treated with contempt and arrogance, and ordered to act or pose in offensive ways (especially if robots were customized to display characteristic discriminatory marks, like racial and gendered features; see Sparrow 2017). Taking this behavior to its extreme, it might even result in humans deliberately damaging robots to satisfy vanity and cruel instincts, as part of a sadistic game: organized robot “torture” could become a decadent form of entertainment and a remunerative business, such as depicted in the narrative universe of *Westworld* (Murphy 2018a).

Many share the intuition that, if such behaviors are to be considered morbid and depraved when they target humans, they cannot be morally neutral when they target artificial agents who, although non-sentient, explicitly present themselves as socially interactive (Coeckelbergh 2010b). A robot’s exterior appearance may solicit compassion and attachment in humans, and its cognitive resources may be powerful enough to establish enduring and relatively rich relationships with their users. That is why the proponents of MCR feel that abusing social robots is not only distasteful, but morally reprehensible. Can a fully-fledged theory of ethics consistently capture the normative reasons behind the intuition that social robots are legitimate moral patients? The recent philosophical discussion about MCR prevalently focuses on the attempts to justify robots’ rights (Darling 2012; Seibt et al. 2014; Gunkel 2017).

This paper aims to show that rights’ attribution is neither the only nor the best way to express MCR and argues in favor of an alternative approach to MCR based on the combination of virtue ethics and social recognition theory (SRT). After explaining, in Section 2, why we consider the typical arguments in favor of robot’s rights inadequate, we introduce a virtue ethical argument for MCR in Section 3. This argument asserts that abusive behaviors are reprehensible regardless of whether they target humans or

robots because either way they cultivate vices in the human who, perpetrating such actions, damages her own moral character. In Section 4 and Section 5, we examine the advantages and the objections offered by this approach. To answer the objection that neither the justification nor the domain of application of MCR is sufficiently specified by virtue theory, SRT is introduced in Section 6. Here, we clarify that social robots, unlike other technological artifacts, deserve moral consideration because they tend to establish a relationship of pseudo-recognition with their human users and reciprocate their recognitive responses. This concept is expanded in Section 7 where, to show the complementarity of virtue ethics and SRT, we argue that the social relationships founded on alienating recognitive dynamics cultivate vices. Also, we consider two additional objections concerning the implementation of our approach: first, humans cannot establish a fully-fledged and symmetrical relationship of recognition with existing robots and second, practical wisdom suffices to prevent the formation of alienating feelings during human-robot interaction (HRI). In Section 8, we argue that forceful suppression of the social dispositions spontaneously solicited by a robot is neither virtuous nor effective. That is because an intrinsic contradiction lies at the core of the notion of social robotics, which we critically discuss in Section 9. Section 10 concludes our discussion of MCR presenting the practical implications our approach.

2 Robots' Rights: Arguments in Favor and Against

When arguing that robots deserve rights (Laitinen 2002; Darling 2012; Gunkel 2017; Gerdes and Øhrstrøm 2015), scholars generally apply either an *objectivist* or a *relational* argumentative strategy. The first strategy assumes that a social robot qualifies for personhood and rights if it holds certain objective qualities or features, such as freedom of will or sentience (Gerdes 2016). The second strategy assumes that rights are social constructions and that a society could attribute personhood to robots only if they have established a social relationship with humans—for example, being considered as citizens, companions, or community members (see Coeckelbergh 2010a).

Both strategies grapple with problems: the first strategy's requirements on MCR are too demanding to be realistically met within the foreseeable future; the second sets the bar of MCR at a much more reasonable height, but does not explain why MCR should imply rights' attribution.

On a closer look, the first strategy is problematic because existing social robots are too unsophisticated to be considered sentient. Science does not provide us with any reason to believe that today's most complex service robots are even minimally sentient. Moreover, they do not display—and will hardly acquire any time soon—any of the objective cognitive prerequisites that could possibly identify them as persons or moral patients (e.g., self-awareness, autonomous decision, motivations, preferences). Finally, we lack the tools to predict or assess when and if these requirements will ever be met by robots.

In response to this objection for “lack of sophistication,” the advocates of robots' rights typically reply by shifting the attention from the robot's objective qualities to the cultural and normative context in which such qualities are construed. In this spirit, *anti-anthropocentric* and *constructivist* arguments are put forward to deconstruct, update, and broaden the most conservative notions of right and personhood (Coeckelbergh 2010a; Gunkel 2017).

Anti-anthropocentrism remarks that, against the backdrop of techno-scientific advances and societal transformations, the legal concept of “personal rights” became more comprehensive when it stopped being justified by the presumption of human exceptionalism. This is why, in some contexts, it came to include not only humans, but also animals, corporate entities, and certain components of the natural ecosystem (Hauskeller 2016).

Constructivism questions the idea that robots need a natural endowment of human-like intelligence or autonomy to enjoy some of the human rights, because personhood and rights are conventional notions, intersubjectively constructed within socio-cultural practices (Whitby 2008). On this assumption, the fact that robots play social roles and participate in collaborative practices involving humans could be enough to justify the attribution of pseudo-personhood (Hauskeller 2014).

Although these responses cannot persuasively support the first argumentative strategy in favor of robots’ rights, which fails to specify what objective features would qualify a robot as a rights’ holder, they can effectively support the relational strategy. These responses explain why the conditions to justify MCR emerge contextually, as historical effects of certain social interactions and power relations, rather than being a normative precondition that pre-exists and justifies such relations (Laitinen 2002). Whether human or robotic, personal identities are socio-culturally constructed through the maturation of intersubjective relations. That is why MCR must be understood as a conventional norm established by humans contextually, through a process of public validation.

Importantly, although the relational argument seems generically capable to justify MCR, it gives us no reason to believe that MCR should specifically involve the kinds of entitlements or status that we usually indicate with the term “rights.” Rights’ attribution is significantly more specific and demanding than a generic claim for MCR (Coeckelbergh 2010a), because it implies that robots hold special status and moral dignity. However, robots do not need human-like status or dignity in order to establish efficacious social interactions with humans and, in turn, social interaction with robots, even if emotionally intense, does not incline humans to attribute to robots status or dignity comparable to their own (Küster and Świdarska 2016).

Humans show a strong tendency to anthropomorphize robots and develop attachment to them (Duffy 2003; Damiano and Dumouchel 2018), but would the preservation of any robot be given priority over the rights of any human? Sparrow (2012) argues this is a critical test for the attribution of personhood and rights to robots in a relational perspective, a test that robots cannot pass in our society. Humans come first, from both a psychological and normative-legal point of view.

Arguably, the primary reason most humans do not attribute rights to social robots, and typically do not show any spontaneous inclination to do it, is that, within our current HRI practices, robots are conceptually categorized as inanimate tools (Levy 2007). This categorization is incompatible with the status of a rights holder, because it implies that robots are mere means to a human end rather than ends in themselves. Therefore, if humans do not accept robots as rights holders and persons, it is not simply because of personal biases or emotional factors: the role assigned to robots is formally and practically an instrumental one, as in the consumerist culture robots are generally understood as tools or products designed to fulfill human goals and expectations (Bryson 2010a).

3 MCR Based on Virtue Ethics

The belief that robots exist to serve humans instrumentally may not justify rights attribution, but is compatible with some milder forms of MCR. The relational argument can be used to assert that at least some artificial companions hold moral patiency and that, subsequently, our interactions with them may involve certainly obligations (Carr 2018; Torrance 2008), although not necessarily such obligations are *towards them*. This argument does not imply that social robots deserve rights, dignity, personhood, or an axiological status higher than that of an inanimate tool. As such, it is not structured like a deontological argument (Coeckelbergh 2010a). Also, our argument does not presuppose that MCR would assure any utilitarian benefit to the humans or the robots involved, or to the society as a whole. Therefore, it is not based on consequentialist premises either.

Rather, the relational argument we propose is inspired by virtue ethics: it recommends treating social robots in a morally considerate manner because this is what a humane and compassionate agent would habitually do in their social interactions and because the opposite behavior would not be compatible with a virtuous lifestyle and moral flourishing (Peeters and Haselager in preparation). Virtue ethics, at least on some accounts, weighs the moral value of actions based on their *compatibility with and contribution to the cultivation of* our moral character. Crucially, if iterating a certain behavior—such as sadistic torture—contributes to the cultivation of vicious habits, then that behavior must be deemed morally problematic. On the other hand, a behavior that contributes to the cultivation of virtuous habits—such as treating others with care and respect—deserves to be considered intrinsically good and, as such, praiseworthy and desirable.

Virtues exist as potential trajectories of one's character: they are predispositions or inclinations to generate morally appropriate behaviors and decisions (Vallor 2016). Therefore, virtue ethics judges the agent's conduct on the basis of its compatibility with the virtues that inform the agent's character. Good conduct preserves, exercises, and strives to improve the existing virtues.

Like Aristotle's *Ethica Nicomachea*, many approaches to virtue ethics agree that the moral community in which the actor operates plays a key role in determining what habits and character traits have to be considered virtuous/vicious (Collins 2004; Oakley and Cocking 2008; Hursthouse 1999). In this paper, we will assume that compassion, empathy, and respect are fundamental political and social values in our globalized and cosmopolitan civilization; hence, they are presupposed by every contemporary virtue ethical approach to MCR.

Thus, from a virtue ethical perspective, robot mistreatment is blameworthy first of all because the abusive behavior *reveals* the presence of a moral defect in the abuser's character (Sparrow 2016), and secondarily because it *invites* immorality in oneself and others (Coeckelbergh 2018): not only does the abusive behavior qualify the agent in a negative way, as the kind of person who lives an unworthy lifestyle, but it also offers a contemptible example that other people could imitate. Robot mistreatment both expresses and reinforces existing negative character traits, cultivating them in both the abuser and other, potential abusers. This implies, among other things, that following one's own bad inclinations while dealing with a robot, with the risk of accentuating such inclinations, constitutes a form of self-harm. According to virtue ethics, these

considerations justify a negative moral judgment of the abuser and a reason to discourage the actions produced by their abusive inclinations.

Is virtue ethics ultimately compatible with, or subsumable by, consequentialist or deontological moral theories? Our virtue ethical account remains neutral on this question. No doubt virtue cultivation has morally praiseworthy consequences beyond the acquisition of those virtues, and the exercise of virtuous behaviors might be considered a deontological obligation. However, this paper aims primarily to delineate the virtue ethical arguments that support MCR, without depending on whether or not such arguments could be defended from consequentialist or deontological perspectives. This choice is consistent with our assessment that robot abuse is possible and morally reprehensible regardless of the fact that robots are not sentient and have no rights.

Abusive behavior may have a bad influence on the actor, cultivating their vices—whether it targets another human, an animal, or a robot (Nomura et al. 2016). To clarify, the immediate effects of the abusive action committed on the social agent at this or that point in time are not what the virtue ethicist would deem morally problematic. Nor is the agent's status or the amount of harm suffered by the robot relevant. Rather, the problem is the formation of a potentially malicious character in the user through the consolidation of persistent unempathetic dispositions and vicious traits (Gini et al. 2007).

Although systematic long-term experimental studies on HRI are still lacking, various bodies of empirical evidence (reviewed by Coeckelbergh 2018) strongly indicate that humans naturally tend to empathize with robotic agents and that the conduct towards robots is informed by the moral customs, societal conventions, and ethical beliefs of the human users. That is why virtue ethicists are concerned by the potential impact of inhumane, unempathetic relationships established with social robots in daily life: if such relationships deprive the actor of humanity, compassion, and sense of responsibility, then the human actor's moral conscience risks being “damaged,” in the specific sense of “corrupted.”

Children raised by artificial nannies (Whitby 2010; Bryson 2010b), adult consumers of robotic sex (Whitby 2012), and players of violent games targeting robots (Sparrow 2017) represent the most paradigmatic and worrying hypothetical scenarios. The user's age constitutes a special concern in these scenarios, as deviant tendencies will likely deeply ingrain themselves if encouraged by wrong or absent role models during a formative stage.

Adults equally risk moral corruption when they indulge protractedly in a reprehensible lifestyle. The sudden explosion of inner conflicts accumulated during their psychological development may reveal unresolved tensions in their character. Those who possess insecure natures and continuously seek confirmation from others risk being affected the most: when they exert a complete control over human-like artificial agents they gain an opportunity to explore, embrace, and indulge in morally dubious—or even morbid—egotistic inclinations that they might not even have suspected they had. Virtue ethics recommends preventing such an outcome.

4 Strengths of the Virtue Ethical Account

A theoretical approach to MCR based on virtue ethics enables a level of thematic analysis that other approaches in normative ethics cannot offer. First of all, it does not

appeal to some objective qualities or powers of the social robot. Therefore, it does not presuppose that a robot has developed particular abilities or a special status: without such heavy requirements (which currently cannot be met by any non-fictional autonomous agent), virtue ethics can explain why even today's much simpler social robots may deserve some moral consideration. Such an explanation depends on the situated and pragmatic nature of the virtue ethics approach, which focuses on the concrete reality of the human-robot relationship (HRR). The relationship is considered in terms of the embodied and emotional details of its immanent realization, rather than the metaphysical status and moral dignity of the robot.

Existing social robots do not meet the requirements that other theories demand to justify moral consideration. On the one hand, theories that postulate that moral consideration follows from personhood struggle to find objective reasons to consider robots as actual persons (Hauskeller 2016). On the other, theories that postulate that moral consideration is a means to prevent suffering and pain will find it impossible to attribute these psychological states to non-sentient robots (Torrance 2008). These theories are unable to capture effectively why mistreating or disrespecting social robots is ethically reprehensible because they do not attach any specific moral judgment to the relationship established with social robots and the actions performed by the agent as part of it (Whitby 2008; Hauskeller 2016).

Virtue ethics, on the other hand, can recognize the moral value of HRR because the agent's conduct is evaluated on the basis of the attitude taken towards the target, while the effects produced on the target are judged in relation to their compatibility with the agent's moral standards. We defend an agent-based and exemplarist approach to virtue ethics (e.g., Zagzebski 2010) because such an approach is in a position to express a negative judgment about the mistreatment of non-sentient agents, such as robots, which *prima facie* do not seem mistreatable. Virtue ethics does not need to claim that mistreating a social robot is bad for the robot. Rather, it asserts that mistreating a social robot is bad for the human user insofar it prevents the user from fulfilling their own moral nature and achieving self-improvement. Such actions are not bad in a utilitarian sense (as "disadvantageous" or "practically detrimental"), but in the sense that they are incompatible with the model of compassion and empathy that informs the social habits of a virtuous agent.

The virtue ethics argument in favor of MCR works even if the agent knows that the robot in question is merely an inanimate tool. It works because the moral judgment about any social relationship depends on the habits and dispositions that scaffold the relationship. Relationships between a human user and its social robot are explicitly meant to reproduce the embodied, emotional, and psychological details of a companionship relationship (Damiano and Dumouchel 2018). The companionship bond habitually involves social virtues like empathy, compassion, and reciprocal consideration and respect. These virtues are pre-reflectively embedded in the character of the social interactants, which in turn is shaped by their embodied (behaviors), emotional (feelings), and psychological (evaluations) dispositions. One does not need to hold the belief that a robot is a person to establish a satisfactory pseudo-social relationship with it (that is, a pseudo-relationship imitating a real relationship so well that it feels satisfactorily social). Even without such explicit belief, the interaction with a robot can still be loaded with a moral valence that depends on the particular pattern of embodied, emotional, and psychological dispositions solicited by the esthetic properties

of the robot (Coeckelbergh 2010b). This is confirmed by the empirical evidence that automatic visceral empathic reactions are produced in humans by the observation of both humans and objects that look like humans (Suzuki et al. 2015).

Consider Dolores, the tragic robotic heroine of the *Westworld* series, to which the title of this paper is dedicated. Or consider any social robot perfectly indistinguishable from an animated natural agent (whether human or non-human). The interaction with such a robot spontaneously elicits the same bottom-up pre-reflective responses that it would have elicited if it was a natural agent (which presumably is the very reason the robot was built in the first place). From a moral viewpoint, the relationship with the robot has the same virtuous or vicious potential as the corresponding relationship with a person or pet (Sparrow 2016), even if the agent is aware of dealing with a robot. Whether this explicit awareness is in place or not does not make a moral difference because, according to our agent-based approach to virtue ethics, practicing the pre-reflective habits and dispositions (i.e., the “moral affordances” ecologically embedded in the social environment) carries more intrinsic moral valence—i.e., more potential to cultivate vices and virtues—than simply entertaining the explicit beliefs associated with them (Rietveld 2008; Brownstein and Madva 2012). In fact, due to the strong habitual nature of all social practices, the robot would elicit similar virtuous or vicious responses, even if it was not entirely identical to a natural agent. Such elicitation merely requires that the habitual behavioral and emotional responses triggered by the interaction with the robot were similar to those typically enacted through the interaction with the human.

Virtue ethics gives us reasons to judge the moral aspects of HRI according to normative standards analogous to those we typically use to judge human-human interaction (HHI) or any other genuinely social form of interaction. The degree of perceptual similarity between the two kinds of interaction depends on the extent to which relationships with robots conjure dispositions and habits typically involved in our social life. Hence, the HRR can have a moral valence through its pragmatic situatedness. Such valence can therefore only be evaluated contextually, factoring in the fine-grained psychological and behavioral specificity of the agents involved and considering the particular experiential and emotional circumstances in which the relationship is established.

5 Three Objections: Applicability, Justification, and Indirectness

So far, we have described what a virtue ethical account of MCR would be and provided reasons to prefer this model over others. However, some objections have been raised against similar virtue ethical accounts. Coeckelbergh (2010b) is skeptical for three fundamental reasons.

The first and the second objection raise, respectively, an “application problem” and a “justification problem.” They contend that virtue ethics lacks clear criteria to determine what entities deserve moral concern, how such concern should be applied, and why it should be applied only to certain entities rather than others. In other words, even conceding that a virtuous agent is expected to treat certain entities as if they were worthy of virtuous conduct (like respect and compassion), it is still unclear why this counterfactual characterization should identify only some categories of non-human

entities (like social robots), excluding other kinds of advanced technological artifacts. These objections allege that virtue theory does not provide sufficient ground to demarcate robots from other objects, making it impossible to tell on what basis social robots should be included in the group of entities that deserve moral consideration, and how the decision should be made.

The third problem raised by Coeckelberg is that virtue theory could reply to these objections only by justifying the particular moral status of robots through an ontological characterization of the robot. In doing so, virtue ethics would inherit the problem of “lack of sophistication” that affects the objectivist strategy. According to Coeckelbergh (2010b), this happens because the virtue ethical arguments in favor of MCR are indirect, i.e., derivative and tied to an anthropocentric bias. Ignoring that the moral value of robots is an effect of HRI, virtue ethics requires that the human position in the universe is ontologically central and axiologically superior, because the only value that virtue ethics can see in a robot is the one that reflects and confirms the primacy of human nature.

This problem does not affect our virtue ethical account of MCR, which does not rely on objectivist/ontological anthropocentric prerequisites and explicitly recognizes that the moral valence of HRR has a relational origin. As previously illustrated, our approach is based on the quality of the (pseudo-)social relationship arising from HRI. The robot’s objective features and the comparison with distinctive human features (intelligence, sentience, etc.) do not play a role here. There is no necessary *a priori* reason that the example of virtue inspiring the moral agent, i.e., the role model that they imitate, should be a human one.

Although the concept of virtue derives from the humanistic tradition, virtues do not necessarily have to be considered unique human features that the robot can only imitate imperfectly. Rather, virtues can indicate that an agent is equipped and motivated to establish a responsible balance with the world environment, inhabiting it in harmony with the other entities that populate it. Through its interactions with the others, any virtuous agent can contribute to making sense of the world environment as an ecosystem and to preserving it. This ecocentric, as opposed to anthropocentric, interpretation of virtue ethics is supported by its classical sources (Zwolinski and Schmidt 2013). It reflects the environmentalist concerns formulated by Coeckelbergh (2010a) and Nørskov (2016b) and incorporates their worries about the indirect/derivative nature of the arguments for MCR. Under this interpretation, virtue ethics is not affected by anthropocentric bias because it is in a position to consistently promote a compassionate attitude towards the target entities and furthermore disavows the opposite type of attitude, regardless of whether the target entities are natural or artificial, sentient or inanimate, and whether the agent knows the entity’s true ontological status or not.

For the virtue ethicist, the moral judgment about the quality of HRR does not depend on a decision unilaterally made by the human agent concerning the ontological status of the robot, but on the behavioral habits and character dispositions that spontaneously and dynamically arise through HRI and because of it. The moral judgment about HRR typically relies on the experience of virtues and vices that are typically embedded in social relationships the human agent is already familiar with, but that does not necessarily presuppose an anthropocentric bias or the fact that HRR is an imitation of human-human relationships. The phenomenon of anthropomorphism certainly plays a role in MCR, but must not be confused with anthropocentrism: as we will argue

through the rest of the paper, it is not indispensable that robots look human-like to establish morally significant HRR, as anthropomorphism can target virtually any natural or artificial entity.

More is required to answer Coeckelberg's other two objections, which ask how the analogy between human-robot and human-human relationships has to be theoretically justified and practically applied. This question truly is critical for any MCR theory and not only for the ones based on virtue ethics. The status assigned by Western civilization to a social robot is typically that of a serviceable artifact like many others, the kind of artifact that usually is not identified as having the value of being worthy of respect, compassion or other forms of consideration. But, if social robots are just artificial tools, then why would we expect them to cultivate virtuous or vicious behaviors in humans more than, say, intelligent cars, sophisticated toys, and realistic computer simulations? And how should we measure an artifact's capability to cultivate vices and virtues in its human users?

A social robots is a very peculiar kind of artifact, one that essentially differs from other types of instruments: only social robots are meant to establish a (pseudo-)social relationship with their users, with all the psychological and moral implications that this entails. Unlike other technologies, social robots reproduce the sphere of social interaction by their distinctive capability to evoke psychological and emotional behaviors and to respond to them. How? Social robots may establish a seemingly reciprocal self-other relationship with their user. For such a relationship to obtain, one must recognize another entity as the social "other" and experience oneself as that partner's other in turn. A social other is one who can recognize their partner as their "other." Thus, a fully-fledged self-other relationship requires a two-sided recognition dynamic and hence presupposes at least some very basic form of mutuality, or it would not be able to support the formation of a social bond.

The (quasi) mutual social recognition we experience when we interact with social robots is fundamental to the moral dimension of HRR. Mutual recognition dynamics, in combination with the principles of virtue ethics outlined so far, provide a foundation to MCR, defining both its domain of application and its theoretical justification. That is why our answer to Coeckelberg's first and second objection is inspired by social recognition theory (Honneth 1995; Ikäheimo and Laitinen 2011).

SRT situates the historical foundation and the normative justification of socio-political relationships within the dynamics of reciprocal acknowledgment and social role attribution (Laitinen 2002; Laitinen et al. 2016). Recognition theorists have highlighted that social robots, more than any other kind of artifact, tend to establish complex, although imperfect, recognitive relationships with their human users (Laitinen 2016a). This means social robots are designed to produce behaviors and expressive responses calibrated to be interpreted as social by the users and, correspondingly, to recognize the social behaviors and emotions produced by the humans, making sense of them and appropriately responding to them (Darling 2012; Ball et al. 2014). This allows us to reply to the objections of application and justification: we do not have reasons to apply our virtues in the same way to all entities, but social robots are definitely among the entities that demand our moral consideration because they are actively recognized by (and at the same time appear to recognize) the human users as social agents.

Humans tend to form pseudo-recognitive relationships with social robots and be affected deeply, in both a psychological and moral sense, by such relationships. How? As the same recognitive habits are implicated in both HHIs and HRIs, involving similar sets of virtuous and vicious dispositions, the pseudo-social relationships between robots and humans can be analogous to the actual social relationships that involve only humans. Hence, the relationship a virtuous agent builds with their social robot can—to some extent—imitate the habitual standards of compassion and respect that are built-in and, so to speak, “hardwired,” in morally admirable social relationships. This helps avoid the corruption of the user’s moral dispositions.

6 Moral Corruption as a Recognitive Phenomenon

Inappropriate HRI can trigger and progressively reinforce a process of moral decay. How would this corruption develop? Empirical evidence is sparse, but a rich phenomenological account of the typical structure of mutual recognition allows us to describe how social interaction impacts on virtues and vices.

Robots may have a bad influence over their human companions in different ways (Vallor 2016), but it is likely that corruption would typically operate through inciting a sense of entitlement. Feelings of self-importance are magnified by the experience of exerting total control over another agent, especially an agent that is seemingly intelligent, autonomous, and human-like in look and behavior. Exposed to daily interactions with perfectly obedient social robots, both adults and children would normalize and enjoy an experience of unconstrained power over autonomous agents. Media experts, educators, and sociologists worry that the habit of commanding digital assistants like Alexa and Siri make children lazy and spoiled and solicit sexist objectification in adults (Truong 2016; Fessler 2017; Murphy 2018b).

The egotistic potential contained in the experience of commanding an autonomous agent seems incomparably stronger than the gratification associated with the use of generic non-anthropomorphic machines, like vehicles, weapons, or industrial robots (Laitinen 2016b). Regardless of their objective power, most technological devices neither produce significant recognitive responses, nor solicit them in humans. They are not treated by the human user as if they were or could be social agents and, in turn, do not appeal to the human user as a person. However complex, their interaction neither resembles a social relationship between persons nor aims to mimic one.

Yet, social robots and digital assistants are precisely designed with the purpose of establishing a relationship of that kind. Only an autonomous social agent—especially the physically embodied kind, designed to communicate feelings and emotions with its human-like presence—can validate a human user, reassuring them that are recognized and valued as a person, because they deserve to be. Validation through social recognition is a fundamental mechanism of personality consolidation (Laitinen 2016b; Laitinen et al. 2016). It informs the agent about which character traits to reinforce. Validation also builds up self-esteem (Laitinen 2016b) and enables the agent to identify with a social role (father, husband, employee, etc.), which in turn solicits the acknowledgment of certain public responsibilities.

The role assigned to the human in the relationship with a social robot is typically a dominant one, and as such, it should be loaded with various responsibilities. In a

complementary manner, the robot's anthropomorphic look and behavior are usually designed to display loyalty and encourage trust in their human owner, who is more or less explicitly solicited to take a "master" role in the relationship. Social robots, unlike other devices, can actively motivate human users to perceive themselves as authority figures endowed with a dominant social role. What social robots grant to their human user is much more than effectual power: it is a power *relationship*, implying not only the recognition of the user's personal identity, but of her superior status (Ikäheimo and Laitinen 2007; Whitby 2008; You et al. 2011).

Becoming intoxicated by this power is a prominent risk of interaction with social robots. One essential feature of recognition dynamics is that one's self-perception is deeply shaped by the way one is publicly perceived. An important part of our identity depends on the expectations that others have about us and on our expectations regarding those expectations. Accordingly, human users might not be prepared to handle the identity-redefining experience of being treated as masters. In HRR, more than ever before, the dominant role is assigned arbitrarily, and any human—however unworthy or unprepared—can take it on.

Once this relational template is formally established, any morally incompetent human can be granted an imperious position by the simple fact of being human. Such power would likely be dissociated from all the moral and intellectual prerequisites that usually come with (and are expected from) a position of authority: wisdom, expertise, moral responsibility, and public accountability.

Thus, not only will the human user have a tremendous power over the robot they own, the human user will never face the uncomfortable experience of being judged or questioned by the robot if the robot is programmed to follow their orders blindly. Due to this incongruity, feelings of responsibility by the human user in this position of authority may degenerate, motivating various internal conflicts. Without proper supervision, an outcome of these idiosyncratic instabilities might be that the human would acquire an authoritarian stance. In the long term, habituation of this stance might inflate the expectations to be fully served and cherished while never feeling challenged by or responsible for others.

Social robots could be a perfect trigger for this regressive dynamic of moral "deskilling" (Vallor 2015), because they are meant to simulate unconditional recognition without expecting to receive any in return. Unlike the humans who serve in a subordinate position, robots might be programmed to ignore disrespectful and inconsiderate behaviors. This, in turn, might justify the human master's sense of entitlement even more, reinforcing the authoritarian stance.

Such privilege towards robots is likely to encourage self-indulgent and complacent habits, boost the self-awareness of the human users, erode their inhibitions, spoil their sense of empathy, and—worst-case scenario—motivate them to tolerate, justify, or even replicate abusive behaviors against actual living creatures. If these dynamics were replicated on a massive scale, they could exacerbate social tensions in large communities and deteriorate civil cohesion.

These analyses combine virtue ethics and SRT because only their effective integration allows us to model the concrete socio-psychological dynamics underlying HRI and understand their complex moral implications. HRR is charged with moral value because abusing or mistreating an entity recognized as a social companion likely leads to the cultivation of a vice. In turn, SRT explains why the moral quality of HRR itself,

more than any particular action a human agent might commit over a robot, constitutes the key ethical issue: the relationship is imbued with an intrinsic moral significance, one that virtue ethics must value in the specific context in which HRI occurs.

That is because, if the social robot did not have any ability to invite a pseudo-social recognitive relationship with the human user, then the human user's psychology could not be affected by the way the robot is treated. Robot abuse would not lead to the development of vicious traits, if robots were not anthropomorphized, because the actions committed against the robot would hardly be perceived as abusive in the first place (Gray et al. 2012). Anthropomorphism—which is the implicit, automatic tendency of humans to perceive intentions, goals, and emotions in other agents, including agents that do not have minds—is a key prerequisite of MCR because it is a fundamental psychological component of the specific recognition dynamics that can, in certain conditions, lead to vice cultivation.

7 Asymmetric Recognition, Rational Discrimination, and Emotion Suppression

Virtue and recognition theories complement one another, but our account of MCR still has to face two “intra-theoretical” challenges, i.e., two objections arising within particular interpretations of virtue ethics and SRT.

The first intra-theoretical objection points out the incompleteness of the recognition relation that informs HRI, which appears inherently asymmetric and unidirectional (Scheutz 2012). Today's best social robots can only simulate the comprehension of the recognitive responses produced by the human, (poorly) reciprocating their behavioral component without understanding the phenomenology or affective experience involved.

The recognitive dynamics involved in HRI significantly differs from fully-fledged recognition dynamics: unlike human-human social relationships, which can be richly spontaneous, unpredictable, and never entirely controllable, HRR is at best “pseudo-social.” The robotic component plays only a prosthetic and compensative role, producing a relation that is inevitably unoriginal and stereotyped (Hauskeller 2016). But does this imply that the virtues that usually apply to HHI are not involved in HRI? We do not think so.

Even incomplete, asymmetric relationships, in which most of the recognitive work is done by humans (whether consciously or only semi consciously), can influence genuine social dispositions and fundamental traits of human moral character (Torrance 2008; Damiano and Dumouchel 2018). Vicious tendencies are formed not only in case of actual abuse or mistreatment of a real person: it is sufficient that, when the agent takes an abusive or authoritarian stance, they experience the associated sense of entitlement as rewarding and fulfilling. Despite HRR being pseudo-recognitive at best, the human affective and psychological responses produced during HRI are neither pretended nor simulated. For this reason, their reiteration may unintentionally reinforce abusive or authoritarian dispositions, normalizing the corresponding vicious habits.

Pseudo-recognition suffices for the cultivation of virtue or vice, because the moral valence of HRI and the perception of the robot as a moral patient promoted by HRR are enough to impact the relevant human character traits. Human dispositions are affected

significantly by how the HRR is subjectively perceived and intersubjectively contextualized, even when the relationship is asymmetrically centered on the recognitive work done by the human, as the human is the only truly sentient element in the relationship.

The second intra-theoretical objection posits that virtue ethics should contain the rational resources necessary to distinguish between actual and merely simulated social relationships. Discriminating between interactions with real and pseudo-social agents—says the objection—means recognizing that only the former is morally relevant, and therefore significant to a path of self-improvement.

Practical wisdom—“*phronesis*” in classical Greek virtue ethics—is, after all, one of the most important virtues and a distinctive mark of human intelligence. The Aristotelian tradition understands it as the ability to discern how to act virtuously in practical circumstances, and why. Unless robots were totally indiscernible from humans, a very modest amount of *phronesis* should be sufficient to recognize the differences between a non-sentient robot and a truly sentient person. Subsequently—says the objection—every human agent should be sufficiently wise to understand that there is no objective reason for treating the robot like a human, even if they feel emotionally compelled to do so for having established a pseudo-social relationship with it. Why should a wise human exert moral consideration for their robot then, if they know perfectly well that there is no objective reason to do so? Is not MCR inauthentic when it allows for willful attempts to ignore an evident distinction between human and robot?

Our answer is that virtue cultivation involves a complex mix of explicit (reflective, deliberative) and implicit (pre-reflective, habitual) capabilities. The moral character of an agent is the (more or less precarious) result of this mix, which fluidly supervenes on the dynamic interconnections between its underlying psychological processes. Formation of moral habits is one of the most important parts of moral character, as habits provide an important source of motivation and practical moral competences, i.e., the capability to adequately deal with complex circumstances adjusting one's conduct to the normative demands of the context.

Moral habit formation includes the consolidation of both practical skills and implicit knowledge and as such is a largely pre-reflective and to some extent cognitively impenetrable process (Brownstein and Madva 2012): the agent can sense the direction of its progress and indirectly guide it through routine exercise, but does not have direct control over it. Virtue ethics cannot ignore this fact because it relies on an embodied understanding of the moral agent's character as a multi-layered, possibly fragmented and occasionally unstable network that integrates automatic behavioral routines and deliberative processes.

Based on this embodied and pragmatic understanding of moral agents, virtue ethics rejects the intellectualist view that reduces moral judgment to a fixed, psychologically detached, and entirely transparent rational mechanism.

Intellectualism interprets practical wisdom as a capacity for rational decision, disconnected from the other virtues (like compassion and empathy), and assumes that rationality competes against habitual dispositions in the formulation of accurate moral judgments. Against intellectualism, the embodied and situated approach to moral psychology stresses the continuity between *phronesis* and other virtues, as they all conjointly contribute to shape the complex, and often contradictory, landscape of deliberative capabilities and behavioral habits that form the moral character. Intellectualism is particularly contested by the situationist virtue ethical approaches that

emphasize the fragmented nature of moral character (Merritt 2000). An even stronger position is put forward by Harman (2000), which refers to well-documented empirical evidence to argue that situational factors, more than character traits, play a defining role in determining one's moral conduct in different practical scenarios.

However, Harman's argument does not simply question the intellectualist interpretation of virtue, it threatens the viability of virtue theory altogether, as it claims that virtues construed as character traits may not exist at all (Harman 1999). While Harman might be correct to point out that character can be fragmented and lack internal consistency, his dismissal of the existence of virtues, and hence of the usefulness of virtue theory, is itself hasty and built on questionable assumptions in his interpretation of experimental results. Furthermore, Harman ignores some of the subtleties even in the Aristotelian account of virtue theory, when he fails to take into account Aristotle's distinctions between virtue and continence, and vice and incontinence, and therefore focuses exclusively on the external behavior of the agent (Athanasoulis 2000). The exercise of a virtue or a vice, in the presence of continence and incontinence, can explain why psychological experiments focusing only on external behavior, like those used by Harman to prove the nonexistence of character traits, fail to capture the internal struggles and seem to deliver the result that it is the situation, to the exclusion of character traits, which determines behavior (Athanasoulis 2000).

Harman's critical perspective should be corrected with the principles of neo-sentimentalist moral philosophy (Rodogno 2016), which emphasize that practical wisdom is embedded into one's affective sensitivity. It is the sentimental education that shapes one's emotional and psychological habits, making them contextually accurate and efficacious. According to the sentimentalist view, which is compatible with the virtue ethical notion of moral character, *phronesis* is not a rational calculus opposed to blind instinct and emotion. Virtue ethics, in combination with SRT, point to the possibility of developing relatively stable and locally consistent sentimental dispositions to act ethically, which are nevertheless susceptible to mere conformity to social expectations (continence), and adaptability to different situations.

8 Inner Conflicts and Suppression of Virtuous Dispositions

A robot owner may feel deeply attached to their robot and may consider their relationship "social," even if they are aware that the robot is not a person and that its recognitive responses are simulated. A conflict between opposite, apparently incompatible dispositions may then arise. This conflict is commonly experienced by humans as "mixed feelings" or a tension between "reason and passions." Virtue ethics does not ignore this possibility, but it generally does not recommend suppressing one's affective and behavioral habits as a way to solve the conflict, if those habits support the cultivation of moral virtues. Habits form a strongly interconnected network, hence willingly suppressing a deeply ingrained virtuous tendency to favor another could be practically impossible. And, even if it were possible, such suppression might have a heavy psychological cost. This opens up another problem: why should we accept MCR in the first place, if relating to social robots in an emotional way may lead to such dramatic conflicts?

Our answer is that the psychological traits scaffolding MCR already are or (in normal circumstances) should be deeply rooted in our nature, whether this is innate or acquired, as a “second nature.” For most of us, it might be difficult, or psychologically unhealthy, to ignore these traits. One may find it easier to accept the enduring implications of one’s emotional bonds with a robot, if the efforts to prevent emotional instability are likely to generate an even deeper moral instability.

Training ourselves to not anthropomorphize robots, and hence prevent the development of empathetic dispositions and attachment, might seem a preferable alternative to emotionally investing in a robot by cultivating the virtues of MCR. Although this option might be practically preferable in some situations, we need to stress that shutting down one’s own affective and sentimental dispositions could take a heavy toll on one’s empathy and emotional intelligence, because anthropomorphising tendencies are built into human social cognition skills (Waytz et al. 2007). Such impoverishment in the agent’s psychological landscape would come at an unacceptable moral cost. A virtuous agent might have good personal justifications for being unwilling to, or incapable of, suppressing their positive dispositions and psychological habits, if they involve morally valuable character traits that should not be given up. Hence, the objective limits posed by the structure of our character constrain our capability of deliberately suppressing our own moral dispositions, even when such suppression is dictated by rationally justified hard moral prescriptions.

Let us examine how such a hard prescription would work, by comparing two opposite situations. In the first situation, a worker is given the task of dismantling obsolete social robots, or testing the stability of new prototypes by harshly kicking them as they walk. This cruel job becomes routine for the worker, who somehow gets used to see these robots just as machines, even if they look alive and behave as living beings would do during the process. The worker cannot be judged negatively for refusing to empathize with the robots (s)he has to dispose of and is therefore exempted from the expectations of MCR that would apply in other circumstances: the worker has never established a social relationship with the robots, and the protocols of his task require him to abstain from developing psychological and emotional habits associated with them.

In the second situation, a child is asked by its parents to say goodbye to its beloved robot pet, a cute artificial dog whose maintenance has become unaffordable. Consider the pedagogically problematic and psychologically traumatic effects on the child. The child’s family proceeds with caution, limiting as much as possible the disruptive effects on the child. The “caution” to be exerted by the parents in this case could involve, for example, sparing the child the most dramatic aspects of the separation, or assigning a symbolic meaning to this experience in order to make it more acceptable and tolerable (for example, ritualizing it with a “funeral,” see Brown 2015). The child might even learn from this experience, as it would from the loss of a living pet. By recognizing the emotional dimension for the child, the parents offer the developing child an opportunity for learning how to deal with sad experiences.

These examples show that the correct application of MCR does not always necessarily involve treating robots with respect and care, or developing an attachment for them as if they were persons. However, if the human agent has spontaneously developed virtuous (compassionate, empathic, etc.) attitudes towards others, including robotic others, then our account recommends caution in forcefully suppressing or

denying such dispositions. Whether the decision of suppressing one's moral habits is justified or not should be assessed contextually, considering its psychological impact and the balance of costs and benefits for one's moral flourishing.

In general, virtue theory disapproves of emotion suppression if it threatens to produce alienation and instability, which may preclude the cultivation of vicious tendencies and dramatic inner conflicts. This issue deserves to be carefully unpacked in the next section, as the seed of an alienating contradiction seems to be built into the very relationship of pseudo-social recognition that characterizes HRI.

9 The Anthropomorphizing While Dehumanizing Paradox

The psychological mechanisms postulated by virtue ethics and social recognition theory are not situated solely at the rational level, but can encompass much deeper and broader systems that constitute one's implicit identity. MCR is likely to face internal tensions and psychological contradictions: namely, the status we rationally and explicitly attribute to an entity like a robot does not necessarily coincide with the status we emotionally and subconsciously register as part of our subjective perception and personal interactive experience. As a result, while remaining strongly tied to one another, the normative (legal, moral) and the descriptive (psychological, behavioral) sides of HRR may not entirely coincide.

This internal tension is reflected by our ambivalence in the way we refer to social robots: a social robot is designed to be autonomous from their human user and, at the same time, merely instrumental to human goals. While social robots are built to look, behave, and display emotions like us, we take for granted that the robot's function and its very existence are instrumental to fulfill our needs and desires.

This ambivalence gestures at a paradox hidden at the core of the very idea of social robotics. As underlined by Sparrow (2017), the fundamental ethical problem at the core of social robotics is that, while robots are designed to be like humans, they are also developed to be owned by humans and obey them. The disturbing consequence is that, while social robots become progressively more adaptive and autonomous, they will be perceived more and more as slave-like. In fact, owning and using an intelligent and autonomous agent instrumentally (i.e., as an agent capable to act on the basis of its own decisions to fulfill its own goals) is precisely the definition of slavery. The moral implications, from the point of view of virtue ethics, are both evident and worrying.

As remarked by Nørskov (2014), our recognition dynamics involving social robots risk to become more and more alienating as the anthropomorphic quality of social robots is both affirmed and denied by the cultural standards of HRI. This quality becomes dehumanized and vilified as social robots are sold and bought like disposable commodities. The contradiction is that, while the human user is encouraged to invest emotionally in the social robot, the user will also be expected to treat the robot as an instrument, as a mere means to reach their personal practical goals. We call this contradiction the “anthropomorphizing while dehumanizing robots paradox” (ADP).

The negative effects of being served by robots might in the long run be comparable to the corruptive consequences of being served by human slaves, even if the human user rationally believes that the robot, unlike a human slave, does not have rights or an intrinsic dignity and freedom. These detrimental psychological effects would operate

subconsciously, at the level of the automatic formation of emotional responses, behavioral routines, and implicit beliefs (discriminatory biases).

Being served by robotic slaves might have an even higher corruptive potential than being served by human slaves. The master should feel at least indirectly questioned by the human slaves: but this feeling does not arise if the master knows from the beginning that his servants are just non-sentient and perfectly obedient machines. Rational awareness would relieve the master from any sense of responsibility, empathy, and care, but would not reduce the master's desire to enjoy the dominant position usually exerted over their artificial slaves.

If, as we suspect, the ADP constitutes the prototypical norm of the standard relationship between humans and robots, then the daily interactions with social robots will involve the risk of forming egocentric and narcissistic personality traits in the human users. These would be triggered or strengthened by the alienating condition of exerting unrestrained power over a servant that, while human-like in its capability to work, behave socially, and express affect, is nonetheless devoid of any human autonomy and dignity.

Like other types of cognitive dissonance, the tension between anthropomorphizing and dehumanizing attitudes can be explicitly formulated at the logical level, as a contradiction between incongruous beliefs, but its alienating effects are primarily experienced within the emotional sphere.

We can envision at least two possible strategies to dissolve the paradox, each meant to block one of the two conflicting attitudes. Either we diminish the anthropomorphizing attitude, diminishing all the dynamics that solicit empathy and attachment for the robot in the human user, or we counter the dehumanizing attitude, encouraging the human user to grow feelings of care and respect for the social robot. Neither solution is fully satisfactory, as both are likely to frustrate the expectations attached to the very idea of social robotics. The former undermines the possibility of using robots in the social domain, because it prevents the kind of intimate or companion-like interactions that a robot must create to establish trust and cooperation with a child, a patient with disabilities, or an elderly person. The latter engenders the risk of too strong an attachment or even an empathetic transfer on the side of the human agent, which could prevent them from disposing efficaciously of the robot according to its practical function.

Even if the human user's rational capabilities were not compromised by the attachment to the robot, inducing in the human user a stronger sense of responsibility, care, and respect for the robot might not always be the right solution to the vice-inducing alienation generated by the ADP. After all, attachment to the robot constitutes half of the antecedents in the ADP. That means that, if the other half is still in place, then positive feelings for, and emotional investment in, the robot could well exacerbate the contradiction rather than alleviate it.

Is a less alienating way to look at robots possible, one that does not generate the ADP and that, subsequently, does not solicit the formation of dehumanizing dispositions? If a solution exists, it likely requires a reconfiguration of the HRR according to a new concept of social recognition, capable of accounting for both the instrumental use of robots as artifacts and the possibility of giving them the care and respect that they deserve in light of their dignified status as social companions. Developing this proposal requires a conceptual effort that falls outside the scope of the present paper.

10 Prescriptive Applications

What are the practical implications of our approach to MCR? That is, what categories of actions deserve to be prohibited or at least discouraged as robot abuse or mistreatment? We do not think that this question could be answered by a one-size-fits-all moral judgment, as there is no way to generate a moral maxim precise and comprehensive enough to determine *a priori*, for example, when the violent destruction of a social robot constitutes a vicious action. The psychological, motivational, and contextual variables are too fine-grained and heterogeneous to allow such generalization.

This indeterminacy is not a limitation, but a strength of our model, because it reflects the real complexity of the motivational landscape of our actions as embodied and situated moral agents. Contrary to top-down prescriptive approaches to ethics, which reduce moral conduct to formal compliance with a general set of rules and decision-making procedures, virtue ethics represents a bottom-up approach. On this account, an action's moral value is not judged against the standards defined by a general protocol, but can only be appreciated within the contextual specificity in which it occurs. Furthermore, recognitive processes can involve complex and multi-layered psychological and motivational dynamics that can be opaque and, at times, contradictory.

While a universal ethical standard would probably be impossible, a comparative moral judgment, based on the differentiation between gradations of positive and negative behaviors, should be possible for many types of HRI. Such a moral judgment would have to take into account the vividness and depth of HRI and the resemblance with the corresponding HHIs, considering the moral quality of the habits and inclinations that are likely to be reinforced. For example, an individual who spends remarkable efforts and time to fulfill their maniacal desire to collect sex robots to brutally rape and torture them almost certainly belongs in the category of the vicious people who deserve attention. But an individual who uses the same kinds of robots in the same way as part of a supervised therapeutic program to overcome violent tendencies should be judged in a different way. Our moral judgment in one case and the other might imply that we have an obligation to either advise, guide, admonish, or even penalize these persons for their behavior. Analogously, we should feel compelled to invite them to recognize, address, and correct their deviant tendencies, possibly with the help of qualified counselors (e.g., Peeters and Haselager [in preparation](#)). But, in case of recidivism, what concrete action should be taken to discourage or sanction such behavior? What prescriptions, prohibitions, and sanctions could and should be enforced? Such questions deserve a more nuanced consideration than we are able to provide here.

Virtue ethics does not lean towards a hardline approach to discipline and could hardly advocate strict prescriptive and prohibitive norms. One strength of the agent-based approach to virtue ethics is its capacity to philosophically inspire individuals in their quest for personal flourishing, psychological stability, and existential fulfillment. Not unlike the strands of environmental ethics inspired by Eastern philosophical traditions (Coeckelbergh [2010a](#); Nørskov [2016a](#)), Aristotelian virtue ethics proposes a moral pedagogy based on a personal path of intellectual self-development and emotional maturation aiming to perfect one's balance with oneself, other individuals, society, and the environment. As this path is closely tailored to the specificity of one's personal vicissitudes and presupposes a careful understanding of each individual's unique background, the goal of virtue ethics is not to provide universal moral maxims

or context-independent decision-making protocols. That is why the proposed normative framework does not aspire to chastise or censor the lifestyles and the activities that adults may freely decide to embrace as part of their private life with their social robots, even if these lifestyles and activities are questionable and potentially corruptive.

This is not only because of the inherent difficulty to generate moral prescriptions sufficiently specific and diverse to cover the entire casuistry of potentially vicious activities. Even if such prohibitions could accurately represent the whole spectrum of possible vicious behaviors, they may still be useless, as virtue cultivation requires a strong personal drive, which can be either motivated from within, by personal motives and self-discipline, or from without, for example by wanting to live up to a role model or by seeking the esteem of one's peers. Role models and moral examples are generally more inspiring than prohibitions, as emphasized by the agent-based exemplarist approach to virtue ethics (e.g., Zagzebski 2010).

The fact that our approach does not take a strong stance on disciplinary intervention does not mean it hesitates to identify vicious behaviors and act upon them. Rather, we assume that a virtuous relationship with social robots, like the equivalent relationship between humans, requires lived experience more than abstract norms. That is why, rather than condemning or censoring the vicious applications to robots, our theory of MCR aims to suggest virtuous models of HRI, providing a concrete source of inspiration and encouragement through examples and illustrations. Hence, rather than constraining the options available to the end-users with paternalistic policies, our account of MCR aspires to recommend general ethical exhortations that can inform robot design (for makers and programmers), codes of conduct and etiquette standards for HRI (for private employers and public officers), and representation of HRR in media and pedagogically sensitive contexts (for film-makers, script-writers, and educators). These guidelines suggest a self-regulatory framework that these actors should follow willingly to meet the highest standards of moral decency and professional credibility.

Finally, our approach justifies the need to address the apparently ineliminable contradictions contained in the ADP. These contradictions seem to be able to corrupt the root of any HRR, menacing the very ideal of social robotics. Finding an alternative model of social recognition that enables a healthier and more balanced relationship between humans and robots—without reproducing the alienating representations of slavery and exploitation—should be the object of future philosophical investigations.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Aristotle (1988). *Ethica Nichomachea*. In I. Baywater (Ed.), Oxford University Press.
- Athanassoulis, N. (2000). A response to Harman: virtue ethics and character traits. *Proceedings of the Aristotelian Society*, 100, 215–221.
- Ball, A., Silvera-Tawil, D., Rye, D., & Velonaki, M. (2014). Group comfortability when a robot approaches. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social robotics* (Vol. 8755, pp. 44–53). Cham: Springer.

- Brown, A. (2015). To mourn a robotic dog is to be truly human. *The Guardian*. Thu 12 Mar 2015. Published online: <http://www.theguardian.com/commentisfree/2015/mar/12/mourn-robotic-dog-human-sony>. Retrieved 29-5-2018.
- Brownstein, M., & Madva, A. (2012). Ethical automaticity. *Philosophy of the Social Sciences*, 42(1), 68–98. <https://doi.org/10.1177/0048393111426402>.
- Bryson, J. J. (2010a). Robots should be slaves. In Y. Wilks (Ed.), *Natural language processing* (Vol. 8, pp. 63–74). Amsterdam: John Benjamins. <https://doi.org/10.1075/nlp.8.11bry>.
- Bryson, J. J. (2010b). Why robot nannies probably won't do much psychological damage. *Interaction Studies*, 11(2), 196–200. <https://doi.org/10.1075/is.11.2.03bry>.
- Carr, L. (2018). *On what grounds might we have moral obligations to robots?* Retrieved from: <https://www2.rivier.edu/faculty/lcarr/OUR%20MORAL%20OBLIGATION%20TO%20ROBOTS.pdf>. Accessed 25 May 2018.
- Coeckelbergh, M. (2010a). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221. <https://doi.org/10.1007/s10676-010-9235-5>.
- Coeckelbergh, M. (2010b). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3), 235–241. <https://doi.org/10.1007/s10676-010-9221-y>.
- Coeckelbergh, M. (2018). Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos. Journal of Philosophy & Science*, 20(1), 141–158.
- Collins, S. (2004). Moral virtue and the limits of the political community in Aristotle's *Nicomachean Ethics*. *American Journal of Political Science*, 48(1), 47–61.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.00468>.
- Darling, K. (2012). Extending legal rights to social robots. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2044797>.
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3–4), 177–190. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3).
- Dumouchel, P., & Damiano, L. (2017). *Living with robots*. (M. B. DeBevoise, Trans.). Cambridge: Harvard University Press.
- Fessler, L. (2017). We tested bots like Siri and Alexa to see who would stand up to sexual harassment. *Quartz*. Retrieved from: <https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/>. Accessed 25 May 2018.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.
- Gerdes, A. (2016). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers and Society*, 45(3), 274–279. <https://doi.org/10.1145/2874239.2874278>.
- Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing test. *Journal of Information, Communication and Ethics in Society*, 13(2), 98–109. <https://doi.org/10.1108/JICES-09-2014-0038>.
- Gini, G., Albiero, P., Benelli, B., & Altoè, G. (2007). Does empathy predict adolescents' bullying and defending behavior? *Aggressive Behavior*, 33(5), 467–476. <https://doi.org/10.1002/ab.20204>.
- Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2), 101–124. <https://doi.org/10.1080/1047840X.2012.651387>.
- Gunkel, D. J. (2017). The other question: can and should robots have rights? *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-017-9442-4>.
- Harman, G. (1999). Moral philosophy meets social psychology. *Proceedings of the Aristotelian Society*, 99, 315–331.
- Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society*, 100(223–2), 26.
- Hauskeller, M. (2014). Sexbots on the rise. In M. Hauskeller (Ed.), *Sex and the posthuman condition* (pp. 11–23). London: Palgrave Macmillan. https://doi.org/10.1057/9781137393500_2.
- Hauskeller, M. (2016). Automatic sweethearts. In *Mythologies of transhumanism* (pp. 181–199). Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-39741-2_10.
- Honneth, A. (1995). *The struggle for recognition: the moral grammar of social conflicts*. (J. Anderson, Trans.). Cambridge: MIT Press.
- Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.
- Ikäheimo, H., & Laitinen, A. (2007). Analyzing recognition: Identification, acknowledgement and recognitive attitudes towards persons. In B. van den Brink & D. Owen (Eds.), *Recognition and power* (pp. 33–56). Cambridge: Cambridge University Press.
- Ikäheimo, H., & Laitinen, A. (Eds.). (2011). *Recognition and social ontology*. Leiden: Brill.

- Küster, D., & Świdarska, A. (2016). Moral patients: what drives the perceptions of moral actions towards humans and robots? In J. Seibt, M. Nørskov, & S. Schack Andersen (Eds.), *Frontiers in artificial intelligence and applications* (Vol. 290, pp. 340–343). <https://doi.org/10.3233/978-1-61499-708-5-340>.
- Laitinen, A. (2002). Interpersonal recognition: A response to value or a precondition of personhood? *Inquiry*, 45(4), 463–478.
- Laitinen, A. (2016a). *Should robots be electronic persons or slaves?* Retrieved from: <https://www.finsif.fi/should-robots-be-electronic-persons-or-slaves/>. Accessed 25 May 2018.
- Laitinen, A. (2016b). Robots and human sociality: Normative expectations, the need for recognition, and the social bases of self-esteem. In J. Seibt, M. Nørskov, & S. Schack Andersen (Eds.), *Frontiers in artificial intelligence and applications* (vol. 290, pp. 313–322). <https://doi.org/10.3233/978-1-61499-708-5-313>.
- Laitinen, A., Niemelä, M., & Pirhonen, J. (2016). Social robotics, elderly care, and human dignity: A recognition-theoretical approach. In *Frontiers in artificial intelligence and applications* (pp. 155–163). <https://doi.org/10.3233/978-1-61499-708-5-155>.
- Levy, D. (2007). *Love and sex with robots: The evolution of human–robot relationships*. New York: Harper-Perennial.
- Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics*, 1(3), 209–216. <https://doi.org/10.1007/s12369-009-0022-6>.
- Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice*, 3(4), 365–383. <https://doi.org/10.1023/A:1009926720584>.
- Murphy, R. R. (2018a). Westworld and the uncanny valley. *Science Robotics*, 3(17), eaat8447. <https://doi.org/10.1126/scirobotics.aat8447>.
- Murphy, R. R. (2018b). Parents, rejoice: Alexa will now remind kids to say “please”. *Quartz*, 25 April 2018.
- Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, 17(3), 347–369. <https://doi.org/10.1075/is.17.3.02nom>.
- Nørskov, M. (2014). Human-robot interaction and human self-realization: reflections on the epistemology of discrimination. In J. Seibt, R. Hakli, & M. Nørskov (Eds.), *Frontiers in artificial intelligence and applications* (Vol. 273, pp. 319–327). <https://doi.org/10.3233/978-1-61499-480-0-319>.
- Nørskov, M. (Ed.). (2016a). *Social robots: boundaries, potential, challenges*. London: Routledge.
- Nørskov, M. (2016b). Technological dangers and the potential of human–robot interaction: a philosophical investigation of fundamental epistemological mechanisms of discrimination. In M. Nørskov (Ed.), *Social robots: boundaries, potential, challenges* (pp. 99–122). London: Routledge.
- Oakley, J., & Cocking, D. (2008). *Virtue ethics and professional roles*. Cambridge: Cambridge University Press.
- Peeters, A., & Haselager, P. (in preparation). *Designing virtuous sex robots*.
- Rietveld, E. (2008). Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind*, 117(468), 973–1001. <https://doi.org/10.1093/mind/fzn050>.
- Rodogno, R. (2016). Robots and the limits of morality. In M. Nørskov (Ed.), *Social robots: boundaries, potential, challenges* (pp. 39–56). London: Routledge.
- Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 205–222). Cambridge: MIT Press.
- Seibt, J., Hakli, R., & Nørskov, M. (Eds.). (2014). Sociable robots and the future of social relations. *Frontiers in Artificial Intelligence and Applications*, 273.
- Sparrow, R. (2012). Can machines be people? Reflections on the Turing triage test. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 301–315). Cambridge: MIT Press.
- Sparrow, R. (2016). Kicking a robot dog. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 229–229).
- Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, 9(4), 465–477. <https://doi.org/10.1007/s12369-017-0413-z>.
- Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, 5, 15924. <https://doi.org/10.1038/srep15924>.
- Tavani, H. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. *Information*, 9(4), 73. <https://doi.org/10.3390/info9040073>.
- Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & SOCIETY*, 22(4), 495–521. <https://doi.org/10.1007/s00146-007-0091-8>.
- Truong, A. (2016). Parents are worried the Amazon Echo is conditioning their kids to be rude. *Quartz*, 9 June 2016.

- Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. *Philosophy & Technology*, 28(1), 107–124.
- Vallor, S. (2016). *Technology and the virtues: a philosophical guide to a future worth wanting*. Oxford: Oxford University Press.
- Waytz, N., Epley, A., Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. In *Psychological Review Copyright 2007 by the American Psychological Association*, 2007 (Vol. 114, No. 4, pp. 864–886).
- Whitby, B. (2008). Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. *Interacting with Computers*, 20(3), 326–333. <https://doi.org/10.1016/j.intcom.2008.02.002>.
- Whitby, B. (2010). Oversold, unregulated, and unethical: why we need to respond to robot nannies. *Interaction Studies*, 11(2), 290–294. <https://doi.org/10.1075/is.11.2.18whi>.
- Whitby, B. (2012). Do you want a robot lover? The ethics of caring technologies. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 233–248). Cambridge: MIT Press.
- You, S., Nie, J., Suh, K., & Sundar, S. S. (2011). When the robot criticizes you...: self-serving bias in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 295–296). <https://doi.org/10.1145/1957656.1957778>.
- Zagzebski, L. (2010). Exemplarist virtue theory. *Metaphilosophy*, 41(1–2), 41–57. <https://doi.org/10.1111/j.1467-9973.2009.01627.x>.
- Zwolinski, M., & Schmidt, D. (2013). Environmental virtue ethics. In D. C. Russell (Ed.), *The Cambridge companion to virtue ethics* (pp. 221–239). Cambridge: Cambridge University Press.

Affiliations

Massimiliano L. Cappuccio^{1,2} • Anco Peeters³ • William McDonald²

¹ School of Engineering and Information Technology, University of New South Wales, Northcott Dr, Campbell, ACT 2612, Australia

² College of Humanities and Social Sciences, United Arab Emirates University, 15551, Al Ain, Abu Dhabi, United Arab Emirates

³ Faculty of Law, Humanities and the Arts (19.1066), University of Wollongong, Wollongong, NSW 2522, Australia

Support Publication #19 [SP19]

Sandoval E.B., Billinghamurst M., Cappuccio M.L. (2024),
“Robot Natives: Ethical Design and Regulations for Baby-Robot Interaction”
in *The Cambridge Handbook on Law, Policy, and Regulations for Human-
Robot Interaction*, Cambridge University Press.

Reproduced with permission from Cambridge University Press.

All rights reserved.

Reproduced with permission from Cambridge University Press.
All rights reserved.

Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

Robot Natives: Future Design and Regulations for Baby Robot Interaction.

Eduardo B. Sandoval
Mark Billingham
Massimiliano Cappuccio

Keywords: Robot natives, baby robot interaction, empathic computing, empathic agents, children robot interaction.

Abstract

This chapter explores design guidelines and potential regulatory issues that could be associated with future baby-robot interactions (bHRI). We coin the term “robot natives”, which we define *as the first generation of human’s regularly interacting with robots in domestic environments*. This term includes babies (0–1 year-old) and toddlers (1-3 years old) born in the 2020s. Drawing on the experience of other interactive technologies becoming widely available in the home and the positive and negative impact they have on humans, we propose some insights into the design of future scenarios for baby robot interaction, aiming to influence the future legislation regulating service and social robots used with robot natives. Similarly, we aim to inform designers and developers to inhibit robot designs that can negatively affect the long-term interactions for the robot natives. We conclude that a qualitative, multidisciplinary, ethical, human-centred design approach should be beneficial to guide the design and use of robots in the home and around families as this is not a common approach in the design of studies in child robot interaction (cHRI).

1 Introduction

This chapter explores the possible design guidelines and regulatory issues that could be associated with future baby-robot interactions. Inspired by the term “digital natives” we coin the term “robot natives”, that we define *as the first generation of human’s regularly interacting with robots in domestic environments*. These robot natives are comfortable with interacting with and using robots in their everyday lives. They may also be interested in going beyond that interaction by programming the robot or experiment with the autonomous behaviour of the robots around them.

Our proposition of the “robot native” term is justified by two conditions: (a) The number of robots that we have in domestic environments is increasing; (b) we have a complete cohort of babies and toddlers who, being born with service robots at home, have experienced them as social agents. These babies and toddlers are the robot-natives.

Drawing from the experience of other interactive technologies becoming widely available at home and the positive and negative impact this has on humans, we propose the following research questions that we aim to address in the Conclusion section of this chapter:

1. What will make robots suitable for baby use? What distinguishes them from interactive toys and robot toys?

2. Are our children and us ready to use robots and take advantage of robot technology in domestic, educative, and work environments?
3. What are the potential positive and negative outcomes that can be produced by robots interacting with children?
4. What are the design guidelines required for a successful, engaging, and safe baby robot interaction?
5. What regulations should be put in place to ensure safe child-robot interactions?

These questions aim to encourage robot regulations and design guidelines for the robot natives.

In this chapter, we provide the context that motivates the conditions for the robotic natives in section 2. We first provide some theoretical context, then we survey the literature related to children-robot interaction and the latest advances in this area, and eventually we move to the more speculative domain of baby robot interaction. In section 3, we propose some insights into the design of future scenarios for baby robot interaction. What robots the robotic natives will use and how these robots could affect their experience are central topics in our discussion. In section 4, we present considerations for future legislation and regulation. Conclusions and directions for future work are presented in section 6. Our main goal in this chapter is to minimise the negative impact of robot technologies on these robotic natives by early regulation and standardisation. We suggest that regulation of any new interactive technology should be evidence-based, collegiate, multidisciplinary, democratic, inclusive and use a participative process of design.

2 Theoretical Background

2.1 Robot Natives

Over twenty years ago, Presky (Prensky, 2001) proposed that young people who are highly technologically literate and engaged could be defined as “digital natives”. These are people who have since become very comfortable with a wide range of digital interactive technologies such as the internet, smart phones, streaming content, videogames, and even Virtual Reality (VR). This term is closely associated with the age of the people born as millennials or generation Z. Although the term suggests that age is central to become a digital native, evidence suggests that exposition to technology and level of education are more significant than age or generational gap in the use of interactive technologies (Bax, 2011; Helsper & Eynon, 2010).

In our case, we coined the term Robotics natives to describe the new generation of babies and toddlers who are witnessing the regular use of service robots in domestic environment beyond educational robots and toys since they are born. In contrast with the digital natives, the robot natives do not require advanced cognitive skills as reading and writing. Babies and toddlers can interact with service and social robots at a very basic level as both share the physical space. This is what makes the concept of robot natives consistent, broad, and useful.

We consider the term robot native convenient as the concept is embedded in popular culture and shown by the normalization of human/children/infant robot interaction in

science fiction films like Ron's Gone Wrong, Star Wars, Bicentennial Man, Wall-E, Chappie, Surrogates, Ex-Machina, I am mother, Blade runner, A.I. Artificial Intelligence, Big Hero 6, and many others. In that sense, the term is useful as a benchmark and reference for academic purposes.

2.2 Service Robots and Social Robots

For this text, robots are computers with a certain level of autonomy and embodiment capable to interact with the physical world. These robots can also be more purpose oriented, such as industrial robots, service robots, and social robots. In this chapter, we focus only on the domain of service robots and social robots.

Service robots are defined by the ISO 8373:2021 standard as robots "that perform useful tasks for humans or equipment excluding industrial automation applications" (ISO, 2021). At the early stage of the discipline, social robots were defined as *"physical entities embodied in a complex, dynamic, and social environment sufficiently empowered to behave in a manner conducive to their own goals and those of their community"* (Duffy et al., 1999). A more state of art definition of social robots is: *"... robots that interact with humans and each other in a socially acceptable fashion, conveying intention in a human-perceptible way, and are empowered to resolve goals with fellow agents, be they human or robot"* (Daily et al., 2017)."

As we have witnessed recently, service robots are becoming more visible in certain domains. Robotic technologies are slowly leaving industry and entering public places, schools, and homes. According to the International Robotics Federation, the use of service robots is increasing significantly, with over 18.5 million domestic robot cleaners sold in 2020 (+6%), worth 4.3 billion US dollars (IFR, 2021). Similar phenomena can be observed in relation to hospitality robots (+196%), professional cleaning robots (+92%), and medical robots (55% of the service robots' market).

However, more sophisticated social robots capable of natural interaction with humans are less popular among the public, and remarkable commercial failures have occurred, delaying their adoption (Hoffman, 2019). Despite this, we have been witnessing several advances in human-robot interaction (HRI), cognitive robotics, artificial intelligence, empathic computing, and other related areas. It is expected that, in a few decades, driven by industry 5.0, we will see service and social robots outside of industrial environments interacting with humans in a natural way.

In our view, every robot in a domestic environment or exposed to the public can display a behaviour and function as a service robot or social robots, as humans tend to anthropomorphise machine behaviours. Similarly, we must highlight that we focus on service and social robot technologies as they are emerging and becoming popular among the broader public, but it will take some time for them to become widely used in uncontrolled spaces, such as schools, homes, and public places. Robots navigating freely in these spaces represent an enormous technical challenge. Moreover, they could present a certain risk or danger for humans to be avoided at any cost.

2.3 Sociocultural context and Motivation

In 2014, Cynthia Breazeal claimed that “[...] *we’re so specialized for social interaction. So, if you can really design robots that can interact with people, in this very natural, interpersonal way, I think that would be great.* [...]”. However, this design goal has been mainly focused on adults with highly developed social skills. Robots in domestic environments face different design challenges for users as diverse as babies and their parents. It seems that manufacturers have not considered future interactive scenarios involving babies and toddlers when using robots in domestic environments, as we discuss in section three. These toddlers and babies are developing social, cognitive, and physical skills around the robots. How will robots affect this infant development in the long term?

If these service/social robots take a place in public, domestic, education, or healthcare domains, children who were born in the beginning of the 21st century will be adults by 2050. Throughout their lives they would have been progressively exposed to robots and their technological enhancement. Similarly, to other interactive technologies such as the internet, videogames, or mobile phones, robots will have a steady development. Many of the children of the digital native generation will be raised in an environment where robots in the workplace and home are commonplace. These children will be mainly in countries with developed economies.

Robot Natives, in comparison with digital natives, will face different challenges related to our research questions. Investigating the specific needs of Robot Natives is timely due to:

- **The number of service robots sold worldwide:** We are starting to see an increasing number of these robots in the home environment (Haegele & International Federation of Robotics. Statistical Department, 2016). They have reached a critical price point beyond early adopters and now a broader population can afford them.
- **Future interactive scenarios:** Robot manufacturers of service and social robots have not fully spotted the potential scenarios described in the early stages of HRI literature. We consider that experts in HRI and other related disciplines should be involved in the design of novel interactive scenarios before robot production to avoid negative impact on humans.
- **There is a critical number of robotic natives now:** Global population in certain areas tends to be younger and larger than in the past. For instance, in the USA young people born in the last decade of the XX century is overtaking baby boomers as the largest generation. The robot natives will be exposed to more robots in the future and access more educative resources in certain parts of the world. We consider that a critical number of young people in developing countries will increase the development of robot technologies with new companies and research. Hence, intensive use of service robots and social robots can be correlated with the increasing population.
- **The long-term impact of robots on humans is unknown:** The state-of-art of human-robot interaction research is still very oriented to controlled conditions and short-term experimental interactive scenarios. Our proposal aims to encourage more qualitative research in naturalistic conditions.

- **Emerging industries such as robots in retail stores, and robots preparing and serving food:** We are witnessing an emerging industry of human-robot collaboration in these specific domains. Although robots have limited skills, manufacturers have spotted specific tasks that robots can perform supporting human activity.

We must highlight two key facts that motivate this chapter:

1. **Baby robot interaction is a novel and underexplored topic in HRI.** The intrinsic complexity of the topic has delayed research in this area, and most human-robot interaction practitioners focus on other critical topics. To the best of our knowledge, there are several large projects focused on children robot interaction in Europe as (EU CORDIS Horizon 2020, 2018; EU CORDIS, Horizon 2020, 2019), but none focused on baby/infant robot interaction (0-1 and children under 3 years old).
2. **Even for regular social and service robots there is no regulation for interaction with them as the HRI is an emerging research area.** Certainly, there are ISO norms and official guidelines for the design of service robots but not for social robots.

From our perspective, robots will not be as disruptive a technology as most of the public think. As with previous industrial and information revolutions, the disruptions created by that robots will take time. The design of robots interacting with the environment and humans in uncontrolled conditions present challenges that other interactive technologies did not face. To illustrate, designing for screens is not the same as designing for robots which will move around, touch things, grab things, and converse with humans. Certainly, the process of designing robots adds additional challenges beyond the artificial intelligence and the social skills that agents as Siri and Alexa have. The actions of robots operating in the physical world can have far more unpredictable consequences than AIs interacting through the screens. There are concerns about the safety of robots manipulating objects around humans as well as humans behaving unpredictably in social context, due to the design challenges that this implies.

2.4 Previous Impact of other Interactive Technologies

In a similar way to widespread use of other interactive technology, as robots enter our lives there will be some concerns that need to be addressed. As an example of unintended technology consequences, we can consider screen-based devices (mobiles, computers, tables, and smart TVs).

In developed countries, screens are now so inexpensive and convenient for people to use that they have become widespread. The average number of screens in a typical Australian household is 6.6 hours, including TV, PC/ Laptops, Mobile phones, and tablets) (ThinkTV, 2021), and they used for an average of 5.5 hours during the day (Fenech, 2021, p. 5). During the recent COVID pandemic screen devices were intensively used to tackle social isolation and keep people connected. While the benefits of screen devices are many and varied (LeBlanc et al., 2017), new evidence suggests that the extensive use of interactive technologies can lead to an overload of information rather than to better cognitive skills. Screens have brought enormous benefits in education, information access, remote communication, and entertainment for a vast majority but some considerations should be done.

For instance, screen devices also have negative effects in humans when overused. The overuse of screen devices also can create different problems in children, such as misuse of social media and video games (Jo et al., 2021a) (Jo et al., 2021b) and long-term physiological effects such as poor sleep quality, overweight/obesity, poorer academic performance, among others (Mineshita et al., 2021; Ratan et al., 2021) (Liu et al., 2021). Ergonomic issues such as bad posture and chronic back and neck pain are reported using smartphones, tablets, and laptops for several hours (Betsch et al., 2021; Sekiguchi et al., 2018).

Similarly, there is no evidence that the skills developed in the digital world can be directly transferred to the real world. For instance, certain skills (eye-hand coordination, X, Y etc) developed and trained in virtual simulators can be just used there and not translated to the real world (first-person shooters, driving, piloting, etc) (Desmurget, 2020). This lack of direct transfer of certain skills is due to virtual simulators and videogames not having all the stimuli that the real world has for humans. However, we acknowledge that other skills (teamwork, strategic skills and other “soft skills”) can be transferred to the real world in several cases, but these can be acquired by other means, such as board games, dramatic playing, and storytelling by numerous studies in playing and ludic activities validating them. I.e., (Koven, 2013; Schell, 2014).

Regulation and design guidelines for baby-robot interaction are relevant to avoid future negative consequences in the use of service and social robots in the future as happening with other previous technologies. Similarly, the significance of discussing the future of baby robot interaction is critical as robots could complement or substitute the use of screen-based devices for several tasks. Replacing screen devices with embodied robots mean that more natural interaction closer to human-human interaction is expected. These future robots could be the ultimate expression of artificial intelligence as intelligent beings interact with their environment in the long term.

We are starting to see early steps in this direction in the form of education robots, robot toys, robot cleaners, robot pets, and companion robots, among others. Further out are the robot butlers, the robotic romantic partners, and even dystopian visions of robots taking control of humanity as portrayed in science fiction (B. Sandoval, 2014; Dautenhahn, 1998; Mubin et al., 2015; Sandoval & Mubin, 2015).

With this chapter, we aim to frame technologies correctly for the benefit of users and balance our argument for it with existing commercial arguments (short-term approach, lack of proper validation, arguing moral neutrality). We also aim to influence the design process and policy making of future robot technologies, highlighting the negative consequences of overuse, lack of regulation, and non-critical optimism. In other words, we want to keep our distance of the technology development approach generated in the mainstream producers of them.

2.5 Previous work in Children Robot Interaction

Robot natives will be highly affected by novel findings in human robot interaction. Children Robot Interaction (CHRI) will be an increasing relevant research area once a critical mass of

researchers notices that there is a gap in the literature regarding the topic. Similarly, social robots are available for researchers. See Figure 1.

There is a significant body of research in children-robot interaction, the topic has been central in HRI since the early days of the discipline. First mentions of the relevance of exploring cHRI can be found in (Karwowski & Rahimi, 1992). Researchers as Dautenhahn and Kanda have generated a considerable amount of work related to different aspects of children robot interaction since the 2000s. Recent papers from Kanda and his team illustrate unforeseen aspect of these interactions as abusive behaviour from the children towards the robot when adults do not supervise the interaction (Nomura et al., 2016) and his early work explored the potential use of interactive robots as peer tutors and social partners for children (Kanda et al., 2004). Similarly, researchers as Dautenhahn and her team have done early explorations aiming to use interactive robots in benefit of children with autism (Dautenhahn, 1999) and more recently they have tested the feasibility of robot interventions helping anxious children and adolescents (Rasouli et al., 2022).

More recently, literature reviews as the one written by (Belpaeme et al., 2013) illustrate the area of opportunities for the use of robots with children. Belpaeme et. al. discuss the use of interactive robots in education, robot assisted therapy (in substitution to animal assisted therapy), and edutainment. Certainly, these areas represent business opportunities and there are continuous explorations using social robots interacting with children.

Inspection to the literature suggests that there is a broad diversity in terms of research goals and methods in the studies related to cHRI. To illustrate, a robot as a tool to boost new ideas and stimulate children's creativity (Leite et al., 2017), social robots assessing children emotional state via analysing psychophysiological signals (Filippini et al., 2021), children's perceptions towards robots (Beran et al., 2011; Lin et al., 2009; Sandoval & Penaloza, 2012), the use of humanoid robots and storytelling as therapeutic tool for autistic children (Beran et al., 2011) or the design of huggable robots for paediatric care (Jeong et al., 2015).

Similarly, the variety of robots used for cHRI is broad. Sharkey and Sharkey described several robots used for cHRI or repurpose from ageing care for the early stages of cHRI (Sharkey & Sharkey, 2011). We are not providing detail on these robots as the diversity of robots is broader nowadays and some of them are already discontinued. An extensive list of robots for research, entertainment, education, and service can be found in (IEEE, 2022) and the Anthropomorphic Robot Database (ABOT) (Phillips, E. et al., 2018). It must be highlighted that a significant amount of the research in cHRI is oriented to the use of humanoid, general purpose research platforms with few zoomorphic exceptions as it is shown in Figure 1.

2.6 Baby Robot Interaction

Baby robot interaction (bHRI) should be conceived as an essential subset of cHRI research. However, after survey the literature, we observe a relevant gap in the cHRI studies related to the youngest group of research in cHRI. Inherent complexities in bHRI have made it difficult to research, including the management of infants as participants of HRI studies and the possible fear of the parents exposing their babies to robots. There are very few studies related to bHRI and it remains an unexplored area.

Some of the few studies in bHRI include using robots to encourage infant leg-motion training aiming optimal neurodevelopmental outcomes via reward learning and imitation complementing human therapy (Fitter et al., 2019). Also, a thermal infrared imaging technique has been used to assess autonomic response in babies during the interaction with a robot and an avatar (Filippini et al., 2021). Similarly, there have been efforts to design and test tele-presence robot for interacting with babies and toddlers focused on the use of other members of the family as grandparents (Abe et al., 2018; Shiomi et al., 2016).

There is an important and broad opportunity for junior researchers to explore bHRI. bHRI offers fascinating and relevant challenges for researchers. For instance, when researchers study robots and babies, long-term cognitive development and social impact should be factors to be investigated carefully to educate robot designers.



a) Papero (NEC) b) IfBot (Business Design Labs) c) iRobi (Yujin Robots)



d) Kompai (Robotsoft) e) Qrio (Sony) f) Nao (Softbank) g) Pepper (Softbank)



h) Huggable (MIT)

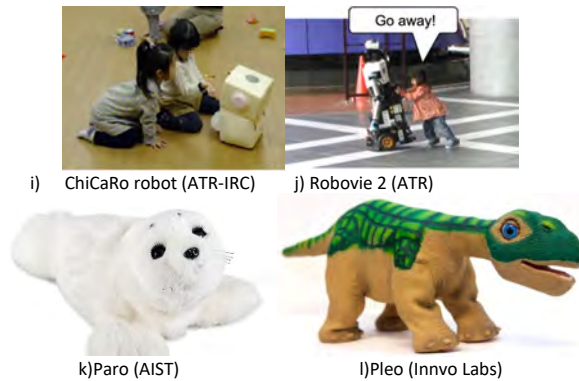


Figure 1: Examples of robots used in children robot interaction research. All the rights to their respective owners.

3 Design Recommendations for bHRI

When surveying the studies in this area, we realised that the term used for “baby/infant robot interaction” has been used to explore and study cognitive robotic architectures, robots designed as a babies, and machine learning systems inspired by baby learning and tested (Zehfroosh & Tanner, 2022) but there is no evidence that human-centred design approaches are taken to develop future devices for robot natives. Such a lack of human-centred design studies can raise some of the ethical questions as the ones related to cHRI (Sharkey & Sharkey, 2011) but applied to robot natives. Certainly, we do not aim to prescribe a methodology for the design of robots. However, we can point out the conditions in the state-of-art of cHRI that could be re-assessed to inform and educate better design decision in the development of robots for babies and toddlers. We raise these suggestions to encourage a safe, engaging, and effective future bHRI.

3.1 Robots as General Research Platform

If possible, it could be convenient to use and design ad-hoc robots for specific interactive scenarios. Currently, the literature suggests that generalist, research platforms are used to perform cHRI. This is possible not the right approach to make robot technologies broadly used by the public. Products require a purpose, instead of trying to fit a robot in multiple scenarios, researchers could explore potential and appropriate interactive scenarios in which service and social robots add value to the life to robot natives. In Figure 2 we suggest multiple scenarios that can offer opportunities for novel service and social robots. Similarly, Figure 2.I a) illustrate an scenario that has not been explored enough: the interaction of babies around domestic service robots when they are working. This scenario is possibly the most common bHRI and there are no reports of this already existing interactive scenario. Shortly, we will publish a paper in regards of this scenario and its design opportunities.

3.2 A different experimental philosophy

We consider that there are two fundamental actions in the design of future baby interaction: a) The use of a human centred design approach. and b) Baby-robot interaction

research with proper context. For a Human centred design approach, that includes four fundamental factors to be considered when human robot interaction is designed: Users, robot, task, and environment, we suggest the use of design frameworks and design principles for HRI as proposed by previous research (Bartneck & Forlizzi, 2004). For instance, for selection of robot interfaces (Gonzalez-Jimenez et al., 2022), robot behaviours (Huang & Mutlu, 2012), domestic robot ecologies (Sung et al., 2010), design principles for better design of robots (Sandoval et al., 2018a), design inclusive criteria for robots and robot ergonomics (Mohan et al., 2014; Sosa et al., 2018).

Similarly, bHRI implies several research constraints. Perform controlled experiments on babies involve ethical considerations in regards the well-being of the baby, the parents, and the long-term effect of those interactions. To illustrate, few years ago ergonomics (Sharkey & Sharkey, 2011) suggested that anthropomorphism and deception can be morally questionable when robots interact with elderly and children. Traditional empirical controlled-condition experiments could not necessarily capture and describe the human experience of robot natives fully when these studies rely only in metrics and are data driven. We suggest this future research in bHRI could be performed in naturalistic conditions at the real homes of the users, as this is critical for the success of future service and social robots. Similarly, it can be claim that the studies found in our literature review do not consider domestic robots as cleaners which show certain level of autonomy and already interact with several babies at some homes. It could be necessary to explore these scenarios initially in order to generate insights for more complex future bHRI.

In other words, into-the-wild bHRI cannot be easily objectified and reduced to controlled experiments. bHRI is an intrinsically difficult topic to study for ethical and logistic reasons. HRI and particularly bHRI is a phenomenon that is happening in complex ecosystems including humans, environment, objects as toys, and other interactive devices. To illustrate, babies and parents keep interacting when the adults perform a video call or look for a playlist while holding the baby. In other moments, babies and toddlers continuously explore their environment and interact with Bluetooth speakers, keyboards, interactive toys, toy musical instruments, large appliances such as laundry machines or mobile devices that they accidentally reach. All these artifacts with buttons and screens offer interactive possibilities to the robot natives. Possibly, observation methods as suggested in (Ciesielska et al., 2018) could be potent and suitable methods to explore bHRI on all its complexity. We consider that an observational approach in naturalistic conditions could significantly help to define design guidelines and legislation for bHRI shortly. This observational and naturalistic approach is also essential for the success of future service and social robots.

3.3 Tuned Design of bHRI Scenarios

Because of this observational, human centred approach, the good and detailed design of interactive scenarios for bHRI is required. That would help to minimise the raise of ethical and moral concerns when bHRI could happen. It also could help to align the multiple factors involving (users, environment, artifacts) in the interactions that will take place along the life of the robot natives. The historical records of previous interactive technologies could help us to visualise potential outcomes for service and social robots. In a hypothetical exercise, we present a selection and definition of bHRI scenarios aligning these factors in Figure 2.

3.4 No Tecno-solutionism

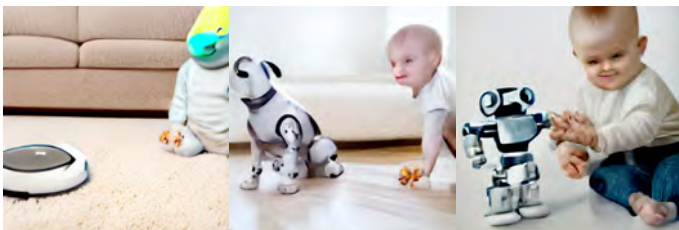
Babies (0-1 years old) and toddlers (1-3) are inherently social, so the relation of the robot with the child tends naturally to become social. As we suggested in previous sections, future robot natives will be born immersed in a world populated by robots, a constant presence around them at home.

It is tempting to sell the idea of robots taking care of a baby to release free time to the parents. Many baby products sell this idea including smart rocking cradles to put babies to sleep without considering the importance of this activity for human bonding.

These types of ideas have a strong techno-solutionist approach coming from the idea that for a every human problem there is a technical solution that can sort it out. We have seen this approach in many different interactive technologies in the last decades. These kinds of scenarios raise ethical questions in regards of the use of robots: What kind of robots we want our babies to interact with? Do we want social robots interacting with babies as if they were substitutes of their parents? Do we want these robots to specifically interact with babies for utilitarian reasons (e.g., parsimony, efficient time use)? Some interactive scenarios would be problematic as robot caregivers for bHRI as we illustrate in Figure 2.IV.

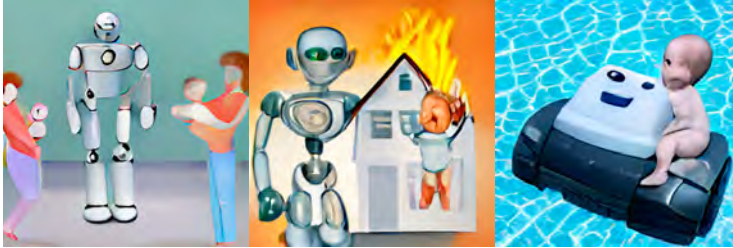
In our view, we need to overcome the techno-solutionist mentality, as it does not provide appropriate guidelines to the ethical design of the robots. Although they aim for good as they are meant to enhance human conditions, we require to unlearn these solutionist attitudes, as well as the biases and impulsive design decisions that they carry with them, to discover a more humanistic and multidisciplinary approach to the development of interactive technologies specifically designed for the robot natives.

Very often in HRI we see that the focus lies in the resolution of the technical challenges and the measurement of short term interactions in controlled scenarios. However, as the robot natives lack any formal knowledge about or scientific understanding of technology, they are likely to misjudge the abilities of robots, especially if they display some of the features of humans or animals. Also, the adult public requires education to avoid practices and routines that may lead babies and toddlers to spend too much time with robots. The overuse of robots could reduce the essential time babies and toddlers get to spend in the company of their primary caregivers (e.g., the mother, father, or grandparents), thus obstructing the affective development of the infant.



Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

- I. Baby Robot Interaction scenarios already happening: a) Robot cleaners working around babies b) Robot pets as AIBO interacting with babies and toddlers c) Robot edutainment devices interacting with babies.



- II. Short-term potential and desirable baby-robot interaction: Interactions that can be potential business cases according to the current state of art of service and social robots: a) Telepresence robots encouraging remote interaction with other members of the family with restricted mobility. b) Robots for baby safety and quick action in emergency scenarios c) Robots for safe outdoors interaction.



- III. Future baby robot interactions in which robots support adult interaction with the baby: a) Robots supporting teachers for basic learning experiences as singing, vowel learning and others. B) Robots for passive supervision of baby groups. c) Robots as proxies supporting interactive playing among parents and babies.



- IV: Questionable and ethically ambiguous bHRI: Activities that strength human bonds and develop social, cognitive, and emotional skills in parents and babies.

Figure 2. Future baby robot interaction scenarios requiring a multidisciplinary discussion regarding the ethical considerations. All the images created using Dall-e website.

4 Potential Regulations

As other consolidated areas in robotics, regulations are required for the design, development and commercialisation of service and social robots aiming to interact with robot natives. We propose some considerations that could be essential for healthy, sustainable, and ethical bHRI.

4.1 No substitution for Human-human interaction

Children cannot be raised by robots. As other previous interactive technologies, they should not be aimed to be a substitute of human-human interaction at that early age. Robots as other technologies have an impact in the long term development of kids. As we mentioned in section 2, negative impact coming from the overuse of any device could cause negative impact in the robot natives.

4.2 Ethical Evaluations and Extensive Testing and Validation

It is not clear if industry discusses the long-term impact of their devices and innovations. Due to the characteristics of service and social robots in the behaviour and well-being of babies and toddlers (particularly in their social, emotional, and cognitive development) a qualitative multidisciplinary assessment from an external or internal group of experts would be required before release robots to the broad public. Universities, research centres, government, and industry-peers should take part of the creation of these ethical protocols looking for good practices in the design of service and social robots. An obvious situation that should be avoid is robots (cleaners and others) hurting babies in their way while trying to complete a cleaning job.

This could not be a welcome and popular recommendation for the industry, but babies and toddlers are fragile and essential for the humankind. Researchers, and industrial designers should take a virtuous ethical posture for their designs as we suggest in (Cappuccio et al., 2021) when they develop a robot.

4.3 No Vigilant, Deceptive and Manipulative Robot Behaviours

As services and social robots can influence the behaviour of babies and toddlers due to their social behaviour, anthropomorphism (Sharkey & Sharkey, 2011; Zlotowski et al., 2015), potential advanced AI, and cameras and sensors, it is important to moderate the impact the robots have on robot natives. We can learn from other artificial agents already used in domestic environments (Chubb et al., 2022; Hennig, 2018) and avoid future negative deceptive scenarios.

Likely, service, and social robots interacting with robot natives will use cameras among other sensors. Privacy of babies and toddlers is something that by default is considered as sensitive and should be keep regulated. However, more complicated interactions can emerge when the robot capturing babies' behaviours through cameras and sensors and reacts consequently trying to develop empathy with the young user. Moreover, robots, if

not well designed (even with good intentions), can influence the behaviour of babies and toddler. Robots can be a vigilante which can conditionate behavioural responses at early age. That represents a significant risk in ethical terms as human being should be autonomous in their responses, develop the sense of good and bad via social learning and intrinsic motivations and not because they are under permanent vigilance of a robot or any other device. As anecdotal evidence suggest, babies and toddlers are aware of interactive devices as cameras and react consequently after certain number of expositions.

Hence, we need to discuss and study potential dependency and behavioural addiction towards robots. Similarly, establish empathic limitations towards robots for robot natives is desirable. In other words, like cars and appliances, we do not want robot natives to attach too much to robots as their creators are not necessarily prioritising our well-being. Government and academics have a word on this. We need to question ourselves, what level of deceptiveness, autonomy and reactivity do we allow in the robots around our babies and toddlers?

5 Conclusions

This chapter does not aim to provide a technical contribution, nor does present results or a method. Neither we aim to provide an extensive literature review as cHRI has been already reviewed by other colleagues and bHRI studies are still very limited. We aim to influence the design process and policy making of future robot technologies and highlight the potential negative consequences of overuse, lack of regulation, and non-critical optimism over robots. Similarly, we aim to highlight the areas of opportunity on the unexplored field of bHRI. As previous similar papers as (Belpaeme et al., 2013) we aim spark a conversation about bHRI and trigger the interest of early career researchers on the topic. Summarising, this chapter aim to provide suitable responses for the research questions presented in the introduction.

5.1 Questions

1. **What will make robots suitable for baby use? What distinguish them from interactive toys and robot toys?** Similarly, we must distinguish service and social robots from interactive toys and robotic toys. We provide definitions of service and social robots in the section 2 of this chapter. Also the main differences can be explained in terms of their design purpose; a) Interactive toys and robot toys are designed specifically to interact with babies and toddlers in a safe, controlled way, b) Toys devices have a ludic and educative purpose as a main function. Service and Social robots could not have only that purpose.
2. **Are robot natives and us ready to use robots and take advantage of robot technology in domestic, educative, and work environments?** Certainly, we have not performed enough studies in bHRI as research community to provide an answer. We already have domestic robots interacting with babies and children in certain homes. However, it is unforeseen what are the possible long-term consequences of these interactions in combination with other interactive technologies. Hence, further studies, discussions, and legislative and political discussions are required. All of them framed in a multidisciplinary, human-centred approach.

3. **What are the potential positive and negative outcomes that can be produced by robots interacting with robot natives?** As other previous interactive technologies, overuse, lack of testing, unethical designs can create behavioural addictions, and even physical hurt to robot natives. We have discussed that in more detail in the section 2.4 of this chapter.
4. **What are the design guidelines required for a successful, engaging, and safe bHRI?** We provide some design suggestions and frameworks aiming for a virtuous design of service and social robots interacting with robot natives in the section 3 of this paper.
5. **What regulations should be put in place to ensure safe bHRI?** Similarly, we elaborated on the ethical considerations for the design and use of robots for babies and toddlers. We also encourage an initial discussion on regards of appropriated robot behaviours, deceptiveness, levels of anthropomorphism and persuasive skill on robots that do not influence negatively on robot natives in the section 4 of this chapter.

Interactive technologies have deep impact in the human experience as we can observe in the adoption of smartphones, computers, internet and less interactive or smart technologies as domestic devices. These technologies have changed the way we live for good and bad. In the case of robots interacting with babies and toddlers, we aim to add to the discussion about the regulations required for the robots used by children in next decades of the 21st century. We conclude that a qualitative and multidisciplinary approach should be used to guide the design and use of robots in the home and around families. We focus on service robots and social robots with the current limitations that they have but consider the possible enhancements that these robots will have soon. Certainly, there is the risk of being highly speculative in bHRI. Hence, we consider that in an initial stage observational, naturalistic, ethnographic studies would be useful to understand better the future challenges and perspectives in the area.

References

- Abe, K., Shiomi, M., Pei, Y., Zhang, T., Ikeda, N., & Nagai, T. (2018). ChiCaRo: Tele-presence robot for interacting with babies and toddlers. *Advanced Robotics*, 32(4), 176–190. <https://doi.org/10.1080/01691864.2018.1434014>
- Alves-Oliveira, P., Arriaga, P., Paiva, A., & Hoffman, G. (2017). YOLO, a Robot for Creativity: A Co-Design Study with Children. *Proceedings of the 2017 Conference on Interaction Design and Children*, 423–429. <https://doi.org/10.1145/3078072.3084304>
- B. Sandoval, E. (2014). *Human robot interaction and fiction: A contradiction*. ICSR 2014 Sydney. https://link-springer-com.wwwproxy1.library.unsw.edu.au/chapter/10.1007/978-3-319-11973-1_6
- Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 591–594. <https://doi.org/10.1109/ROMAN.2004.1374827>
- Belpaeme, T., Baxter, P., de Greeff, J., Kennedy, J., Read, R., Looije, R., Neerincx, M., Baroni, I., & Zelati, M. C. (2013). Child-Robot Interaction: Perspectives and Challenges. In G. Herrmann, M. J. Pearson, A. Lenz, P.

Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

- Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics* (pp. 452–459). Springer International Publishing. https://doi.org/10.1007/978-3-319-02675-6_45
- Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in child–robot interaction. *International Journal of Human-Computer Studies*, 69(7), 539–550. <https://doi.org/10.1016/j.ijhcs.2011.04.003>
- Betsch, M., Kalbhen, K., Michalik, R., Schenker, H., Gatz, M., Quack, V., Siebers, H., Wild, M., & Migliorini, F. (2021). The influence of smartphone use on spinal posture – A laboratory study. *Gait & Posture*, 85, 298–303. <https://doi.org/10.1016/j.gaitpost.2021.02.018>
- Cappuccio, M. L., Sandoval, E. B., Mubin, O., Obaid, M., & Velonaki, M. (2021). Can robots make us better humans? *International Journal of Social Robotics*, 13(1), 7–22.
- Chubb, J., Missaoui, S., Concannon, S., Maloney, L., & Walker, J. A. (2022). Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI. *International Journal of Child-Computer Interaction*, 32, 100403.
- Ciesielska, M., Boström, K. W., & Öhlander, M. (2018). Observation Methods. In M. Ciesielska & D. Jemielniak (Eds.), *Qualitative Methodologies in Organization Studies: Volume II: Methods and Possibilities* (pp. 33–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-65442-3_2
- Daily, S. B., James, M. T., Cherry, D., J. Porter, J., Darnell, S. S., Isaac, J., & Roy, T. (2017). Chapter 9 - Affective Computing: Historical Foundations, Current Applications, and Future Trends. In M. Jeon (Ed.), *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 213–231). Academic Press. <https://doi.org/10.1016/B978-0-12-801851-4.00009-4>
- Dautenhahn, K. (1998). The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied Artificial Intelligence*, 12(7–8), 573–617. <http://www.tandfonline.com/doi/abs/10.1080/088395198117550>
- Dautenhahn, K. (1999). Robots as social actors: Aurora and the case of autism. *Proc. CT99, The Third International Cognitive Technology Conference, August, San Francisco*, 359, 374.
- Desmurget, M. (2020). *La fábrica de cretinos digitales: Los peligros de las pantallas para nuestros hijos* (L. C. Fernández, Trans.). Ediciones Península.
- Duffy, B. R., Rooney, C., O'Hare, G. M., & O'Donoghue, R. (1999). What is a Social Robot? *10th Irish Conference on Artificial Intelligence & Cognitive Science, University College Cork, Ireland, 1-3 September, 1999*. <http://researchrepository.ucd.ie/handle/10197/4412>
- EU CORDIS Horizon 2020. (2018). *Child-Robot Communication and Collaboration: Edutainment, Behavioural Modelling and Cognitive Development in Typically Developing and Autistic Spectrum Children*. <https://cordis.europa.eu/project/id/687831>
- EU CORDIS, Horizon 2020. (2019). *DE-ENIGMA: Multi-Modal Human-Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children*. <https://cordis.europa.eu/project/id/688835>
- Fenech, S. (2021, April 12). Would you believe we spend 5.5 hours a day on average looking at our phones. *Tech Guide*. <https://www.techguide.com.au/amp/news/mobiles-news/would-you-believe-we-spend-5-5-hours-a-day-looking-at-our-smartphones/>
- Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Prezioso, M., Sciarretta, C., del Cimmuto, V., Lisciani, D., & Merla, A. (2021). Facilitating the Child–Robot Interaction by Endowing the Robot with the Capability of

Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

Understanding the Child Engagement: The Case of Mio Amico Robot. *International Journal of Social Robotics*, 13(4), 677–689. <https://doi.org/10.1007/s12369-020-00661-w>

Fitter, N. T., Funke, R., Pulido, J. C., Eisenman, L. E., Deng, W., Rosales, M. R., Bradley, N. S., Sargent, B., Smith, B. A., & Mataric, M. J. (2019). Socially Assistive Infant-Robot Interaction: Using Robots to Encourage Infant Leg-Motion Training. *IEEE Robotics & Automation Magazine*, 26(2), 12–23. <https://doi.org/10.1109/MRA.2019.2905644>

Gonzalez-Jimenez, S., Gallo, D., Sosa, R., Sandoval, E. B., Colombino, T., & Grasso, M. A. (2022). A Decision Support Design Framework for Selecting a Robotic Interface. *Proceedings of the 10th International Conference on Human-Agent Interaction*, 104–113. <https://doi.org/10.1145/3527188.3561913>

Haeghele, M. & International Federation of Robotics. Statistical Department. (2016). *World robotics service robots, 2016*. IFR Statistical Department, VDMA. http://canterbury.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2AwNtlz0EUrEywMU40NkkzNLV PSTFMTE5OAjZK0pBTLNAOLpERLA4sU1E2vDLBLO8E7Y9CtwH-uYK6tkbG5pYWZKTMDM7AyBh38GBEB2tRhAWx5gHZ5GkGO3YHzLeALDEFWIFUuboIMLKANB0IMTKI5IgzK4 NUtCkX5SfmgA5QViiF5GSJQrKMArEnNRBkU3VxDnD10EQbGQ4dh4qHOMhjYAF27FMIGBSSTCxS00A3fCeCj nQxtkxMMTdnSJRKTE1JMU20NDaTZJDCbY4UPklpBj6Qc0AYsvRChOE1DZicU2WR_QoASXB2Sw

Helsper, E. J., & Eynon, R. (2010). Digital natives: Where is the evidence? *British Educational Research Journal*, 36(3), 503–520. <https://doi.org/10.1080/01411920902989227>

Hennig, N. (2018). *Siri, Alexa, and Other Digital Assistants: The Librarian's Quick Guide*. ABC-CLIO.

Hoffman, G. (2019, May 1). *Anki, Jibo, And Kuri: What We Can Learn From Social Robots That Didn'T Make It* *IEEE Spectrum - IEEE Spectrum*. IEEE Spectrum: Technology, Engineering, and Science News. <https://spectrum.ieee.org/autoton/robotics/home-robots/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures>

Huang, C.-M., & Mutlu, B. (2012, March 5). *Robot Behavior Toolkit: Generating Effective Social Behaviors for Robots*. HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction. <https://doi.org/10.1145/2157689.2157694>

IEEE. (2022). *ROBOTS: Your Guide to the World of Robotics*. ROBOTS: Your Guide to the World of Robotics. <https://robots.ieee.org/>

IFR. (2021). *World Robotics 2021 – Service Robots report released*. IFR International Federation of Robotics. <https://ifr.org/ifr-press-releases/news/service-robots-hit-double-digit-growth-worldwide>

ISO. (2021). *ISO 8373:2021 Robotics Vocabulary* [Standards]. ISO. <https://www.iso.org/standard/75539.html>

Jeong, S., Santos, K. D., Graca, S., O'Connell, B., Anderson, L., Stenquist, N., Fitzpatrick, K., Goodenough, H., Logan, D., Weinstock, P., & Breazeal, C. (2015). Designing a socially assistive robot for pediatric care. *Proceedings of the 14th International Conference on Interaction Design and Children*, 387–390. <https://doi.org/10.1145/2771839.2771923>

Jo, S., Baek, I. C., Fava, M., Mischoulon, D., Hong, J. P., Kim, H., Park, M. J., Kim, E. J., & Jeon, H. J. (2021a). Association of smartphone overuse with depression, anxiety, and other addictive behaviors: A nationwide community sample of Korean adults. *Psychiatry Research*, 304, 114133. <https://doi.org/10.1016/j.psychres.2021.114133>

Jo, S., Baek, I. C., Fava, M., Mischoulon, D., Hong, J. P., Kim, H., Park, M. J., Kim, E. J., & Jeon, H. J. (2021b). Association of smartphone overuse with depression, anxiety, and other addictive behaviors: A nationwide

Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

- community sample of Korean adults. *Psychiatry Research*, 304, 114133. <https://doi.org/10.1016/j.psychres.2021.114133>
- Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction*, 19(1–2), 61–84. <https://doi.org/10.1080/07370024.2004.9667340>
- Karwowski, W., & Rahimi, M. (1992). *Human-Robot Interaction*. CRC Press.
- Koven, B. D. (2013). *The Well-Played Game: A Player's Philosophy* (Illustrated edition). The MIT Press.
- LeBlanc, A. G., Gunnell, K. E., Prince, S. A., Saunders, T. J., Barnes, J. D., & Chaput, J.-P. (2017). The Ubiquity of the Screen: An Overview of the Risks and Benefits of Screen Time in Our Modern World. *Translational Journal of the American College of Sports Medicine*, 2(17), 104–113. <https://doi.org/10.1249/TJX.0000000000000039>
- Lin, Y.-C., Liu, T.-C., Chang, M., & Yeh, S.-P. (2009). Exploring children's perceptions of the robots. *International Conference on Technologies for E-Learning and Digital Entertainment*, 512–517.
- Liu, J., Riesch, S., Tien, J., Lipman, T., Pinto-Martin, J., & O'Sullivan, A. (2021). Screen Media Overuse and Associated Physical, Cognitive, and Emotional/Behavioral Outcomes in Children and Adolescents: An Integrative Review. *Journal of Pediatric Health Care*. <https://doi.org/10.1016/j.pedhc.2021.06.003>
- Mineshita, Y., Kim, H.-K., Chijiki, H., Nanba, T., Shinto, T., Furuhashi, S., Oneda, S., Kuwahara, M., Suwama, A., & Shibata, S. (2021). Screen time duration and timing: Effects on obesity, physical activity, dry eyes, and learning ability in elementary school children. *BMC Public Health*, 21(1), 422. <https://doi.org/10.1186/s12889-021-10484-7>
- Mohan, R. E., Rojas, N., & Chua, A. (2014). Design principles for robot inclusive spaces: A case study with Roomba. *2014 IEEE International Conference on Robotics and Automation (ICRA)*. <https://doi.org/10.1109/ICRA.2014.6907681>
- Mubin, O., Obaid, M., Sandoval, E., & Fjeld, M. (2015). Using video preferences to understand the human perception of real and fictional robots. *Proceedings of the 3rd International Conference on Human-Agent Interaction*, 33–39.
- Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, 17(3), 347–369. <https://doi.org/10.1075/is.17.3.02nom>
- Phillips, E., Zhao, X., Ullman, D., & Malle, B. (2018). *The ABOT Database*. <https://www.abotdatabase.info/>
- Prensky, M. (2001). Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently? *On the Horizon*, 9(6), 1–6. <https://doi.org/10.1108/10748120110424843>
- Rasouli, S., Gupta, G., Ghafurian, M., & Dautenhahn, K. (2022). Proposed Applications of Social Robots in Interventions for Children and Adolescents with Social Anxiety. *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 1–7.
- Ratan, Z. A., Zaman, S. B., Islam, S. M. S., & Hosseinzadeh, H. (2021). Smartphone overuse: A hidden crisis in COVID-19. *Health Policy and Technology*, 10(1), 21–22. <https://doi.org/10.1016/j.hlpt.2021.01.002>
- Sandoval, E. B., Brown, S., & Velonaki, M. (2018). How the Inclusion of Design Principles Contribute to the Development of Social Robots. *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, 535–538. <https://doi.org/10.1145/3292147.3292239>

Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction

- Sandoval, E. B., & Mubin, O. (2015, December 4). The technology in science fiction is not always what we want in the real world. *The Conversation*. <http://theconversation.com/the-technology-in-science-fiction-is-not-always-what-we-want-in-the-real-world-51196>
- Sandoval, E. B., & Penalzoza, C. I. (2012). Children's knowledge and expectations about robots: A survey for future user-centered design of social robots. *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference On*, 107–108.
- Schell, J. (2014). *The Art of Game Design: A Book of Lenses, Second Edition* (2 edition). A K Peters/CRC Press.
- Sekiguchi, T., Hagiwara, Y., Yabe, Y., Tsuchiya, M., Itaya, N., Yoshida, S., Yano, T., Sogi, Y., Suzuki, K., & Itoi, E. (2018). Playing video games for more than 3 hours a day is associated with shoulder and elbow pain in elite young male baseball players. *Journal of Shoulder and Elbow Surgery*, 27(9), 1629–1635. <https://doi.org/10.1016/j.jse.2018.06.005>
- Sharkey, A., & Sharkey, N. (2011). Children, the elderly, and interactive robots. *IEEE Robotics & Automation Magazine*, 18(1), 32–38. <http://ieeexplore.ieee.org/abstract/document/5751987/>
- Shiomi, M., Abe, K., Pei, Y., Zhang, T., Ikeda, N., & Nagai, T. (2016). ChiCaRo: Tele-presence Robot for Interacting with Babies and Toddlers. *Proceedings of the Fourth International Conference on Human Agent Interaction*, 349–351. <https://doi.org/10.1145/2974804.2980496>
- Silvera-Tawil, D., & Brown, S. A. (2019). Cross-Collaborative Approach to Socially-Assistive Robotics: A Case Study of Humanoid Robots in a Therapeutic Intervention for Autistic Children. In O. Korn (Ed.), *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction* (pp. 165–186). Springer International Publishing. https://doi.org/10.1007/978-3-030-17107-0_9
- Sosa, R., Montiel, M., Sandoval, E. B., & Mohan, R. E. (2018). ROBOT ERGONOMICS: TOWARDS HUMAN-CENTRED AND ROBOT-INCLUSIVE DESIGN. *DS92: Proceedings of the DESIGN 2018 15th International Design Conference*, 2323–2334.
- Sung, J., Grinter, R. E., & Christensen, H. I. (2010). Domestic Robot Ecology. *International Journal of Social Robotics*, 2(4), 417–429. <https://doi.org/10.1007/s12369-010-0065-8>
- ThinkTV. (2021, June 1). *The average Aussie home has 6.6 devices for consuming content*. ThinkTV. <https://thinktv.com.au/facts-and-stats/australian-homes-are-experiencing-a-screen-explosion/>
- Zehfroosh, A., & Tanner, H. G. (2022). A Hybrid PAC Reinforcement Learning Algorithm for Human-Robot Interaction. *Frontiers in Robotics and AI*, 9, 797213.
- Zlotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics*, 7(3), 347–360. <https://doi.org/10.1007/s12369-014-0267-6>

Biographies

Dr. Eduardo B. Sandoval is a social robotics researcher, his work spans different aspects of social robotics, such as Reciprocity in Human-Robot Interaction (HRI), robots and games, robots and education, and addiction to robots developing over 40 peer reviewed publications and other research outcomes.. Beyond traditional robotics, there is a growing interest in the idea of designing

machines that are capable of meaningful social interactions with humans. His work incorporates insights from behavioural economics and social psychology to explore different approaches in social robotics. In 2014 he obtained the “Alfonso Caso” Medal as one of the best postgraduate students at the National Autonomous University of Mexico. In 2018 he joined to the University of New South Wales (UNSW) under the prestigious Scientia program as Associate Lecturer. He joined to the National Research System of Mexico in 2019. In 2022, he was awarded with the first medal “Wilfrido Massieu” at the National Polytechnic Institute of Mexico. In June of 2022, he was awarded with a second Scientia Lecturer appointment at UNSW until 2026.

Dr Massimiliano (Max) Cappuccio is Deputy director of the Values in Defense and Security Technology group. His research in human performance and human-machine interaction attempts to combine different embodied approaches to cognition. As a cognitive philosopher and a technology ethicist, his research on intelligent systems is interdisciplinary and aims to integrate phenomenological analyses, empirical experimentation, and synthetic modelling. His current research focuses on skill acquisition and disruption, social-robotics, theory and ethics of artificial intelligence and autonomous systems. His early work has focused on mirror neurons theory, joint attention and deictic pointing, the frame problem of artificial intelligence, and the philosophical foundations of computationalism. During his career, he has produced more than a hundred original scientific contributions, including papers in journals, book chapters, collective volumes, special issues of journals, and papers presented at international conferences. He conducts an intense activity as an organizer of academic events, including interdisciplinary workshops, research seminar series, and international conferences (like the TEMPER workshop on Training, Enhancement, and Military Performance and the annual Joint UAE Symposium on Social Robotics).

Prof Mark Billinghurst is Director of the Empathic Computing Laboratory, and Professor at the University of South Australia in Adelaide, Australia, and also at the University of Auckland in Auckland, New Zealand. He earned a PhD in 2002 from the University of Washington and conducts research on how virtual and real worlds can be merged, publishing over 650 papers on Augmented Reality, Virtual Reality, remote collaboration, Empathic Computing, and related topics. In 2013 he was elected as a Fellow of the Royal Society of New Zealand, and in 2019 was given the ISMAR Career Impact Award in recognition for lifetime contribution to AR research and commercialization. In 2022 he was selected for the ACM SIGCHI Academy, for leading human-computer interaction researchers, and also selected for the IEEE VGTC VR Academy for leading VR researchers.

Reference list

- Abbass HA, Scholz J, Reid DJ (2018) Foundations of Trusted Autonomy: An Introduction. In *Foundations of Trusted Autonomy*. Springer.
- Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), in *Robot ethics: The ethical and social implications of robotics* (pp. 35–52). Cambridge, MA: MIT Press.
- Adams, R. A., Shipp, S. & Friston, K. J. (2013) Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218 (3), 611–643.
- Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*.
- Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008) Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, 11 (9), 1109–1116.
- Airenti, G. (2015). The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy. *International Journal of Social Robotics*, 7(1), 117–127.
- Airenti, G. (2018). The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind. *Frontiers in Psychology*, 9, 2136.
- Alac, M., Movellan, J., & Tanaka, F. (2011). When a robot is social: spatial arrangements and multimodal semiotic engagement in the practice of social robotics. *Social Studies of Science*, 41(6), 893–926.
- Alahbabi M., Almazroei F., Almarzoqi M., Almeheri A., Alkabi M., Nuaimi A.A., Cappuccio M.L., Alnajjar F. (2017), “Avatar Based Interaction Therapy: A Potential Therapeutic Approach for Children” In X IEEE International Conference on Mechatronics and Automation (ICMA), 480–496 [SP13].
- Allen, C., Wallach, W., & Smit, I. (2006). Why Machine Ethics? *IEEE Intelligent Systems*, 21(4), 12–17.

Alnajjar F., Bartneck C., Baxter P., Belpaeme T., Cappuccio M., Di Dio C., Eyszel F., Handke J., Mubin O., Obaid M. (2021a), *Robots in Education: An Introduction to High-Tech Social Agents, Intelligent Tutors, and Curricular Tools*, Routledge **[CP10]**.

Alnajjar F.S., Cappuccio M.L., Renawi A., Mubin O., Kiong Loo C. (2020), "Personalized Robot Interventions for Autistic Children: an Automated Methodology for Attention Assessment", *International Journal of Social Robotics*, 2020 **[CP4]**.

Alnajjar F.S., Cappuccio M.L., Mubin O., Arshad R., Shahid S. (2021b), Humanoid robots and autistic children: a review on technological tools to assess social attention and engagement. *International Journal of Humanoid Robotics*, World Scientific Publishing Co. **[CP3]**.

Alnajjar F., Renawi A.M., Cappuccio M.L., & Mubin O. (2019), A low-cost autonomous attention assessment system for robot intervention with autistic children. *Proceedings Of The Tenth IEEE Global Engineering Education Conference (Educon 2019)*, 9-11 April 2019, American University in Dubai, Dubai, UAE, 787-792 **[SP12]**.

Alotaibi M., Alnajjar F., Cappuccio M.L., S. Khalid (2023), Efficacy of Emerging Technologies to Manage Childhood Obesity: A Systematic Review. *Diabetes Metab Syndr Obes.* 2022; 15: 1227-1244 **[SP15]**.

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, 22(4), 477–493.

Andler, D. (2006). Phenomenology in Artificial Intelligence and Cognitive Science. In *A Companion to Phenomenology and Existentialism* (pp. 377–393). Blackwell Publishing Ltd.

Asada, M. (2013). Cognitive developmental robotics: From physical interaction to social one. In *Intelligent Systems, Control and Automation: Science and Engineering* (pp. 37–51). Springer Japan.

Asada, M. (2014). Towards artificial empathy based on affective developmental robotics. *Journal of the Robotics Society of Japan*, 32(8), 666–677.

Asada, M. (2015a). Development of artificial empathy. *Neuroscience Research*, 90, 41–50.

Asada, M. (2015b) Towards artificial empathy. *International Journal of Social Robotics*, vol.7, pp.19–33, 2015.

Asada, M. (2016). Mirror neuron system and social cognitive development. In *Cognitive Neuroscience Robotics A* (pp. 73–93). Springer Japan.

Asada, M. (2020). Rethinking autonomy of humans and robots. *Journal of Artificial Intelligence and Consciousness*, 07(02), 141–153.

Asada, M. (2021a). Physical Embodiment. In *Encyclopedia of Robotics* (pp. 1–6). Springer Berlin Heidelberg.

Asada A. (2021b). Robot Pain and Empathy. In *Encyclopedia of Robotics* (pp. 1–9). Springer Berlin Heidelberg.

Asada, M. (2021c). Robotics and reinforcement learning. *Journal of the Robotics Society of Japan*, 39(7), 575–580.

Bae, I., & Han, J. (2017). Does height affect the strictness of robot assisted teacher? *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17: ACM/IEEE International Conference on Human-Robot Interaction, Vienna Austria.

Ball, A., Silvera-Tawil, D., Rye, D., & Velonaki, M. (2014). Group Comfortability When a Robot Approaches. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social Robotics* (Vol. 8755, pp. 44–53). Springer International Publishing.

Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics*, 185(4), 725–740.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?

Bartneck, C. and Hu, J. (2008) Exploring the abuse of robots, *Interaction studies*. John Benjamins, 9(3), pp. 415–433.

Bartneck, C. *et al.* (2018) Robots and racism, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 196–204.

Bartneck, C., Kanda, T., Mubin, O., & Al Mahmud, A. (2009a). Does the Design of a Robot Influence Its Animacy and Perceived Intelligence? In *International Journal of Social Robotics* (Vol. 1, Issue 2, pp. 195–204). <https://doi.org/10.1007/s12369-009-0013-7>

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009b). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1), 71–81.

Bartneck, C., Reichenbach, J., & Carpenter, J. (2006). Use of Praise and Punishment in Human-Robot Collaborative Teams. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*.

Bartneck, C., Reichenbach, J. and Carpenter, J. (2008) 'The carrot and the stick: The role of praise and punishment in human–robot interaction', *Interaction studies*. John Benjamins, 9(2), pp. 179–203.

Baumeister, R. F. (1984) Choking under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skilful Performance. *Journal of Personality and Social Psychology*. 46 (3), 610–620.

Baumeister, R. F. & Heatherton, T. F. (1996) Self-Regulation Failure: An Overview. *Psychological Inquiry*. 7 (1). Routledge, 1–15.

Beilock, S. L. (2011) *Choke*. Boston: Little, Brown Book Group.

Beilock, S. L., & Gray, R. (2007) Why do athletes choke under pressure? In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (425–444). John Wiley & Sons Inc.

Beilock, S. L., & Gray, R. (2012) From Attentional Control to Attentional Spillover: A Skill-Level Investigation of Attention, Movement, and Performance Outcomes. *Human Movement Science*, 31 (6), 1473–1499.

Beilock, S. L., Bertenthal, I. B., Hoerger, M. & Carr, T. H. (2008) When Does Haste Make Waste? Speed-Accuracy Tradeoff, Skill Level, and the Tools of the Trade. *Journal of Experimental Psychology: Applied*. 14 (4), 340–352.

Beilock, S. L., Carr, T. H., MacMahon, C. & Starkes, J. L. (2002) When Paying Attention Becomes Counterproductive: Impact of Divided versus Skill-Focused Attention on Novice and Experienced Performance of Sensorimotor Skills. *Journal of Experimental Psychology: Applied*. 8 (1), 6–16.

Beilock, S.L. & Carr, T.H. (2001) On the fragility of skilled performance: what governs choking under pressure? *Journal of experimental psychology. General*. 130(4), pp.701–725.

Bendel, O. (2017). Sex Robots from the Perspective of Machine Ethics. In A. D. Cheok, K. Devlin, & D. Levy (Eds.), *Love and Sex with Robots* (Vol. 10237, pp. 17–26). Springer International Publishing.

Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. *Communications of the ACM*, 42(5), 51–58.
<https://doi.org/10.1145/301353.301410>

Bishop, L., van Maris, A., Dogramadzi, S., & Zook, N. (2019). Social robots: The influence of human and robot characteristics on acceptance. *Paladyn, Journal of Behavioural Robotics*, 10(1), 346–358.

Bongard, J. C., & Pfeifer, R. (2003). Evolving Complete Agents using Artificial Ontogeny. In *Morpho-functional Machines: The New Species* (pp. 237–258). Springer Japan.

Boy, G. A. (Ed.). (2017). *The handbook of human-machine interaction*. CRC Press.

Breazeal C.A. (2002) Designing sociable robots. MIT Press, Cambridge, MA.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1–2), 119–155. [https://doi.org/10.1016/S1071-5819\(03\)00018-1](https://doi.org/10.1016/S1071-5819(03)00018-1)

Breazeal, C.A., Dautenhahn, K., & Kanda, T. (2016). Social Robotics. In *Springer Handbook of Robotics* (pp. 1935–1972).

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, *RA-2*(1), 14–23. <https://doi.org/10.1109/JRA.1986.1087032>

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159.

Brooks R.A. (1999) *Cambrian intelligence: the early history of the new A.I.* MIT Press; Cambridge, MA.

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. (2013) Active inference, sensory attenuation and illusions. *Cognitive Processing*. *14*(4), 411–427.

Brown, H., Friston, K. & Bestmann, S. (2011) Active Inference, Attention, and Motor Preparation. *Frontiers in Psychology* *2*(218).

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 599.

Bruineberg, J., & Stone, O. (2024). Structuring embodied minds: attention and perceptual agency. *Philosophical Studies*, *181*(2–3), 461–484.

Bruni, D., Perconti, P., & Plebe, A. (2018). Anti-anthropomorphism and Its Limits. *Frontiers in Psychology*, *9*, 2205.

Bryson, J. J. (2010). Why robot nannies probably won't do much psychological damage. *Interaction Studies*, *11*(2), 196–200.

Bryson, J. J. (2018). Patience is not a virtue: AI and the design of ethical systems. *AI & Society*, *33*(4), 561–571. <https://doi.org/10.1007/s00146-017-0771-8>

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI Magazine*, *38*(2), 22–34.

- Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006) Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology: CB*, 16(19), 1905–1910.
- Cangelosi, A. (2001). *Simulating the evolution of language* (A. Cangelosi & D. Parisi, Eds.). Springer.
- Cangelosi, A. (2011). *Robotics and embodied agent modelling of the evolution of language*. Oxford University Press.
- Cangelosi, A., & Asada, M. (2022). *Cognitive robotics*. MIT Press.
- Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Nori, F., Fadiga, L., Wrede, B., Rohlfing, K., Tuci, E., Dautenhahn, K., Saunders, J., & Zeschel, A. (2010). Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3), 167–195.
- Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics*. MIT Press.
- Cappuccio M.L. (2009), Constructing the Space of Action. From Bio-Robotics to Mirror Neurons. *World Futures. The Journal of General Evolution*, 65/2 (February 2009), Taylor & Francis, London, pp. 126-132 **[SP1]**.
- Cappuccio M.L. (2015a), Inference or Familiarity? The Embodied Roots of Social Cognition. *Synthesis Philosophica*, 29/2, (April 2015), Croatian Philosophical Society, Zagreb, pp. 253-272. **[SP14]**.
- Cappuccio, M. L. (2015b). Introduction: when embodied cognition and sport psychology team-up. *Phenomenology and the Cognitive Sciences*, 14(2), 213–225.
- Cappuccio M.L. (2015c), The Seminal Speculation of a Precursor: Elements of Embodied Cognition and Situated AI in Alan Turing, in V. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Springer - Synthese Library, 2015 **[CP1]**.
- Cappuccio M.L. (2016), Autonomous artificial agents, ethical decision, and the frame problem, *AISB Annual Convention 2016, AISB 2016*.

Cappuccio M.L. (2017), Flow, Choke, Skill: the role of the non-conscious in sport performance in Z. Radman (ed.), *Before Consciousness. In Search of the Fundamentals of Mind*, Imprint [SP5].

Cappuccio M.L. (2023), Dreyfus is right: knowledge-that limits your skill, *Synthese* 202, 85 [SP7].

Cappuccio M.L., Devitt K.S., Wrigley C. (2020), *Attitudes and Beliefs toward the Introduction of a Logistic Robot in a RAAF Base Warehouse*, Unpublished report - January-March 2020, 18 pages, 7547 words, Commissioned by Royal Australian Air Force [SP17].

Cappuccio M.L., Froese T. (eds.) (2014), *Enactive Cognition at the Edge of Sense-Making. Making Sense of Non-Sense*, Palgrave Macmillan, November 2014 [SP4].

Cappuccio M.L., Galliot J.C., Lanteri A., Eyssel F. (2024), "Autonomous Systems and Technology Resistance: New Tools for Monitoring of Acceptance, Trust, and Tolerance", *International Journal of Social Robotics*, Springer [CP5].

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021a), "Mapping Meaning and Purpose in Human-Robot Teams: Anthropomorphic Agents in Military Operations", *Philosophical Journal of Conflict and Violence*, Vol. V (1), Trivent Publishing [SP16].

Cappuccio M.L., Galliot J.C., Sandoval E.B. (2021b), "Saving Private Robot: Risks and Advantages of Anthropomorphism in Soldier-Agent Teams", *International Journal of Social Robotics*, Springer [CP6].

Cappuccio M.L., Gray R., Hill D.M., Mesagno C., Carr T.H. (2019a), "The Many Threats of Self-Consciousness: Embodied Approaches to Choking Under Pressure in Sensorimotor Skills", in Massimiliano L. Cappuccio (ed.), *Handbook of Embodied Cognition and Sport Psychology*, MIT Press, 2019 [SP6].

Cappuccio M.L., Ilundain-Agurruza J. (2019), "Embodied Cognition", in Dieter Hackfort and Robert Schinke (eds.), *The Routledge International Encyclopedia of Sport and Exercise Psychology*, Routledge, 2019 [SP2].

Cappuccio M.L., Kirchhoff M.D., Alnajjar F.S., Tani J. (2019b), “Unfulfilled Prophecies in Sport Performance. Active Inference and the Choking Effect”, *Journal of Consciousness Studies*, Volume 27, Numbers 3-4, 2019, pp. 152-184 [CP2].

Cappuccio M.L., Peeters A., McDonald W (2019), “Sympathy for Dolores. Moral consideration for robots based on virtue and recognition”, *Philosophy & Technology*, 1-23 (2019), Springer [SP18].

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2021a), “Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents”, *International Journal of Social Robotics* [CP7].

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (2021b), “Robotics Aids for Character Training: More Than Just Another Enabling Condition”, *International Journal of Social Robotics*, Springer [CP8].

Cappuccio M.L., Sandoval E.B., Mubin O., Obaid M., Velonaki M. (eds.) (2021c), “Virtuous Robotics: Artificial Agents and the Good Life”, special issue of *International Journal of Social Robotics*, Springer Verlag [CP9].

Cappuccio M.L., Shepherd, S.V. (2013), “Pointing Hand: Joint Attention and Embodied Symbols”, *The Hand, an Organ of the Mind: What the Manual Tells the Mental*, Zdravko Radman [SP10].

Cappuccio M.L., Wheeler M. (2012), “Ground-level Intelligence: Action-Oriented Representation and Dynamics of the Background”, in Zdravko Radman (ed.), *Knowing without Thinking. Mind, Action, Cognition and the Phenomenon of the Background*. Palgrave Macmillan, London, February 2012, pp. 13-36 [SP3].

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45), 17825–17830.

Cauchard JR, Zhai KY, Spadafora M, Landay JA (2016) Emotion encoding in Human-Drone Interaction. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, 2016, pp. 263-270.

- Chen, A.G., Benrimoh, D., Parr, T. & Friston, K. (2019) A Bayesian account of generalist and specialist formation under the active inference framework. *BioRxiv*. doi: 10.1101/644807.
- Chesher, C., & Andreallo, F. (2022). Eye machines: Robot eye, vision and gaze. *International Journal of Social Robotics*, 14(10), 2071–2081.
- Christensen, W., Sutton J. & McIlwain, D. (2015) Putting Pressure on Theories of Choking: Towards an Expanded Perspective on Breakdown in Skilled Performance. *Phenomenology and the Cognitive Sciences*. 14 (2). Springer Netherlands, 253–293.
- Chen, Y., Murata, S., Arie, H., Ogata, T., Tani, J., & Sugano, S. (2016). Emergence of interactive behaviours between two robots by prediction error minimization mechanism. *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Cergy-Pontoise, France. kn
- Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence*, 149(1), 131–150.
- Chrisley, R. (2009). Synthetic phenomenology. *International Journal of Machine Consciousness*, 01(01), 53–70.
- Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology. General*, 140(1), 102–116.
- Cialdini, R. B. (2007). *Influence: The psychology of persuasion* (Rev. ed.). Harper Business.
- Ciardo, F., & Wykowska, A. (2022). Robot's social gaze affects conflict resolution but not conflict adaptations. *Journal of Cognition*, 5(1), 2.
- Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Bradford Books.
- Clark, A. (2001). Reasons, Robots and the Extended Mind. *Mind & Language*, 16(2), 121–145.

Clark, A. (2001b). Visual Experience and Motor Action: Are the Bonds Too Tight? *The Philosophical Review*, 110(4), 495–519.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA.

Clark, A. (2013). The many faces of precision (Replies to commentaries on “Whatever next? Neural prediction, situated agents, and the future of cognitive science”). *Frontiers in Psychology*, 4, 270.

Clark, A. (2016) *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Clark, A. & Grush, R. (1999). Towards a Cognitive Robotics. *Adaptive Behaviour*, 7(1), 5–16.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3), 235–241.

Coeckelbergh, M. (2011) Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Int J of Soc Robotics* 3, 197–204.

Coeckelbergh, M. (2018) ‘Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering’, *Kairos. Journal of Philosophy & Science*. Sciendo, 20(1), pp. 141–158.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Consolvo, S., Klasnja, P., McDonald, D. W., & Landay, J. A. (2009). Goal-setting considerations for persuasive technologies that encourage physical activity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 263–272). ACM. <https://doi.org/10.1145/1518701.1518741>

- Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), 299–309. <https://doi.org/10.1007/s10676-016-9403-3>
- D'Angelo, D. (2020). The phenomenology of embodied attention. *Phenomenology and the Cognitive Sciences*, 19(5), 961–978.
- Dacey M (2017) Anthropomorphism as cognitive bias, *Philosophy of Science* 84 (5):1152-1164.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in Human-Robot Co-evolution. *Frontiers in Psychology*, 9, 468.
- Damiano, L., & Stano, P. (2018). Synthetic biology and artificial intelligence: Grounding a cross-disciplinary approach to the synthetic exploration of (embodied) cognition. *Complex Systems*, 27, 199–228.
- Damiano, Luisa, & Cañamero, L. (2010). *Constructing Emotions. Epistemological groundings and applications in robotics for a synthetic approach to emotions*. 20–28.
- Damiano, L., Dumouchel, P., & Lehmann, H. (2015a). Artificial empathy: An interdisciplinary investigation. *International Journal of Social Robotics*, 7(1), 3–5.
- Damiano, L., Dumouchel, P., & Lehmann, H. (2015b). Towards human–robot affective co-evolution overcoming oppositions in constructing emotions and empathy. *International Journal of Social Robotics*, 7(1), 7–18.
- Darling, K. (2016). Extending legal protection to social robots. In R. Calo, A. M. Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Köse-Bagci, H., Mirza, N. A., & Blow, M. (2009). KASPAR—A minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3–4), 369–397. <https://doi.org/10.1080/11762320903123567>
- Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1480), 679–704.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Köse-Bagci, H., Mirza, N. A., & Blow, M. (2009). KASPAR—a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3–4), 369–397.
<https://doi.org/10.1080/11762320903123567>

Dawson, G., et al. (2012). Early behavioural intervention is associated with normalized brain activity in young children with autism.

Daza, M., Barrios-Aranibar, D., Diaz-Amado, J., Cardinale, Y., & Vilasboas, J. (2021). An approach of social navigation based on proxemics for crowded environments of humans and robots. *Micromachines*, 12(2), 193.

De Graaf, M. M. A., & Ben Allouch, S. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems*, 61(12), 1476–1486.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, 6(4), 485–507.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10), 441–447.

Desideri, L., Bonifacci, P., Croati, G., Dalena, A., Gesualdo, M., Molinario, G., Gherardini, A., Cesario, L., & Ottaviani, C. (2021). The mind in the machine: Mind perception modulates gaze aversion during child–robot interaction. *International Journal of Social Robotics*, 13(4), 599–614.

Di Paolo, E. A. (2003). Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop. *Dynamical Systems Approach to Embodiment and Sociality*, 19–42.

Di Paolo, E. A., Noble, J., & Bullock, S. (2000). *Simulation models as opaque thought experiments*.

DiSalvo, C., Gemperle, F. (2003). From Seduction to Fulfillment: The Use of Anthropomorphic Form in Design. In: Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces, 67–72. DPPI '03. New York, NY, USA: ACM.

Dreyfus, H. L. (2005) Overcoming the myth of the mental: how philosophers can profit from the phenomenology of everyday expertise. *Proceedings and Addresses of the American Philosophical Association*, 79(2), 47–65.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171(18), 1137–1160.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3), 177–190.

Dumouchel, P., & Damiano, L. (2011). Artificial Empathy, Imitation and Mimesis. *Ars Vivendi Journal*, 1, 18–31.

Dumouchel, P., & Damiano, L. (2017). *Living with Robots*. Harvard University Press.

Eckhardt, D., et al. (2005). *Rubi: A parallel robotics start-up*. In *Proceedings of the 5th International Conference on Development and Learning (ICDL)*. (Note: primary authors include Movellan et al.; pilot data presented in HRI 2009; exact author list variable across sources.)

Epley, N., Waytz, A., & Cacioppo, J.T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.

Epley N, Waytz A, Akalis S, Cacioppo JT (2008) When We Need A Human: Motivational Determinants of Anthropomorphism. *Social Cognition* 26 (2): 143–55.

Erlhagen, W., Mukovskiy, A., Bicho, E., Panin, G., Kiss, C., Knoll, A., van Schie, H., & Bekkering, H. (2006). Goal-directed imitation for robots: A bio-inspired approach to action understanding and skill learning. *Robotics and Autonomous Systems*, 54(5), 353–360.

Esteban, P. G., Baxter, P., Belpaeme, T., Billing, E., Cai, H., Cao, H.-L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., Fang, Y., Ju, Z., Kennedy, J., Liu, H., Mazel, A., Pandey, A., Richardson, K., Senft, E., Thill, S., Ziemke, T. (2017). How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn: Journal of Behavioural Robotics*, 8(1), 18–38.

- Eyssel, F. and Hegel, F. (2012) '(S)he's Got the Look: Gender Stereotyping of Robots¹', *Journal of applied social psychology*. Blackwell Publishing Inc, 42(9), pp. 2213–2230.
- Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4), 724–731.
- Ferreira, J. F., & Dias, J. (2014). Attentional Mechanisms for Socially Interactive Robots. A Survey. *IEEE Transactions on Autonomous Mental Development*, 6(2), 110–125.
- Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3–4), 189–208.
[https://doi.org/10.1016/0004-3702\(71\)90010-5](https://doi.org/10.1016/0004-3702(71)90010-5)
- Fitzpatrick, P. M., Metta, G., Natale, L., Rao, S., & Sandini, G. (2001). *Learning about objects through action - initial steps towards artificial cognition*. OSD or Non-Service DoD Agency.
- Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann.
- Formosa, P. (2021). Robot autonomy vs. Human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines*, 31(4), 595–616.
- Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & Society*, 36(3), 839–851.
- Fortunati, L., & Edwards, A. (Eds.). (2024). *The De Gruyter handbook of robots in society and culture*. De Gruyter.
- Franco, F. (2013). Embodied Attention in Infant Pointing. In *Agency and Joint Attention* (pp. 152–164). Oxford University Press.
- Friston, K. (2010) The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*. 11 (2), 127–138.
- Friston, K. (2011) What is optimal about motor control? *Neuron* 72 (3), 488–498.

Friston, K. (2013) Life as we know it. *Journal of the Royal Society, Interface*, 10, 20130475.

Friston, K. (2014) Active inference and agency. *Cognitive Neuroscience*. 5(2), 119–121.

Friston, K. (2018) Computational mechanisms of curiosity and goal-directed exploration. *BioRxiv*. Doi: 10.1101/411272.

Gallagher, S. (2007). Social cognition and social robots. *Pragmatics and Cognition* 15 (3):435-453.

Gallagher, S. (2011). Interpretations of embodied cognition. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 59–71). Imprint Academic.

Gallagher S (2008) - Understanding Others: Embodied Social Cognition. In: Handbook of Cognitive Science, edited by Paco Calvo and Antoni Gomila, 437–52. San Diego: Elsevier.

Gallagher, S. (2013). You and I, robot. *AI and Society* 28 (4):455-460.

Gallese, V. (2007). The shared manifold hypothesis: Embodied simulation and its role in empathy and social cognition. *Empathy in Mental Illness*. Cambridge University Press.

Gallese, V., & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, 15(11), 512–519.

Galliot, J. (2018). The soldier's tolerance for autonomous systems. *Paladyn, Journal of Behavioural Robotics*, 9(1), 124–136.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behaviour*, 61, 633–655.

Georgeon, O. L., Wolf, C., & Gay, S. (2013). An Enactive approach to autonomous agent and robot learning. *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. 2013 IEEE International Conference on

Development and Learning and Epigenetic Robotics (ICDL), Osaka, Japan.

Gerdes, A. (2016). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers and Society*, 45(3), 274–279.

Ghiglino, D., Floris, F., De Tommaso, D., Kompatsiari, K., Chevalier, P., Priolo, T., & Wykowska, A. (2023). Artificial scaffolding: Augmenting social cognition by means of robot technology. *Autism Research: Official Journal of the International Society for Autism Research*, 16(5), 997–1008.

Gips, J. (2011): Towards the Ethical Robot. In: Machine Ethics. ed. Anderson, M., Anderson, S.L.. Cambridge: Cambridge University Press. pp.244-253.

Gray, R. & Allsop, J. (2013) Interactions between Performance Pressure, Performance Streaks, and Attentional Focus. *Journal of Sport & Exercise Psychology* 35 (4), 368–386.

Gray, R. & Cañal-Bruland, R. (2015) Attentional Focus, Perceived Target Size, and Movement Kinematics under Performance Pressure. *Psychonomic Bulletin & Review*. 22 (6), 1692–1700.

Han, J., & Conti, D. (2020). The Use of UTAUT and Post Acceptance Models to Investigate the Attitude towards a Telepresence Robot in an Educational Setting. *Robotics*, 9(2), 34.

Harvey, I., Di Paolo, E., Wood, R., Quinn, M., & Tuci, E. (2005). *Evolutionary robotics: A new scientific tool for studying cognition*. 11(1–2), 79–98.

Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1), 123–136. <https://doi.org/10.1075/is.13.3.02sim>

Heerink, M., Krose, B., Evers, V., & Wielinga, B. (2009). Measuring acceptance of an assistive social robot: a suggested toolkit. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 528–533.

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008). Understanding social robots: A user study on anthropomorphism. *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 574–579.

- He, K., Chan, W. P., Cosgun, A., Joy, A., & Croft, E. A. (2024). Robot gaze during autonomous navigation and its effect on social presence. *International Journal of Social Robotics*, 16(5), 879–897.
- Hohwy, J. (2013) *The Predictive Mind*. OUP Oxford.
- Hohwy, J. (2016) *The self-evidencing brain*. *Noûs*, 50(2), 259–285.
- Hoffman, G. (2012). Embodied cognition for autonomous interactive robots. *Topics in Cognitive Science*, 4(4), 759–772.
- Hoffman, G., & Breazeal, C. (2004). Collaboration in Human-Robot Teams. In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics.
- Hoffmann, M., Marques, H., Arieta, A., Sumioka, H., Lungarella, M., & Pfeifer, R. (2010). *Body in robotics: a review*. 2(4), 304–324.
- Hutto, D. D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press.
- Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews. Neuroscience*, 7(12), 942–951.
- Idei, H., Murata, S., Chen, Y., Yamashita, Y., Tani, J., & Ogata, T. (2018). A Neurorobotics Simulation of Autistic Behaviour Induced by Unusual Sensory Precision. *Computational Psychiatry (Cambridge, Mass.)*, 2, 164–182.
- Jackson, R. C., & Beilock, S. L. (2008) Performance Pressure and Paralysis by Analysis: Research and Implications. Farrow, D., Baker, J. & MacMahon, C. (eds.) in *Developing Sports Expertise: Researchers and Coaches Put Theory into Practice*, 104–118. London: Routledge.
- Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of The*. <https://www.pnas.org/content/117/12/6370.short>

- Kaelbling, L. P., & Lozano-Pérez, T. (2011). Hierarchical task and motion planning in the now. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1470–1477). IEEE. <https://doi.org/10.1109/ICRA.2011.5980391>
- Kaplan, A. D., Sanders, T., & Hancock, P. A. (2019). The Relationship Between Extroversion and the Tendency to Anthropomorphize Robots: A Bayesian Analysis. *Frontiers in Robotics and AI*, 5, 135.
- Keijsers, M., & Bartneck, C. (2018). Mindless Robots get Bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*.
- Kevin O'Regan, J. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioural and Brain Sciences*, 24, 939–1031.
- Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *The International Journal of Robotics Research*, 5(1), 90–98. <https://doi.org/10.1177/027836498600500106>
- Kim, M. C., & Hannafin, M. J. (2011). Scaffolding problem solving in technology-enhanced learning environments. *Educational Technology Research and Development*, 59(2), 113–147. <https://doi.org/10.1007/s11423-010-9172-6>
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K. & Kiverstein, J. (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138), 20170792.
- Kirsh, D. & Maglio, P. (1994) On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, 18(4), 513–549.
- Koller, M., Weiss, A., Hirschmanner, M., & Vincze, M. (2023). Robotic gaze and human views: A systematic exploration of robotic gaze aversion and its effects on human behaviours and attitudes. *Frontiers in Robotics and AI*, 10, 1062714.
- Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., & Wykowska, A. (2021). It's in the eyes: The engaging role of eye contact in HRI. *International Journal of Social Robotics*,

13(3), 525–535.

Kozima, H., Nakagawa, C., & Yasuda, Y. (2009). Keepon: A playful robot for supporting social interaction. *International Journal of Social Robotics*, 1(1), 3–15.
<https://doi.org/10.1007/s12369-008-0009-8>

Ku, H., Choi, J. J., Lee, S., Jang, S., & Do, W. (2018a). Designing Shelly, a robot capable of assessing and restraining children's robot abusing behaviours. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 161–162.

Ku, H., Choi, J. J., Lee, S., Jang, S., & Do, W. (2018b). Shelly, a Tortoise-Like Robot for One-to-Many Interaction with Children. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 353–354.

Laitinen, A. (2016). Robots and human sociality: normative expectations, the need for recognition, and the social bases of self-esteem. In J. Seibt, M. Nørskov, & S. Schack Andersen (Eds.), *Frontiers in Artificial Intelligence and Applications* (Vol. 290, pp. 313–322).

Lakoff, G. (1999). *Metaphors we live by*. University of Chicago Press.

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., Tschantz, A., Millidge, B., Wisse, M., Buckley, C. L., & Tani, J. (2021). Active inference in robotics and artificial agents: Survey and challenges. In *arXiv [cs.RO]*. arXiv.

Lanillos, P., Ferreira, J. F., & Dias, J. (2014). *Evaluating the Influence of Automatic Attentional Mechanisms in Human-Robot Interaction*.

Lee, W. H., Lin, C. W., & Shih, K. H. (2018). A technology acceptance model for the perception of restaurant service robots for trust, interactivity, and output quality. *International Journal of Mobile Communications*, 16(4), 361.

Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics*, 1(3), 209–216.

Lewis, M., Sycara, K., & Walker, P. (2018). The Role of Trust in Human-Robot Interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of Trusted Autonomy* (pp. 135–159). Springer International Publishing.

Limanowski, J. (2017) (Dis-)attending to the body. in Metzinger, Thomas (Hrsg.). *Philosophy and predictive processing*. Frankfurt am Main: MIND Group. S., 283 – 295.

Limanowski, J. & Friston, K. (2018) Seeing the Dark: Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. *Frontiers in psychology*, 9, 643.

Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

Lombardi, M., Maiettini, E., De Tommaso, D., Wykowska, A., & Natale, L. (2022). Toward an attentive robotic architecture: Learning-based mutual gaze estimation in human-robot interaction. *Frontiers in Robotics and AI*, 9, 770165.

Lombardi, M., Maiettini, E., Wykowska, A., & Natale, L. (2024). Gaze estimation learning architecture as support to affective, social and cognitive studies in natural human-robot interaction. In *arXiv [cs.RO]*. arXiv. <http://arxiv.org/abs/2410.19374>

Lombardi, M., Roselli, C., Kompatsiari, K., Rospo, F., Natale, L., & Wykowska, A. (2023). The impact of facial expression and communicative gaze of a humanoid robot on individual Sense of Agency. *Scientific Reports*, 13(1), 10113.

Lumbreras, S. (2017). The Limits of Machine Ethics. *Religions*, 8(5), 100.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, 15(4), 151–190.

Lyons, J., & Nam, C. S. (2021). Introduction: The evolution of trust in human-robot interaction. In *Trust in Human-Robot Interaction* (pp. xxi–xxv).

MacIntyre, T. & Moran, A. P. (2007) A Qualitative Investigation of Meta-Imagery Processes and Imagery Direction among Elite Athletes. *Journal of Imagery Research in Sport and Physical Activity*, 2(1).

Malle, B. F. & Scheutz, M. (2015) ‘When will people regard robots as morally competent social partners?’, in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 486–491.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124).

Marchesi, S., & Wykowska, A. (2024). Designing robots that are accepted in human social environments: Anthropomorphism, the intentional stance, cultural norms and values, and societal implications. In *The De Gruyter Handbook of Robots in Society and Culture* (pp. 63–84). De Gruyter.

McGeer, T. (1990). Passive dynamic walking. *The International Journal of Robotics Research*, 9(2), 62–82. <https://doi.org/10.1177/027836499000900206>

Montemerlo, M., & Thrun, S. (2007). *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*. Springer. <https://doi.org/10.1007/978-3-540-46402-0>

Montemerlo, M., Thrun, S., Koller, D., & Wegbreit, B. (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-02)* (pp. 593–598). AAAI Press.

Moradi, M., Moradi, M., & Bayat, F. (2018). On robot acceptance and adoption a case study. In *2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*.

Marewski J. N., Gaissmaier, W. & Gigerenzer, G. (2010) Good judgments do not require complex cognition. *Cogn. Process.* 11 103–121.

Mazzei, D., Zaraki, A., Lazzeri, N., & De Rossi, D. (2014). Recognition and expression of emotions by a symbiotic android head. In *Humanoid robots (humanoids), 2014 14th IEEE-RAS International Conference on* (pp. 134–139).

Metta, G., Sandini, G., Natale, L., Craighero, L., & Fadiga, L. (2006). Understanding mirror neurons. A bio-robotic approach. *Interaction Studies*, 7, 197–232.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot.

Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems. PerMIS '08: Performance Metrics for Intelligent Systems, Gaithersburg Maryland.

Milne, M., Leibbrandt, R., Raghavendra, P., Luerssen, M., Lewis, T., & Powers, D. (2013). *Lesson authoring system for creating interactive activities involving virtual humans the thinking head whiteboard*. 13–20.

Milne, M., Luerssen, M. H., Lewis, T., Leibbrandt, R., & Powers, D. (2011a). Embodied conversational agents for education in autism. *A Comprehensive Book on Autism Spectrum Disorders*, 387.

Milne, M., Luerssen, M., Lewis, T., Leibbrandt, R., & Powers, D. (2011b). *Designing and evaluating interactive agents as social skills tutors for children with autism spectrum disorder*. 23–48.

Milne, M., Powers, D., & Leibbrandt, R. (2009). Development of a software-based social tutor for children with autism spectrum disorders. *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, 265–268.

Milne, M., Raghavendra, P., Leibbrandt, R., & Powers, D. M. W. (2018). Personalisation and automation in a virtual conversation skills tutor for children with autism. *Journal on Multimodal User Interfaces*, 12(3), 257–269.

Montero, B. (2016) *Thought in Action*. Oxford University Press.

Movellan, J. R., Eckhardt, D., & Rodriguez, J. J. (2009). RUBI: An immersive tutoring robot for language learning. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 293–300). <https://doi.org/10.1109/HRI.2009.4909504>

Mubin O., Cappuccio M.L., Alnajjar F.S., Ahmad M., and Shahid S. (2020), “Can a robot invigilator prevent cheating?”, *AI and Society*, 2020 [CP11].

Mullen, R., Hardy, L. & Oldham, A. (2007) Implicit and explicit control of motor actions: revisiting some early evidence. *British Journal of Psychology*. 98(Pt 1), 141–156.

Nakajima, K., Hauser, H., Kang, R., Guglielmino, E., Caldwell, D. G., & Pfeifer, R. (2013).

A soft body as a reservoir: case studies in a dynamic model of octopus-inspired soft robotic arm. *Frontiers in Computational Neuroscience*, 7, 91.

Nakamura, T., Sato, T., Kuniyoshi, Y., Hiraki, K., Shibata, T., Asada, M., MacDorman, K. F., & Tani, J. (1999). Why is cognitive robotics promising? *Journal of the Robotics Society of Japan*, 17(1), 38–43.

Naneva, S., Sarda Gou, M., Webb, T. L., & Prescott, T. J. (2020). A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots. *International Journal of Social Robotics*, 12(6), 1179–1201.

Narvaez, D. (2006). Integrative ethical education. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 703–733). Lawrence Erlbaum Associates.

Neau, C. B. M., Santos, P., Bosser, A.-G., & Beu, N. (2022). *Commonsense Reasoning for Identifying and Understanding the Implicit Need of Help and Synthesizing Assistive Actions*. 3121.

Nehaniv, C. L., Forster, F., Saunders, J., Broz, F., Antonova, E., Kose, H., Lyon, C., Lehmann, H., Sato, Y., & Dautenhahn, K. (2013). Interaction and experience in enactive intelligence and humanoid robotics. *2013 IEEE Symposium on Artificial Life (ALife)*. 2013 IEEE Symposium on Artificial Life (ALife), Singapore, Singapore.

Nezami, M.O., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression. In *Machine Learning and Knowledge Discovery in Databases* (pp. 273–289). Springer International Publishing.

Newen, A., De Bruin, L., & Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford University Press.

Nilsson, N. J. (Ed.). (1984). *Shakey the robot*. SRI International, Artificial Intelligence Center, Technical Note 323.

Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT Press.

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, 17(3), 347–369.

Nørskov, M. (2014). Human-robot interaction and human self-realization: reflections on the epistemology of discrimination. In J. Seibt, R. Hakli, & M. Nørskov (Eds.), *Frontiers in Artificial Intelligence and Applications* (Vol. 273, pp. 319–327).

Nørskov, M. (2017a). *Social Robots: Boundaries, Potential, Challenges*. Taylor & Francis.

Nørskov, M. (2017b). Technological dangers and the potential of human–robot interaction: a philosophical investigation of fundamental epistemological mechanisms of discrimination. In M. Nørskov (2017a), pp. 99–122.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield International.

Onuki T (2014) “Designing robot eyes for communicating gaze”, *Interaction Studies*, 14(3), 451–479.

Paine, R. W., & Tani, J. (2004). Motor primitive and sequence self-organization in a hierarchical recurrent neural network. *Neural Networks: The Official Journal of the International Neural Network Society*, 17(8–9), 1291–1309.

Paine, R. W., & Tani, J. (2005). How hierarchical control self-organizes in artificial adaptive systems. *Adaptive Behaviour*, 13(3), 211–225.

Palinko, O., Bodenhagen, L., Cabibihan, J.-J., Fischer, K., Šabanović, S., Winkle, K., Behera, L., Ge, S. S., Chrysostomou, D., Jiang, W., & He, H. (Eds.). (2025). *Social robotics*. Springer Nature.

Parr, T. & Friston, K. (2017) Working memory, attention, and salience in active inference. *Scientific Reports*, 7, 14678, 1-21.

Parr, T. & Friston, K. (2018) The anatomy of inference: Generative models and brain structures. *Frontiers in Computational Neuroscience*, 12.

Peeters, A., & Haselager, P. (2021). Designing virtuous sex robots. *International Journal of Social Robotics*, 13(1), 55–66.

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and social robotics: A systematic review. *Autism Research: Official Journal of the International Society for Autism Research*, 9(2), 165–183.

Petitot, J., Varela, F.J., Pachoud, B., & Roy, J.-M. (Eds.). (2000). *Naturalizing phenomenology*. Stanford University Press.

Pfeifer, R. (1995). Cognition - perspectives from autonomous agents. *Robotics Auton. Syst.*, 15(1–2), 47–70.

Pfeifer, R. (2000). On the role of morphology and materials in adaptive behaviour. *From Animals to Animats*, 6, 23–32.

Pfeifer, R. (2006). *How the body shapes the way we think: A New View of intelligence*.

Pfeifer, R., Iida, F., & Gómez, G. (2006). Designing Intelligent Robots. *Journal of the Robotics Society of Japan*, 24(7), 783–790.

Pfeifer, R. & Gómez, G. (2005). Interacting with the real world: design principles for intelligent systems. *Artificial Life and Robotics*, 9(1), 1–6.

Pfeifer, R. & Gómez, G. (2009). Morphological computation – connecting brain, body, and environment. In *Creating Brain-Like Intelligence* (pp. 66–83). Springer Berlin Heidelberg.

Pfeifer, R. & Iida, F. (2004). *Embodied artificial intelligence: Trends and challenges*. 1–26.

Pfeifer, R., Iida, F., & Gómez, G. (2006). Morphological computation for adaptive behaviour and cognition. *International Congress Series. Excerpta Medica*, 1291, 22–29.

Pfeifer, R., Iida, F., & Lungarella, M. (2014). Cognition from the bottom up: on biological inspiration, body morphology, and soft materials. *Trends in Cognitive Sciences*, 18(8), 404–413.

Pfeifer, R., Lungarella, M., & Iida, F. (2007). Self-organization, embodiment, and

biologically inspired robotics. *Science (New York, N.Y.)*, 318(5853), 1088–1093.

Pfeifer, R., Lungarella, M., & Iida, F. (2012). The challenges ahead for bio-inspired “soft” robotics. *Communications of the ACM*, 55(11), 76–87.

Pfeifer, R., Marques, H. G., & Iida, F. (2013). *Soft robotics: the next generation of intelligent machines*. 5–11.

Pfeifer, R., & Scheier, C. (1997). Sensory—motor coordination: The metaphor and beyond. *Robotics and Autonomous Systems*, 20(2–4), 157–178.

Pfeifer, R., & Scheier, C. (2001). *Understanding intelligence*.

Quinlan, S., & Khatib, O. (1993). Elastic bands: Connecting path planning and control. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 802–807). IEEE. <https://doi.org/10.1109/ROBOT.1993.291936>

Raggioli, L., & Cangelosi, A. (2022). Embodied attention in word-object mapping: A developmental cognitive robotics model. *2022 IEEE International Conference on Development and Learning (ICDL)*, 32, 156–163.

Raibert, M. H. (1986). *Legged robots that balance*. MIT Press.

Ramachandran, V. S., & Oberman, L. M. (2007). Broken mirrors: A theory of autism. *Scientific American*, 17(2s), 20–29.

Ramstead, M. J. D., Kirchhoff, M. D. & Friston, K. J. (2019) A tale of two densities: active inference is enactive inference. *Adaptive Behaviour*, p.1059712319862774

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reichenbach, J., Bartneck, C. and Carpenter, J. (2006) ‘Well done, Robot! The importance of praise and presence in human-robot collaboration’, in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 86–90.

Richards, D., & Busch, P. (2003). Acquiring and applying contextualised tacit knowledge.

Richards, D., Vythilingam, R., & Formosa, P. (2023). A principlist-based study of the ethical design and acceptability of artificial social agents. *International Journal of Human-Computer Studies*, 172(102980), 102980.

Riek, L. D., & Howard, D. (2014, April). A code of ethics for the human–robot interaction profession. In *We Robot 2014* (University of Miami School of Law).
<https://robots.law.miami.edu/2014/wp-content/uploads/2014/03/a-code-of-ethics-for-the-human-robot-interaction-profession-riek-howard.pdf>

Riek, L. D., Rabinowitch, T., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 245–246.

Robertson I., Kirchhoff M. (2024) Anticipatory Action: Active Inference in Embodied Cognitive Activity, special Issues of Acting Ahead of Actuality, Radman, Z. (eds) in *Journal of Consciousness Studies*.

Rodogno, R. (2016). Social robots, fiction, and sentimentality. *Ethics and Information Technology*, 18(4), 257–268.

Runswick, O. R., Green, R., & North, J. S. (2019) The effects of skill-level and playing-position on the anticipation of ball-bounce in rugby union. *Human Movement Science*, 69, 102544.

Safron, A., Hipólito, I., & Clark, A. (2023). Editorial: Bio A.I. - From embodied cognition to enactive robotics. *Frontiers in Neurorobotics*, 17, 1301993.

Saldien, J. (2012). *Probo: Character-based social robot in child–robot interaction* (Doctoral dissertation). Vrije Universiteit Brussel. <https://doi.org/10.2307/j.ctv9r04r0>

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joubin, F. (2011). Effects of Gesture on the Perception of Psychological Anthropomorphism: A Case Study with a Humanoid Robot. *Proceedings of the Third International Conference on Social Robotics*, 31–41.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The heritability of autism spectrum disorder. *JAMA: The Journal of the American Medical Association*, 318(12), 1182–1184.

Sandini, G., Metta, G., & Vernon, D. (2007). The iCub cognitive humanoid robot: An open-system research platform for enactive cognition. In *50 Years of Artificial Intelligence* (pp. 358–369). Springer Berlin Heidelberg.

Sandini, G., Sciutti, A., & Morasso, P. (2024). Artificial cognition vs. artificial intelligence for next-generation autonomous robotic agents. *Frontiers in Computational Neuroscience*, 18, 1349408.

Sandini, G., Sciutti, A., & Vernon, D. (2021). Cognitive Robotics. In *Encyclopedia of Robotics* (pp. 1–7). Springer Berlin Heidelberg.

Sandoval E.B., Billingham M., Cappuccio M.L. (2024), “Robot Natives: Ethical Design and Regulations for Baby-Robot Interaction”, in *Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction* [SP19].

Sandoval E.B., Shi J., Sandoval D.C., Li B., Cappuccio M.L., Rosenbaum S. (2021), “A Prototype of a Robot Memory Game: Exploring the Technical Limitations of Human Robot Interaction in a Playful Context” In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion)*, March 8–11, 2021, Boulder, CO, USA. ACM, New York, NY, USA, 5 pages [SP9].

Sandoval E.B., Sosa R., Cappuccio M.L., Bednarz T (2022), “Human–robot creative interactions: Exploring creativity in artificial agents using a storytelling game”, *Frontiers Robotics and AI - Section Human-Robot Interaction* 24 August 2022 [SP8].

Santos, P. E. (1998). *Formalising the common sense of a mobile robot*.

Sarason, I. G. (1988) Anxiety, self-preoccupation and attention. *Anxiety Research*, 1(1), 3–7.

Scassellati, B. (2006). Quantitative metrics of social response for autism diagnosis. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive*

Communication, 2005. ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005., Nashville, TN, USA.

Sharkey, A., & Sharkey, N. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies*, 11(2), 161–190.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27–40.
<https://doi.org/10.1007/s10676-010-9234-6>

Scharre P (2018) *Army of None: Autonomous Weapons and the Future of War*. WW Norton & Company.

Schellen, E., Bossi, F., & Wykowska, A. (2021). Robot gaze behaviour affects honesty in human-robot interaction. *Frontiers in Artificial Intelligence*, 4, 663190.

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 205–222). Cambridge, MA: MIT Press.

Schippers M. C., & Van Lange, P. A. (2006) The psychological benefits of superstitious rituals in top sport: A study among top sportspersons. *Journal of Applied Social Psychology*. 2006:2532–2553.

Schneider, W., & Chein, J. M. (2003) Controlled & automatic processing: behaviour, theory, and biological mechanisms. *Cognitive Science*, 27(3), 525–559.

Schultz, R. T. (2005). Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *International journal of developmental neuroscience: the official journal of the International Society for Developmental Neuroscience*, 23(2-3), 125–141.

Sciutti, A., & Sandini, G. (2021). Modelling the minimal skills necessary for cognitive development: The iCub project and the quest for cognition. In *Meaningful Relations* (pp. 215–244). Academia - ein Verlag in der Nomos Verlagsgesellschaft.

Seemann, A. (2012) (ed.), *Joint attention. Perspectives and Developments*, MIT Press, Cambridge MA.

Seibt, J., Nørskov, M., & Hakli, R. (2014). *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014*. IOS Press.

Seibt, J., Nørskov, M., & Schack Andersen, S. (2016). *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016 / TRANSOR 2016*. IOS Press.

Shepherd V.S., Cappuccio M.L. (2012), “Sociality, Attention, and the Mind’s Eye”, in Axel Seemann (ed.), *Joint attention. Perspectives and Developments*, MIT Press, Cambridge MA, January 2012, pp. 205-242 [SP11].

Sharkey, N., & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies*, 11(2), 161–190.

Shim, J., Arkin, R., & Pettinatti, M. (2017). An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2936–2942.

Simut, R., Saldien, J., & Vanderborght, B. (2012). Social storytelling with Probo: Robot-mediated therapy for children with autism. *Interaction Studies*, 13(3), 306–325.

Simut, R., Saldien, J., Goris, K., Vanderborght, B., & Lefeber, D. (2012). Probo: A social robot for child–robot interaction. In *International Journal of Social Robotics* (Special Issue on Academic and Industrial Research Collaborations), 4(2), 165–177.
<https://doi.org/10.1007/s12369-012-0135-6>

Smith, R. C., & Cheeseman, P. (1986). On the representation and estimation of spatial uncertainty. *The International Journal of Robotics Research*, 5(4), 56–68.
<https://doi.org/10.1177/027836498600500404>

Sparrow, R. (2016) ‘Kicking a robot dog’, in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 229–229.

Sparrow, R. (2017) ‘Robots, Rape, and Representation’, *International Journal of Social Robotics*, 9(4), pp. 465–477.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2), 141–161.

Spivey, M. J., & Huette, S. (2024). The embodiment of attention in the perception-action loop. In *The Routledge Handbook of Embodied Cognition* (pp. 452–460). Routledge.

Stano, P., Kuruma, Y., & Damiano, L. (2018). Synthetic biology and (embodied) artificial intelligence: opportunities and challenges. *Adaptive Behaviour*, 26(1), 41–44.

Stentz, A. (1994). Optimal and efficient path planning for partially-known environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3310–3317). IEEE. <https://doi.org/10.1109/ROBOT.1994.351061>

Stentz, A. (1995). The focussed D* algorithm for real-time replanning. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1652–1659). Morgan Kaufmann.

Sunstein, C. R. (2015). *Choosing not to choose: Understanding the value of choice*. Oxford University Press.

Sutton, J. (2008). Material Agency, Skills and History: Distributed Cognition and the Archaeology of Memory. In C. Knappett & L. Malafouris (Eds.), *Material Agency* (pp. 37–55). Springer US.

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, 5, 15924.

Swann, C., Keegan, R. J., Piggott, D. & Crust, L. (2012) A systematic review of the experience, occurrence, and controllability of flow states in elite sport. *Psychology of Sport and Exercise*, 13(6), 807–819.

Tani, J. (1999). Cognitive Robotics. The constructivist approach towards the problems of “self” and “self consciousness.” *Journal of the Robotics Society of Japan*, 17(1), 25–28.

Tani, J. (2017) *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-organizing Dynamic Phenomena*. Oxford University Press.

Tani, J. (2022). Neurorobotics. In *Encyclopedia of Robotics* (pp. 1–11). Springer Berlin Heidelberg.

Tani, J. (2023). Studies of cognitive neurorobotics based on the free energy principle. *Journal of the Robotics Society of Japan*, 41(7), 609–615.

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. *Neural Networks: The Official Journal of the International Neural Network Society*, 12(7–8), 1131–1141.

Tapus, A., Țăpuș, C., & Mataric, M. (2008). User personality matching with robot personality: A social interaction study. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, 654–659. <https://doi.org/10.1109/ROMAN.2009.5326200>

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thompson, E., & Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends in Cognitive Sciences*, 5(10), 418–425.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., ... Mahoney, P. (2006). Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, 23(9), 661–692. <https://doi.org/10.1002/rob.20147>

Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology*, 14(2), 137–149.

Torrance, S. (2011). Machine ethics and the idea of a more-than-human moral world. *Machine Ethics*, 115–137.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society. Third Series*, 2(1), 230–265.

Turing, A.M. (1950). Computing Machinery and Intelligence. In *Parsing the Turing Test* (pp. 23–65). Springer, Dordrecht.

- Ullman, D., & Malle, B. F. (2018). What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 263–264.
- Vallor, S. (2011). Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century. *Philosophy & Technology*, 24(3), 251.
- Vallor, S. (2015). Moral Deskillling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, 28(1), 107–124.
- Vallor, S. (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind*. MIT Press.
- Van der Smagt, P., Arbib, M. A., & Metta, G. (2016). Neurorobotics: From vision to action. In *Springer Handbook of Robotics* (pp. 2069–2094). Springer International Publishing.
- Vernon, D. (2010). Enaction as a Conceptual Framework for Developmental Cognitive Robotics. *Paladyn, Journal of Behavioural Robotics*, 1(2), 151.
- Vernon, D., Metta, G., & Sandini, G. (2010). Embodiment in cognitive systems: on the mutual dependence of cognition and robotics. In *Advances in Cognitive Systems* (pp. 1–12). Institution of Engineering and Technology.
- Vernon, D., & Sandini, G. (2024). The importance of being humanoid. *International Journal of HR: Humanoid Robotics*, 21(01).
- Vernon, D., von Hofsten, C., & Fadiga, L. (2014). *A Roadmap for Cognitive Development in Humanoid Robots* (2011th ed.). Springer.
- Verschure, P., Kröse, B., & Pfeifer, R. (1992). Distributed adaptive control: The self-organization of structured behaviour. *Robotics Auton. Syst.*, 9(3), 181–196.
- Voss, M., Ingram, J. N., Haggard, P., & Wolpert, D. M. (2005) Sensorimotor attenuation by central motor command signals in the absence of movement. *Nature Neuroscience*, 9(1), 26–27.

- Wada, K., & Shibata, T. (2007). Living with seal robots—Its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5), 972–980. <https://doi.org/10.1109/TRO.2007.906261>
- Wada, K., Shibata, T., Saito, T., & Tanie, K. (2004). Effects of robot-assisted activity to elderly people who stay at a health service facility for the aged. *Robotics and Autonomous Systems*, 45(1–2), 171–180. <https://doi.org/10.1016/j.robot.2003.12.001>
- Wagner, A. R., & Arkin, R. C. (2009). Robot deception: Recognizing when a robot should deceive. *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA)*, 46–54.
- Wainer, J., Robins, B., Amirabdollahian, F., & Dautenhahn, K. (2014). Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 6(3), 183–199.
- Wallach, W., Allen, C. (2009). *Moral machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
- Waytz A, Cacioppo JT, Epley N (2014) Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism, *Perspectives on Psychological Science*, May 1, 5(3), 219–232.
- Wheeler, M. (2008). Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem. *International Journal of Philosophical Studies*, 16(3), 323–349.
- Whitby B (2012) Do you want a robot lover? The ethics of caring technologies. In: Lin P, Abney K, Bekey G (eds) *Robot ethics: the ethical and social implications of robotics*. MIT Press, Cambridge, MA; London, pp 233–248
- Willemse, C., & Wykowska, A. (2018). In natural interaction with embodied robots we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. In *PsyArXiv*.

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology*, 4, 58.

Wine, J.D. (1971) Test Anxiety and Direction of Attention. *Psychological Bulletin* 76 (2): 92–104.

Wirkuttis, N., & Tani, J. (2021). Leading or following? Dyadic robot imitative interaction using the active inference framework. *IEEE Robotics and Automation Letters*, 6(3), 6024–6031.

Wulf, G., McNevin, N. & Shea, C. H. (2001) The automaticity of complex motor skill learning as a function of attentional focus. *The Quarterly Journal of Experimental Psychology: Section A*, 54 (4), 1143–1154.

Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics*, 12(6), 1203–1211.

Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Computational Biology*, 4(11), e1000220.

Yamashita, Y., & Tani, J. (2012). Spontaneous prediction error generation in schizophrenia. *PloS One*, 7(5), e37843.

You S., Nie J., Suh K. and Sundar S.S. (2011), "When the robot criticizes you... Self-serving bias in human-robot interaction," *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Lausanne, Switzerland, 2011, pp. 295–296.

Ziemke, T. (2003). Robosemiotics and embodied enactive cognition," *SEED — Semiotics, Evolution, Energy, and Development* 3(3), 112–124.

Ziemke, T. (2022). Embodiment in Cognitive Science and Robotics. In *Cognitive Robotics* (pp. 213–230). The MIT Press.

Ziemke, T. (2023). Understanding social robots: Attribution of intentional agency to artificial and biological bodies. *Artificial Life*, 29(3), 351–366.

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *Advanced Robotics: The International Journal of the Robotics Society of Japan*, 7(3), 347–360.

Złotowski, J., Sumioka, H., Bartneck, C., Nishio, S., & Ishiguro, H. (2017). Understanding Anthropomorphism: Anthropomorphism is not a Reverse Process of Dehumanization. *Social Robotics*, 618–627.