CONSCIOUS MACHINES

How a conscious machine may be possible

by

Galo Cheung

A Thesis

Submitted to the College of Humanities, Arts and Social Sciences

Flinders University

in Fulfillment of the Requirements for

the Degree of

Master of Arts

in

Philosophy

June 2023

Supervised by Dr Tom Cochrane

TABLE OF CONTENTS

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and

2. the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and

3. 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.

Signed: Galo Cheung

Date: 20th June 2023

BACKGROUND

The term "Artificial Intelligence" (AI) was first used by John McCarthy and others in 1956 at the Dartmouth Summer Research Project on Artificial Intelligence. The project states:

> "Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[1]

In the beginning, computer scientists were optimistic. They believed AI could be achieved in several years. However, sixty-six years have passed and the goal remains unachieved; we are still studying how a machine, or a computer system could learn to do tasks that are usually done by humans.

The inspiration for this thesis was from the fast-growing technology in the field of applied AI. In many areas, AI can do as well as or even better than a human does. Chess is a good example. AI had already championed the game back in 1997. In other domains such as health care, transportation system control, auto-piloting, etc., AI has achieved amazingly well. The goal of AI is to create an artificial system that can perform all humans' tasks. Historically, we are ambitious to create artificial humans, replicas of our own kind: from Talos the bronze giant in the Greek Mythology, to the Homunculus in the Middle Ages alchemists and robots in modern times. However, among those mysterious attempts, there is one hurdle that AI technology seems unable to overcome: consciousness. An artificial object without consciousness can hardly be considered as a human. In other words, to successfully create an artificial human, we need to create consciousness.

Human minds consist of different cognitive properties: sensation, emotion, different kinds of mental content and ability to reason. Consciousness is a state or quality of being aware, sentient, and able to experience subjective sensations and mental states. Consciousness involves various cognitive functions, such as perception, attention, memory, reasoning, and self-awareness.

[1] John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon, (1956), "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf

Consciousness is closely intertwined with these cognitive processes and is often regarded as an emergent property of complex cognitive systems. One perspective on the relationship between consciousness and cognition is that consciousness provides a global workspace where information from various cognitive processes can be integrated and accessed by other processes. In this view, consciousness may play a crucial role in attention. In this thesis, I consider the term 'intelligence' in the broadest sense to include various mental activities. Yet as the title of this thesis suggests, my primary concern is consciousness. More precisely, it is about how to construct artificial consciousness.

One way to approach this question is with an analogy with other human capacities. For example, the question 'Can we build an artificial heart?' is, in fact, the question "Can we build an artificial device that pumps blood in an identical way to a heart?". To give a plausible answer, we need to understand what a real heart is. We further need to understand the mechanism of blood pumping, how the heart works, and how its function can be achieved. We have to know the physical and biological structure of a heart. If we can fully understand the principle of how a heart works, then the final hurdle of building an artificial heart is merely a technological problem.

Similarly, we may think that consciousness is something that our brains do just as our hearts pump blood. By the same token, the approach to consciousness is to answer the question of "How can we build a machine that does what a brain does?". No doubt, we need to know the mechanism of mind and how it can be achieved. We also have to know the physical and biological structure of the brain. If we can fully understand the principle of how a brain works, then what remains are purely technological problems.

In this way, this paper hopes to make a meaningful contribution to the study of the theory of mind. With no doubts, understanding how consciousness is achieved is one of the main difficulties in this study. We will be exploring different theories of mind, i.e. the computational theory of mind, Representationalism, Behaviourism and Functionalism, theories of content, and evaluating and analyzing their strengths and weaknesses; and applying the most plausible theory to build a conscious machine. Optimistically, I believe the conscious machine is possible. Hence, I propose this study.

CHAPTERS SUMMARY

This thesis argues that it is possible to build a conscious machine. To do so, we need to know what consciousness is. When we say "we are conscious of something", we tend to mean "we are aware of something". Consciousness allows us to be aware of our surroundings and inner state. However, once we try to pinpoint what exactly consciousness is, it leaves us grasping at thin air. We are conscious of both the external world and our internal selves, but we do not understand what makes us aware of ourselves and the world. Consciousness is such a puzzling and controversial concept; that both philosophers and scientists struggle to define what it is.

Part of the problem here is that consciousness is neither a simple nor a single concept. It consists of different kinds of referent when we use the term "consciousness". When we say a mental state is conscious, we may refer to when it is verbally reportable or internally accessible; when we say we have sensory experiences or feelings, we may refer to our awareness of that experience. "Consciousness" may also refer to thinking processes, such as reasoning. On many occasions, when we use the term "consciousness", we do not clearly distinguish what concept of consciousness we refer to. Nevertheless, as I will detail below, philosophers have tended to converge on the idea of 'phenomenal qualities '.

The controversy of consciousness is that some argue that consciousness is ineffable such that it is not programable (Block 1995), or consciousness is a fundamental property such that it cannot be reduced (Chalmers 1996) further into a lower-order description. Either of these claims implies the failure of the computationalization of consciousness. Without a doubt, this thesis rejects those claims. The first step of our rejection is to argue that consciousness is programable. To do so, as presented in the first chapter, I will aim to give a functional account of consciousness, such that we manage to describe consciousness objectively.

The second chapter then seeks an adequate theory for consciousness. After reviewing several options, I defend higher-order representationalism, a theory of consciousness in terms of higher-order representations. In my account, consciousness can be explained by combining first-order representations, attention and higher-order representations.

After we understand what consciousness is, I shift our focus to computationalizing the mind in chapter three. I present the computational theory of mind, explain how computers manipulate symbols and point out that minds are computations. The introduction of the Turing Machine aims to point out that minds can be reduced into a formal step-by-step computational process. This chapter also raises the need for gaining content for symbols, i.e., representations, which will be the aim of the next chapter.

In Chapter Four, I examine different theories of mental content and argue that success semantics is the true theory of mental content and is also adequate to apply to computers.

In the last chapter, I argue that being alive is necessary for being conscious, which means if we want to build a conscious machine, we must build a living machine first. A *living machine* is a machine that can maintain itself autonomously. The machine has preferences to maintain itself physically as a whole and not fall apart. Furthermore, the machine's preference motivates the machine to use the information it receives to perform different actions to stay alive. At this point, the machine can represent, use information, has an internal preference (to stay alive), and has higher-order representation (success semantics grants its content). Thus, we can claim that the machine is conscious.

CHAPTER ONE

Functionalism

Functionalism is perhaps the most prominent theory of the mind-body relation in philosophy today. Potentially, it allows us to define consciousness in a functional way. Functionalism is concerned with explaining mental states and processes in terms of their functional roles, rather than specific physical or mechanistic details. It emphasizes the idea that mental states are determined by their relations to other mental states, input stimuli, and behavioral outputs. According to functionalism, what matters is the functional organization or pattern of activity that characterizes a particular mental state, not the specific physical or mechanical properties that implement it. It suggests that mental states are defined by their causal relations to other mental states, sensory inputs, and behavioral outputs. Additionally, functionalism allows for the possibility of multiple realizations of mental states. It suggests that different physical systems, not limited to the human brain, could potentially exhibit the same functional organization and give rise to the same mental states.[2] As we want to build a conscious machine, we at least need a tool to objectively justify whether that machine is conscious or not. In our project to build a conscious machine, the "examination" will be based on the behavioural output of the machine. i.e. if the machine's behaviours fulfil all our functional criteria, we can claim that that machine is conscious.

[2] Overall, functionalism focuses on the functional aspects of mental states rather than their physical composition. In other words, mental states are defined by their functions and roles within the overall cognitive system.

In contrast, identity theory asserts a strict one-to-one correspondence between mental states and brain states, implying that mental states can only be realized by specific neurophysiological processes. According to identity theory, mental states are not merely correlated with brain states; they are the same thing.

While both functionalism and identity theory seek to explain the relationship between the mind and the physical world, they differ in their emphasis. Functionalism emphasizes the functional organization and causal relations of mental states, whereas identity theory focuses on the identity between mental states and brain states.

In section two, I will discuss Ned Block's (1995) distinction between phenomenal consciousness and access consciousness, and narrow down our focus upon the former. More importantly, Block claims that functionalism about phenomenal consciousness is false (1995, P.230), and phenomenal-conscious states are distinct from functional properties. In Block's account, phenomenal consciousness is not functionally definable (i.e. explainable) in terms of a computer program. If Block is right, it is a threat to our goal of giving a functional description of consciousness. I will evaluate Block's claim and argue that phenomenal consciousness can indeed be described in a functional way.

In section three, I will present how David Chalmers (1996) proposes another problem for the theory of consciousness, namely, the Hard problem of consciousness. The Hard Problem is a problem of how physical processes give rise to subjective experiences. Chalmers claims that even if we can explain all physical processes or cognitive activities in the brain, the question "why physical processes are accompanied by subjective experiences" would remain. In other words, there is an explanatory gap between objective physical processes and subjective experiences. In response to the hard problem, the phenomenal concept strategy offers a plausible reply. In section four, I will present the phenomenal concept strategy and see how it handles the hard problem. Once we "dissolve" the hard problem, the obstacle on our track has been cleared, and we can turn our focus towards searching for the right theory of consciousness.

Section One: The functional account of consciousness

Traditionally, the philosophical study of consciousness has divided into two branches: dualism and materialism. Dualists claim that the mind is a non-physical phenomenon, while materialists claim that the mind is physical; all mental states are physical states. Meanwhile, functionalism is, at least initially, neither dualist nor materialist; it gives an account on a more abstract level than physical processes happening in the brain. On a functionalist's view, what a human mind is does not depend on what substance it is made of, rather it depends on its causal role in relation to the circumstances and other states of mind.

However, somethings are defined by what they are made of, for example, water. Water is H2O. It may be in liquid form or ice form, or gas form, but fundamentally it has to be H2O. Anything that is not H2O, then that is not water. Nevertheless, many things are defined not in terms of what they are made of, but in terms of what their functions are. Keys are a good example. A key is not defined in terms of what it is made of, e.g. it can be made of metal, wood, plastic, or even be constituted by digital code. Being a key is not a matter of being a specific material, but it is a matter of what it can do or is supposed to do, that is, to open a lock. The job that a key can do, or is supposed to do, is the function of the key. Note that, the "key job" can be realised in multiple ways. Multiple realizability of a property implies that there are multiple ways to instantiate the property. What makes two keys have in common is that they both perform the same job, i.e. to open a lock. Meanwhile, the way that how they open a lock can be different, i.e. Key A can open its corresponding lock by inserting it into the keyhole and turning clockwise. Key B can open its corresponding lock by swiping it through the key reader on the lock.

Let us now apply this same concept to minds. According to functionalism, what defines minds can be considered in terms of what their functions are, but not what they are made of. As Ned Block notes, "Each type of mental state is a state consisting of a disposition to act in certain ways and to have certain mental states, given certain sensory inputs and certain mental states." (Block 1978, P.262).

Something is a mental state in virtue of having a particular function. In other words, what makes a mental state the particular state it is, is that it does the job associated with that mental state. For example, let us consider pain. In general, feeling pain is genuinely perceiving tissue damage in a body part. Functionalists claim that a thing is in pain if it has a state inside it that occupies the pain role or does the pain-job (compare to the key-job in the above key example). What is pain-job? Suppose that, someone X accidentally cuts his finger. This circumstance causes pain (pain typically is caused by bodily injury, and it is a signal of bodily damage). The pain-state in X then causes him to perform some behaviours, e.g. saying "ouch", checking the wound, looking for a band-aid, etc. The pain-state in X also causes some other mental states, such as the state of distress, desire to make the pain go away, belief about the location of the injury, etc. The pain-state causes conjunctions of those beliefs and desires. In short, pain is the detection of bodily damage that is evaluated as bad and triggers responses that tend towards the reduction of further harm.

Given this functional role, we can imagine different creatures occupying it. We can imagine pain in humans, in octopuses (Putnam 1967) and in Martians (Lewis 1980), as sharing this functional profile. In humans, pain is a state that has a particular type of sensory input (e.g. the cut on my finger), has a particular relation with other mental states (e.g. distress), and has a particular type of behavioural output (e.g. screaming). Similarly, the pain in an octopus is caused by body damage, and it causes distress. It also causes to escape from the harmful stimuli.[3] Thus, any state with this causal profile is a pain state, where it does not matter if this is realised in a human (octopus' brain structure could be very different from ours) a different type of nervous system state, or a state of silicon chips.

The example of pain and its role aim to illustrate how functionalism defines an object. In fact, the role of pain in human's body is more complicated. For instance, there are different kinds of pain, e.g. burning, different stabbings, paper cut versus knife cut. Different kinds of pain present as a distinctive bodily feeling (Cochrane 2018, P.43). The role of pain is sometimes difficult to explain. One of the essential characteristics of pain is that it is evaluated as bad. However, in some cases, pain even will not cause any unpleasant feelings. A rare condition known as pain asymbolia is such a case. Asymbolia is a malfunction in the brain, such that patients are able to detect pain but report that they do not experience it as unpleasant. The patients are able to tell a stimulus is painful. And even though the stimulus is harmful to their body, they do not experience it as bad.

Although asymbolia individuals do not associate pain and unpleasantness in the ordinary way, their bodily responses to noxious stimuli are similar to normal individuals, e.g. increased heart rate, sweating (2018 P.46). Seemingly, our body tends to keep us alive from bodily damages even though we do not think there is any harm.

On the other hand, when someone sincerely claims he is in a pain state, that does not necessarily mean he is receiving physical tissue damage. Phantom pain is pain felt in a part of the body that has been amputated (phantom limb pain), or from which the brain no longer receives signals. Phantom pain brings out a problem of the commonsense theory of pain. We generally

[3] We have good reasons to believe octopus can feel pain as we do, by observing them and studying their nervous system, which is similar to humans.
Jane A. Smith, Ph.D., A Question of Pain in Invertebrates,ILAR Journal, Volume 33, Issue 1-2, 1991, Pages 25–31, https://doi.org/10.1093/ilar.33.1-2.25

think that pain is an unpleasant sensory experience associated with actual or potential tissue damage. However, in the case of phantom pain, e.g phantom limb, as the limb does not exist, there is no tissue damage. But the patient claims he feels the pain from his absent limb. It contradicts our concept of pain that feeling pain is genuinely perceiving tissue damage in a body part. There is no sensory perception. The phenomenon of phantom pain may lead us to reconsider the nature of pain. The commonsense concept of pain considers pain is something in a body part. However, the phenomenon of phantom pain shows that pain is not something in a body part, instead it is more like a psychological event, we have to experience it. Under this concept, pain is a subjective experience. The existence of pain seems to depend on feeling it.

To fully discuss theories of pain is beyond the scheme of this thesis. Yet what else could pain be if the “pain” does not have the pain-jobwe mentioned above? If a “pain” does not cause any distress, unpleasant feeling, or desire to escape, could that “pain” be the pain we ordinarily refer to? All that matters about a mental state (our example is “the pain state”) is its relation with the sensory inputs, other mental states, and the behavioural outputs. Thus we know that individuals with asymbolia experience pain differently because their functional profile is different. A state of an entity is a pain in virtue of having a particular function in the overall system. In principle, many different sorts of things (human brain, octopus “brain”, even Martian “brain”, machinery silicon chip network, etc.) can do the pain-job.

With this example of pain, we can see that functionalism potentially offers an account of how consciousness works. According to functionalism, a conscious state is a state that is constituted solely by its functional role. In other words, whether a mental state is conscious or not is a matter of its causal relations with other mental states, sensory inputs and behaviors. However, Block (1995) has identified two types of consciousnesses, i.e. Access consciousness and phenomenal consciousness, and he has argued that only access consciousness can be given such a functional account.

Section Two: Access Consciousness and Phenomenal Consciousness

1: Access Consciousness

According to Block, access consciousness is defined by its “availability for use in reasoning and rationally guiding speech and action.” (1995, P.227).

According to Block (1995), a mental state is access conscious if :

1. It is a representation;
2. Inferentially promiscuous (Stich 1978), that is, poised for use as a premise in reasoning;
3. Poised for rational control of the action;
4. Poised for rational control of speech.

Block states that those four conditions are sufficient, but not all necessary. As he regards (4) is not necessary to allow non-linguistic animals, such as chimpanzees, to have an access-conscious state. Access consciousness plays a role in our thought, talk and action. For example, my occurrent belief that “Tomorrow is Thursday” is fully accessible according to Block’s criteria. “Tomorrow is a Thursday” is a (1) representation. “Tomorrow is Thursday” fulfils (2) as it certainly can be a premise in all sorts of reasoning; My thought “Tomorrow is Thursday” is available for rational control of action (3), that is, if you ask me to prove to you that Tomorrow is Thursday, I can prove it by showing you that today is Wednesday. My belief “Tomorrow is Thursday” is available for (4) rational control of speech, as if you ask me what the day tomorrow is, I can clearly say “Thursday”. In general, access consciousness involves thinking and control of behaviours. These can be identified as mechanical functions. In terms of mechanical functions, it can be studied objectively. It also implies that access consciousness, in Block’s view, can be scientifically studied, examined, or experimented.

2: Phenomenal Consciousness

Besides access consciousness, there is a distinctive consciousness that Block calls Phenomenal Consciousness. Phenomenal consciousness involves subjective experiences or qualia. Those experiences come from our sensory perceptions, such as our visual experiences, auditory experiences, sense of taste, sense of smell, and sense of touch. Block adapts Nagel's notion (1974) that "what makes a state phenomenally conscious is that there is something "it is like" "(Block 1995, P.228). This allows for subjective differences in experience. For example, the visual experience can be different between a person who has normal colour vision and a person who has only black-and-white vision, even they are looking at the same object. What the Sun is like to me is red, but what it is like to a person who only has black-and-white vision will be black and white only (perhaps grey as well). That is a difference in phenomenal consciousness.

Block further adds that phenomenal consciousness is distinct from the information-processing function, but rather it is part of the information-processing implementation (P.229). Phenomenal consciousness does not do any information process, but it does have a role. What role is that? "The idea is that phenomenal consciousness really does something, that it is involved somehow in powering the wheels and pulleys of access to the Executive System" (P.229). "Perhaps there is something about P-consciousness (Phenomenal consciousness) that greases the wheels of accessibility." (P.242) Block explains this view with his hydraulic machine metaphor, "For example, phenomenal consciousness might be like water in a hydraulic computer. You do not expect the computer to work normally without the water." (p.229)

The water may not perform any information-processing function in the hydraulic operation. e.g. the water itself does not help in doing any calculation, but the liquid itself is necessary for the hydraulic system. Nevertheless, the water does perform some functions in the system. For instance, the water can function as supporting the structure of the machine, or as amplifying the input force. In the same way, perhaps phenomenal consciousness functions to support access consciousness operating. Block's main point is that we may not know the function of phenomenal consciousness, but it is necessary for the information-processing system, as the water to the hydraulic machine.

3: Access Consciousness vs Phenomenal Consciousness

Block distinguishes three differences between phenomenal consciousness and access consciousness (p.232). The first difference is that phenomenal consciousness is phenomenal, and access consciousness is representational. He further adds, "P-conscious states, by contrast [to Access conscious states], sometimes are, and sometimes are not, transitive" [P.232]. Such that phenomenal consciousness can be "not conscious of" (P.232). If a state is transitive, then that state is in relation to other objects, states, or events. Hence, a "transitive" phenomenal consciousness refers to a phenomenal-conscious state that is constituted in relation to other objects, or states, or events.

Meanwhile, Block also points out that a phenomenal-conscious state can be intransitive. If transitive consciousness is about being conscious of something, then intransitive consciousness is about being in a state of consciousness. For example, being awake is being in a state of consciousness. In contrast, access consciousness, essentially plays a role in reasoning, and only representational contents that can figure in reasoning. Representational content refers to the meaning or information that is conveyed by a mental representation, such as a thought or perception. A mental representation is said to have representational contents if it stands for, or represents, some aspect of the world outside of the mind. For example, when I think of a tree, my mental representation of the tree has representational contents because it represents or refers to an actual tree in the world.

The second difference is that access consciousness is a functional notion. In section 2.1 we have discussed that the function of access consciousness involves thinking and control of behaviours. What makes a state access-conscious depends on its representational contents of that function in the system. In comparison, phenomenal consciousness is not a functional notion. The phenomenal-conscious contents of a state are the experiential properties, i.e. what it is like to be. Note, however that Block agrees that the content of an experience can be both phenomenally conscious (subjective feelings) and access conscious (representational properties).

The third difference is that phenomenal conscious state is essentially phenomenal conscious. Where "essentially" refers to the intrinsic qualities, or the fundamental nature of a particular experience. For example, the experience of pain is a phenomenal-conscious state. There can be different kinds of pain, but all kinds of pains are essentially phenomenal. Access-conscious contents may change at different times. It is true at one time does not imply it is true at different times. A representational content that is inferentially promiscuous now may not be so later. For instance, my belief that "Tomorrow is Thursday" can only be true for reasoning when today is Wednesday, but not on other days.

So far, we know that access consciousness and phenomenal consciousness have different characters. In most cases, it seems that they come hand-in-hand. When we are access-conscious of something, this implies we are aware of something by having a mental representation (thought) about it. As such we are aware of the content of that thought. If we are able to have a thought (mental representation) of something, we are as well able to experience that thing. For example, when I am aware of a pain on my finger, I am access-conscious of it as I am aware of the mental representation of the pain. At the same time, if I am aware of the pain, I as well experience the pain what-it-feels-like.

It is fair to say, if we are access-conscious of something X, we are phenomenal-conscious of X too. Is it possible to have access consciousness without phenomenal consciousness? Perhaps we can imagine a machine that is access-consciously identical to human, but not phenomenal-consciously identical. The machine may believe "This is red", (for instance, the machine sensor detects a wavelength around 700 nanometers, it matches the definition of "red" in the machine's database,) and rationally controls its actions and speeches. Nevertheless, the machine is phenomenally unconscious. i.e. the machine does not have any experience, or what-it-is-like, to the belief "This is red". (About the theory of consciousness and mental content, we will discuss in Chapter Two and Chapter Four.)

On the other hand, Block also suggests that it is possible to have phenomenal consciousness without access consciousness, and it might quite often happen in our daily life. For example, suppose you have an intense conversation. Meanwhile, there is construction work next to your house, and you are aware of the noise all the time. Nevertheless, during your conversation with someone, you are so focused on the conversation that the noise is not accessible for you in the ways that Block outlines. You become access-consciously aware of the noise when the conversation is over. In this case, during the conversation, you are phenomenal-conscious of the noise but not access-conscious of it.

Block's example of phenomenal consciousness without access consciousness is debatable. Some may argue that his example is not adequate. For instance, Baars (1995) objects that Block's example is untested. Baars considers that phenomenal consciousness is the same as access consciousness, implying that there is no case we can have phenomenal consciousness without access consciousness. Church (1998) argues that in Block's example, the subject in fact would show some signs of being access-conscious to the noise. For example, the subject may raise his voice, feels irritated, or forms a belief that something is happening outside of the house.

Block next introduces his controversial claim on phenomenal consciousness,

> “The controversial part is that I take P-conscious properties to be distinct from any cognitive, intentional, or functional property. (Cognitive = essentially involving thought; intentional properties = properties in virtue of which a representation or state is about something; functional properties = properties definable.) "(P.230)

The key issue is that if Block’s claim is correct, such that phenomenal consciousness is distinct from functional properties, this implies that we cannot provide a functional description to phenomenal consciousness. What is worse, if he is correct, it potentially implies that we cannot build phenomenal consciousness through computer programming simply because phenomenal consciousness is not definable. “Definable”, where I refer to something that can be explained precisely and unambiguously in an objective way. Phenomenal consciousness is an experience that is totally private. For instance, my experience of seeing the color blueish-grey is totally my own experience. I can tell someone what it feels like and he may have some ideas of what it feels like, but to understand how exactly the experience feels, he needs to experience it in his own way. As Block asks, “how does it explain what it is like to see something as red in the first place?” (P.231). Block points out that explaining subjective experiences as such “what it is like”, or in other terms, qualia, is a difficult task. The difficulty comes in this way: a functional account of phenomenal consciousness is an objective description; if we can give an objective description to phenomenal consciousness, then it implies that we can give an objective description to subjective experiences. But subjective experiences, such as feelings, are essentially individual. No matter how I objectively describe one’s feelings in terms of biological/neuroscientific description, still, it seems to be reasonable for one to ask “but what would that be like from the inside?”.

At this stage, I have outlined the distinction between phenomenal consciousness and access consciousness. Block has pointed out the problem in dealing with phenomenal consciousness. Next, we will discuss how Chalmers explicitly states the problem of phenomenal experiences.

Section Three: The Hard Problem of Consciousness

In his 1996 book, *The Conscious Mind*, Chalmers distinguishes between the easy problems of consciousness and the hard problem of consciousness. Chalmers uses the term “easy”, which does not mean those problems can be explained effortlessly; instead, they are “easy” because, at least, we have a clear picture of how to approach them. In Chalmers’s view, those so-called “easy” problems are easy because they concern the explanation of cognitive abilities and functions (p.237, 1996). To explain cognitive functions, all we need is a functional account to perform that function. In other words, we can reduce those easy problems, i.e. a certain kind of brain state, into a descriptive account. Technically, for any given easy problem, it is just a matter of time until we can explain it. According to Chalmers, the easy problems of consciousness include the ability to respond to the environmental stimuli, the integration of information, the reportability of mental states, the ability to access its internal states, attention, the control of behaviour, the difference between wakefulness and sleep (Blackwell Companion 2007, P.235)[4]. Chalmers claims that we can explain those phenomena in terms of scientific notions. For example, to explain access and reportability, we need a mechanism that can retrieve information (in the brain)

[4] Chalmers, D. (2007). The hard problem of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 225–235). Blackwell Publishing. https://doi.org/10.1002/9780470751466.ch18

and can make it available for verbal reports; to explain the integration of information, we need a mechanism to explain how to bring information together and use it for other processes.

The “hard” problem is the problem about how experiences (e.g. visual experience of Red) arise from physical processes (such as sensations, or perceptions, or emotions). By recording the brain activities, we know that a corresponding area of the cortex layer will be activated when we have a visual stimulus. Still, we do not know how that brain state is accompanied by experience. In Chalmers’ expression,

> “It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.” (P.236)

The “hard” problem is hard because it is not a problem with cognitive activities. Even we can reduce mental states into functional descriptions, the hard problem still remains, i.e. we still can ask the question “what are subjective experiences and how do they arise?”.

In short, the easy problems are the problems about explaining psychological phenomena and their relative states. In contrast, the hard problem is the question about why and how our conscious experiences arise from those cognitive activities.

We may still expect that experiences arise when cognitive functions are performed, but how experiences arise is a mystery. In Chalmers’ words, “There is an explanatory gap between the functions and experiences.” (p. 238, 1996). Thus, if we expect to develop a theory of consciousness, we must construct an explanatory bridge to cross the gap. Chalmers has briefly discussed two theories of consciousness, which are Crick and Koch’s “Neurobiological Theory of Consciousness” (i.e. 35–75 Hz neural oscillation in the cerebral cortex is the basis of consciousness), Baars’ “Global Workplace Theory of Consciousness” (i.e. the contents of consciousness are contained in a global workspace; a central processor used to mediate communication between a host of specialised nonconscious processors) (P.239-240, 1996). Chalmers further emphasises that both theories, in fact, are addressing the easy problems, i.e. the human cognition, instead of addressing how experiences arise. His point is that even if these functions are correlated with consciousness, it does not explain why physical processes should give rise to consciousness, or how an objective mechanism generates subjectivity.

Chalmers also introduces five alternative strategies that are being used by researchers in dealing with the hard problem. The first strategy is to admit that the hard problem is too difficult for now, and researchers simply choose to tackle the more manageable problems (so they belong to the category of "easy problems"), such as reportability. Undoubtedly, that unambitious strategy does not target the hard problem; nevertheless, Chalmers believes it is still worthwhile as it attempts to explain some unsolved questions (e.g. what is reportability) in cognitive science. Then, optimistically, we can learn something new.

The second strategy is to deny there is a phenomenon called "experience". Researchers who hold this strategy believe that once they have explained the "easy problems", the task of explaining consciousness has been completed. Chalmers objects that this strategy is "A theory that denies the phenomenon "solve" the problem by ducking the question" (p.241).

The third strategy is explaining the hard problem in the "full sense" (p.241). Researchers who hold this strategy believe that their functional model or theory could explain the full subjective quality of experiences. Chalmers objects that the explanation is usually passed over too quickly. It works like magic such that after giving details of the information process, experiences suddenly enter the picture. The question "How experiences arise", in Chalmers' view, remains unanswered.

The fourth strategy is explaining the structure of experiences. For example, arguably, explaining how the visual system distinguishes different colours could account for the structural relations between different colour experiences. Potentially, certain facts about information processing systems could correspond to certain experiences. However, Chalmers again objects that this strategy still does not explain why there is experience.

The fifth strategy is to 'isolate the substrate of experience' (P.241). We know that experiences arise from brain processes, but which process? This strategy tries to identify the process that gives rise to consciousness. Experimental analysis is needed. However, one's experience cannot be directly observed. Therefore, justification can only be done indirectly. Chalmers admits that this strategy is incomplete; it may still shed indirect light on the problem of experience. For a complete theory of consciousness, we need to know which process gives rise to consciousness, and we need an account of how and why.

After examining and rejecting the above strategies for tackling the hard problem, Chalmers suspects that, since all purely physical accounts will suffer from the same question "Why does this process give rise to experience?", there is no account of the physical process that will tell us why experience arises. He further argues that "conscious experience is just not the kind of thing that a wholly reductive account could succeed in explaining. The alternative is to build a nonreductive account of consciousness." (P.244). In other words, the explanatory gap is always there.

As I have mentioned at the beginning of this section, the core idea of the hard problem is whether or not subjectivity can be naturalised. Chalmers argues that consciousness is an irreducible property. Therefore, consciousness is as fundamental as basic physical properties. This claim implies that the world is not only constructed by physical material but also by consciousness. To support his claim, he uses the zombie argument (the term "zombie" was first introduced by Robert Kirk in his paper "Sentience and Behaviour", p.43 in 1974)[5] and the knowledge argument, which explain why we are forced to have that conclusion. The zombie argument runs as follows:

Imagine there are zombies that are exactly functionally identical to us but with no phenomenal experiences. They think, act, respond, and behave just like humans, except they do not feel anything as we do.

[5] ROBERT KIRK, Sentience and Behaviour, *Mind*, Volume LXXXIII, Issue 329, January 1974, Pages 43–60, https://doi.org/10.1093/mind/LXXXIII.329.43

The argument is often used to argue against physicalism, which holds that all mental states and experiences can be reduced to physical processes in the brain. Proponents of the zombie argument claim that it is logically possible, or at least it is conceivable for there to be a being that behaves like a human but has no subjective experience. And if zombies exist, since the zombie would be physically identical to a human but lack consciousness, it would mean that there is something more to the mind than just physical processes in the brain. Hence zombies would pose a challenge to our conventional understanding of the relationship between the physical and the mental.

The key premise of the zombie argument is the idea of conceivability. Chalmers argues that if we can conceive of a world in which zombies exist, then it is logically possible for such a world to exist. One major criticism is that the argument relies on the idea of conceivability, which is a controversial and difficult concept to define. Some may argue that just because we can conceive of something, it does not necessarily mean that it is logically possible for that thing to exist in reality. Daniel Dennett (1995) thinks those who accept the conceivability of zombies have failed to imagine them thoroughly enough: 'they invariably underestimate the task of conception (or imagination), and end up imagining something that violates their own definition' (p. 322.)[6]
The Mary's Room argument is a thought experiment in philosophy of mind proposed by Frank Jackson in 1982[7]. The argument challenges physicalism, the philosophical view that all phenomena can be explained in terms of physical processes and properties. The thought experiment goes as follows:

Mary is a brilliant neuroscientist who has spent her entire life in a black and white room, and has never seen any colors. However, she knows everything there is to know about the physics and neurobiology of color perception. She knows all the physical facts about color perception, including the wavelengths of light that correspond to different colors, and the neurological processes that are involved in perceiving color.

[6] Dennett, D. C., 1991, *Consciousness Explained*, Boston, Toronto, London: Little, Brown. 1995, 'The Unimagined Preposterousness of Zombies', *Journal of Consciousness Studies*, 2: 322–6.

[7] Frank Jackson, (1982). Epiphenomenal Qualia. *The Philosophical Quarterly (1950-)*, *32*(127), 127–136. https://doi.org/10.2307/2960077

One day, Mary is allowed to leave the black and white room and see the world in full color for the first time. The question is: Does Mary learn something new when she sees the colors for the first time, or does she already know everything there is to know about color perception?

According to the physicalist view, Mary should already know everything there is to know about color perception, since she knows all the physical facts about it. However, the intuition behind the thought experiment is that Mary does learn something new when she sees the colors for the first time. This suggests that there is something about color perception that is not captured by the physical facts alone, and that physicalism may not provide a complete account of consciousness and mental states.

The Mary's Room argument has been the subject of much debate and discussion in philosophy of mind. Dualists may argue that it shows the limitations of physicalism and suggests that consciousness cannot be reduced to purely physical processes. Others have offered various responses to the argument, such as denying that Mary can really know everything there is to know about color perception, or questioning the intuition that Mary learns something new when she sees the colors for the first time.

The zombie argument and Mary argument, if both are sound, imply that consciousness is non-physical, that we cannot reduce consciousness to other natural terms. An increasingly popular way to accommodate this idea is the theory of panpsychism. Panpsychism is a position that suggests that consciousness or mind is a fundamental aspect of the universe and exists in all things. According to panpsychism, everything in the universe, from the smallest subatomic particles to the largest structures, possesses some degree of consciousness or mind. In other words, the mind is not something that arises from matter, but rather a property of matter itself. There are different versions of panpsychism[8], but the basic idea is that everything has some kind of consciousness or subjective experience. In other words, the mind is not something that arises from matter, but rather

[8] Pan-experientialism is the view that *conscious experience* is fundamental and ubiquitous.
Pan-cognitivism is the view that *thought* is fundamental and ubiquitous.
https://plato.stanford.edu/entries/panpsychism/#VariContPanp

a property of matter itself. Chalmers and other panpsychists[9] claim that consciousness is a fundamental property that is intrinsic to all matter. While more complex organisms, like humans and animals, have more sophisticated forms of consciousness, a stone is conscious in some senses, and so is an iron bar, a drop of water, a molecule, and so on. Nonetheless, claiming that everything is intrinsically conscious seems as mysterious as subjective experiences arising from the physical.

Nevertheless, let's assume Chalmers' panpsychist view on consciousness is correct. It may be compatible with our goal of building conscious machines. As everything has consciousness, what we need to do is to find the right way or method to put all the consciousness-units together, so that the sum of the consciousness-units is at the human level (or animal level). While practical, this would not be philosophically satisfying. We would still not have explained why human-level consciousness is present. The reason behind this statement is that the practical method of combining individual units of consciousness does not address the underlying philosophical questions about the nature and origin of human-level consciousness. It does not provide insight into why human consciousness emerges or how it is specifically related to the complex biological systems, cognitive processes, and neural networks that humans possess. In other words, while it may be possible to achieve a functional equivalent of human-level consciousness by assembling consciousness-units, this approach does not address the deeper metaphysical or explanatory aspects of consciousness. It does not explain why humans possess a particular level or quality of consciousness, nor does it delve into the nature of subjective experience as a meaningful relationship between a subject and the physical world.

In the next section, we will see one of the most plausible responses to the hard problem. If the response is adequate, the hard problem can be dissolved.

Section Four: The Phenomenal Concept Strategy

[9] For example, Philip Goff is a proponent of panpsychism, and has written several articles and a book on the subject, titled "Galileo's Error: Foundations for a New Science of Consciousness." In his book, Goff argues that panpsychism provides a solution to the hard problem of consciousness, which is the problem of explaining how subjective experience can arise from physical processes in the brain. According to Goff, if we take panpsychism seriously, we can see that consciousness is not just an emergent property of complex biological systems, but a fundamental aspect of the universe itself. This view challenges the traditional materialist view of consciousness, which holds that consciousness is an epiphenomenon of brain activity.

Functionalism asserts that consciousness is defined by its functional role, and the functional role is in relation to other functional states, input and output. Therefore, a functionalist view on consciousness would assert that consciousness has causal efficacies for physical events. However, instead of explaining consciousness in a mysteriously non-physical way, functionalism provides a more straightforward and more uncomplicated account for explaining consciousness in terms of causal relations with input/output and other functional states.

Chalmers' hard problem of consciousness points out that an explanatory gap exists between physical processes (i.e. brain functions) and conscious experiences. The explanatory gap is a term used to describe the difficulty in explaining subjective experiences, such as the way things feel, from an objective, scientific standpoint. It is the gap between our subjective experience of something, such as the way we experience pain or pleasure, and the objective scientific understanding of the physical processes that underlie those experiences. The problem arises because our subjective experiences, or qualia, are inherently subjective and cannot be observed or measured directly. In contrast, scientific explanations rely on objective and measurable data. As a result, it can be difficult to bridge the gap between subjective experience and objective explanation. We can scientifically explain brain functions in terms of physical mechanisms, but we are unable to explain why and how those mechanisms accompanied with experiences. The problem is whether and how mental states can be reduced to physical states, or whether there is something inherently different about the nature of subjective experience that cannot be fully explained by physical processes alone.

One response to the explanatory gap problem is to adopt a form of dualism, which posits that there are two fundamental types of substance in the universe: mental and physical. Dualists argue that subjective experience is irreducible to physical processes and cannot be fully explained by physicalism. They suggest that subjective experience is a fundamental aspect of reality, and that it requires a distinct explanation that goes beyond physical processes.

Recall the zombie argument and Mary's experience of gaining new knowledge, both arguments share a common structure: they begin with an epistemic approach to reach an epistemic explanatory gap and proceed to an ontological gap between the physical and the phenomenal. That leads to a significant divide among materialists in responding to the explanatory gap problem. Generally speaking, there are three strategies to deal with the hard problem of consciousness. The first strategy is to deny there is an explanatory gap. Proponents of this strategy believe that once we could fully understand all physical or functional processes of the brain, we could know how conscious experiences arise (e.g. Dennett (1996)). For example, they argue that the qualitative content of pain can be identified with, or realised by, the stimulation of the neural firing process. They deny that Mary lacks any factual knowledge in her black-and-white room; they also deny that philosophical zombies are conceivable. All in all, they deny there is an explanatory gap. Chalmers identifies that kind of materialist as "Type-A Materialists" (Chalmers 2006).

The second strategy is to accept that there is an explanatory gap, but argue that the gap is an epistemic one, and deny that there is an ontological gap. The proponents of the second strategy admit that Mary does have new knowledge epistemically, but she does not have any new facts about the world ontologically; they admit that philosophical zombies are conceivable (i.e. epistemically, it is possible, we can think of it), but metaphysically impossible. Chalmers identifies that kind of materialist as "Type-B Materialists" (2006). Type-B materialists embrace conceptual dualism combined with ontological monism, which is where the phenomenal concept strategy comes on. Type-B materialists, or proponents of the phenomenal concept strategy, hold that phenomenal concepts are distinct from physical concepts. However, they also hold that phenomenal properties are identical to some physical properties, or phenomenal properties supervene on physical properties with metaphysical necessity.

According to the phenomenal concept strategy, the explanatory gap between phenomenal properties and physical properties arises from the conceptual gap between phenomenal concepts and physical concepts. Phenomenal concepts are conceptually isolated (Carruthers and Veillet 2007), such that they cannot be deduced from other concepts. That is to say, for any phenomenal concept, there is no physical or functional or representational knowledge from which that concept can be deduced. That explains how Mary gains new knowledge when she leaves her room: even though she has all knowledge (physical facts) of colour vision, when she leaves her room and sees colour other than black-and-white for the first time, she learns a phenomenal concept of different colours instantly. That phenomenal concept she just newly learns, induces her to have a new thought about colour. For that new thought, the phenomenal concept strategy would claim, is just a new thought concerning the same physical fact (i.e. colour) that she already knew when she was in the black-and-white room. In other words, Mary’s case is an example of how we understand a fact in a different way. When Mary is in the black-and-white room, she understands facts about colour via scientific knowledge; when she is released, she can understand the facts about colour via her first-person recognition. There is no difference in contents among the “facts”, but there is a difference in how she grasps the facts.

The idea of phenomenal concepts can also explain how a philosophical zombie is conceivable but metaphysically impossible. The zombie is physically identical to me, such that I know all physical knowledge about that zombie. However, since phenomenal concepts are isolated, we cannot deduce any phenomenal ideas (and apply them to the zombie) from the physical concepts we know. Therefore, I can conceive that the zombie has distinctive experiences and yet still hold that the zombie is physically identical to me.

Here is why the explanatory gap arises: phenomenal experiences involve phenomenal concepts, and to explain the phenomenal experiences (e.g. feelings) in physical terms is to explain the phenomenal concepts in terms of physical concepts. However, we have seen that phenomenal concepts are isolated concepts. They cannot be deduced from other concepts; hence physical concepts cannot explain phenomenal experiences.

The phenomenal concept strategy gives a reasonable response to the explanatory gap problem. Following that strategy, conscious experiences are compatible with physicalism, which is the best scenario for our project.

Conclusion

In this chapter, I have explained what functionalism is and why a functionalist account of consciousness is so far the best account for our building-a-conscious-machine project. According to functionalism, what defines minds can be considered in terms of their functions. Following Block's idea, there are two types of consciousness, namely access consciousness and phenomenal consciousness. Access consciousness is defined as the availability for use in reasoning and rationally guiding speech and action. We have fewer problems in explaining these functions.

However, the difficulty is the hard problem, which attaches to phenomenal consciousness, and gives us some reasons to think that subjective experiences of a fact and objective knowledge of a fact, indeed, are the different sides of the same token. The so-called “explanatory gap” argument is a matter of fact about how we grasp knowledge about the world from a different perspective. Therefore, the problem of phenomenal consciousness can be regarded as solved once we can explain the subjective experience, or properties, in terms of functionality.

CHAPTER TWO

The Theories of Consciousness

In the previous chapter, we discussed the functional role of consciousness. To be precise, our focus is on phenomenal consciousness, which refers to the subjective experience of being aware of oneself and the external world. This awareness enables us to adapt to our environment and make decisions based on our perceptions. Philosophers attempt to understand the nature of this complex phenomenon. I have presented a functionalist account of consciousness and defended it against its main objection. Next, we need to explain how consciousness works. The objective of this chapter is to seek an adequate theory for consciousness. To do so, I shall begin with setting up a list of criteria for a good theory for consciousness and examine the potential candidates. By the end of this chapter, I will present the theory of consciousness.

Section One: The criteria for the theory of consciousness

What would make a theory of consciousness to be a successful one? In other words, what are the desiderata for a successful theory of consciousness? I suggest that the theory consists of five criteria to explain the most puzzling aspects of consciousness.

The first criterion for the theory we look for is that the theory should be compatible with scientific measures. We study the relationship between mental state and neural activity. To identify consciousness, we may start by listing the common features of conscious activity. Reportability is an essential tool here. Reportability allows a subject to express her inner state. In the case of human beings, a clear and distinct method to identify one's consciousness is by asking the subject for a report of her experience. Suppose the subject can report verbally or behaviorally in a broader sense with relevant content. Potentially, we might infer from these reports the neural correlates of consciousness. e.g. perhaps, some kind of neural synchrony is associated with the overall level of consciousness in an organism (Prinz 2012).

In other words, scientific study has discovered different facts about how the brain functions, and the objective of our theory of consciousness is to match the facts that science discovers. The ideal theory should be able to explain the scientific facts about mental states, why they are like that, and predict or foresee the functional outcome, undergo scientific examination, be justified under different scenarios and resist any challenge, and provide identifiers for consciousness.

The second criterion is a follow-up to the first criterion and provides a mechanical account of consciousness. If we identify the common features among consciousness, then we ought ideally to give a mechanical account of how they function. Grounding functionalism in physical mechanisms provides a bridge between the abstract functional descriptions of mental states and the concrete details of the physical world. This connection allows for a more comprehensive understanding of how mental phenomena emerge from physical processes.

Furthermore, grounding functionalism in physical mechanisms can address concerns about the epistemic gap between functional and physical descriptions. Critics argue that functional accounts alone cannot fully capture the subjective nature of mental states. By emphasizing the underlying physical mechanisms, this approach offers a potential solution to bridge this explanatory gap, as it provides a more tangible link between subjective experience and the physical processes that give rise to it. For example, representationalism is one of the most prominent approaches towards explaining consciousness. It considers all mental states as representations. Not surprisingly, there are different versions of representation[10]. However, in general, they are either first-order

[10] Mental representations can take various forms, depending on the nature of the information being represented. Here are a few examples:

Perceptual Representations: These representations involve the mental encoding of sensory information from the external environment. For instance, when you recall the image of a beach, you create a mental representation that includes visual details such as sand, water, and palm trees.

Imaginative Representations: Imaginative representations involve mentally recreating or imagining sensory experiences that are not currently present. This can include vivid mental images, sounds, smells, tastes, and tactile sensations. For example, when you imagine the taste of your favorite food, you are creating an imaginative representation.

Conceptual Representations: Conceptual representations involve the encoding and storage of abstract knowledge and ideas. These representations capture the essence and characteristics of concepts and categories. For instance, your mental representation of the concept "dog" may include features such as four legs, barking, and fur.

Propositional Representations: Propositional representations involve the encoding of relationships and connections between concepts or ideas. These representations typically take the form of propositional statements or mental sentences. For example, when you think, "If it rains, then I'll bring an umbrella," you are using propositional representation to express a conditional relationship.

representational theory or representations plus some extra factor, such as some way of manipulating representations. When we talk about manipulation, we often compare how the brain manipulates representation to how a computer manipulates its input symbols.

Third, the theory of consciousness needs to explain the character of experience. To be sure, we want to know what makes the difference amongst different mental states. The third criterion requires the theory of consciousness to align with the phenomenal character of our subjective experiences. Conscious experiences have a character that is discernible to careful introspection. For instance, it is plausible that a mental state is not conscious if we are not aware of it. In other words, the introspective characteristic is a feature of consciousness. A good theory should be able to identify this feature.

The fourth criterion is that, ideally, we can explain why, from a functional point of view, we have conscious experiences. What is the purpose of having consciousness? What is the purpose of having phenomenal experiences toward different perceptions, beliefs or thoughts? Or, more precisely, why are we conscious? In nature, it appears that not all organisms are conscious, e.g. bacteria. Nevertheless, without consciousness, bacteria can still survive reasonably well. If consciousness is not necessary for survival, then why do we have conscious experience? A good theory will help us to understand the role of consciousness better.

The fifth criterion aims to justify our intuitions about the consciousness of other beings. We see non-human animals (particularly those that are more complex) seemingly behave consciously, and we see them express what looks like pain when they are injured. So we think they are more or less phenomenally conscious as we are. However, questions arise because some theories of consciousness (e.g. Higher-Order Representationalism), as we will see later, do not recognize non-human animals, and not even human infants, as conscious. But given that the bio-similarity between humans and other animals, and how animals behave and respond to their surroundings, we

Motor Representations: Motor representations involve the mental encoding of actions and movements. They allow individuals to simulate and plan motor actions internally. When you imagine yourself throwing a ball or playing an instrument, you are utilizing motor representations.

These are just a few examples of the different types of mental representations that individuals employ to encode, store, and manipulate information in their minds.

expect our intuition is correct. It is highly unintuitive to most philosophers that animals (and even human infants) are not conscious.[11]

In fact, a survey of philosophers' intuitions about consciousness in other minds shows that only 3.94%[12] of them believe cats do not have any consciousness. Our preferred theory of consciousness should be able to explain consciousness in general cases, but not only in human adult cases.

Short Summary

The five criteria set a frame for our theory of consciousness. A successful theory of consciousness should be able to satisfy those criteria. In the following sections, I will present the most popular approaches to the theory of consciousness, namely, First-Order Representationalism and Higher-Order Representationalism. I will elaborate on their strengths and weaknesses, respectively. Consequently, I will conclude that, in explaining consciousness, both theories cannot cope with the five criteria. Both theories have significant downfalls in their explanatory power. In the following sections, I shall examine the potential candidates for the theory of consciousness. Moreover, by the end of this chapter, I will propose a theory that can better cope with the five criteria for consciousness.

Section Two: First-Order Representationalism

To begin with this section, I shall introduce the idea of representationalism. Representationalism is a position that holds that the world we experience is mediated by mental representations, which are produced by the mind. Mental representation refers to the internal mental states or processes that stand in a certain relationship to the external world. Representationalism asserts that all mental states are representations, for instance, memories, thoughts, mental images, feelings and

[11]"the newborn infant exhibits in addition to sensory awareness specially to painful stimuli, the ability to differentiate between self and nonself touch, sense that their bodies are separate from the world, to express emotions, and to show signs of shared feelings."

Lagercrantz, H., Changeux, JP. The Emergence of Human Consciousness: From Fetal to Neonatal Life. Pediatr Res 65, 255–260 (2009). https://doi.org/10.1203/PDR.0b013e3181973b0d

[12] https://survey2020.philpeople.org/survey/results/5106

desires, are representations. According to representationalism, our senses provide us with information about the world, but it is our mind that constructs a coherent and meaningful mental representation of that information. Mental representations are typically seen as intermediaries between the mind and the world, enabling us to form beliefs, make judgments, and engage in cognitive processes such as perception, memory, and reasoning. This means that mental representations are not just arbitrary or meaningless constructs, but rather they have a structure and content that reflects the structure and content of the world. For example, perceptual states such as seeing a cake, involve the mental representation of sensory information. When you see a cake, your visual system processes the visual stimuli and creates a mental representation that corresponds to the cake's visual attributes, such as its shape, color, and texture. This mental representation enables you to recognize and perceive the cake as an object in your environment.

Desires, such as wanting to eat the cake, also involve representation. A desire is a mental state that represents a particular goal or outcome. In this case, your desire to eat the cake represents your preference, intention, or motivation to consume it. It involves the mental representation of a desired state, where you imagine the experience of eating the cake and the pleasure it would bring.

Both perceptual states and desires can be seen as forms of mental representation because they involve the creation and manipulation of internal models that stand for something external. Perceptual representation captures the perceptual qualities of the cake, allowing you to mentally construct its visual representation. Desire representation captures your subjective inclination or motivation toward the cake, representing your goal or intention to obtain and consume it.

As such, representationalism is used to explain a wide range of mental phenomena, including perception, imagination, memory, language, and reasoning. Similarly, many representationalists argue that thinking process involves the manipulation of mental representations.

In the previous chapter, we have discussed phenomenal consciousness involves subjective experiences. According to Ned Block, phenomenal consciousness is "experience; the phenomenally conscious aspect of a state is what it is like to be in that state." (1995, P.227). Peter Carruthers (2017) further distinguishes phenomenal consciousness “What it is like” into two senses: First, “What it is like” can be in term of the quality of an object that appears to me. For example, an apple is Red in color to me, Carruthers regards this experience (First order) as the property of the world; or Second, “What it is like” can be in terms of the experience that I have of an object that appears to me. For example, I have an experience of how the redness of the apple appears to me, Carruthers regards this experience (Higher order) as the property of the organism's experience of the world. The theory of phenomenal consciousness in this sense is a Higher Order Theory. In short, “Higher-order concept referring to the experience of red rather than a first-order concept referring to red." (2017, p.4).

The key idea of first-order representationalism is that consciousness can be reduced to representations. Moreover, because representations are definable or constructable, they can be naturally linked to our computational theory of conscious machines.

I start by assessing consciousness in First-Order Representational terms, focusing upon Tye's (1995) account. After examining Tye's First-Order Representationalism, I point out that Tye's theory has a vital weakness, his account of consciousness cannot explain the difference between conscious and unconscious experience. According to Tye's theory, conscious experiences are experiences that have impacts on our action-selection, behaviours, or beliefs. However, there are pieces of evidence showing that unconscious experiences will as well have impacts on our behaviours, and cause some meaningful actions, e.g. blindsight patient, as we will see in the later section.

So, what is representation? Literally, representation means to "stand-in for something." For example, when we say, "X represents Y", we mean "X stands-in for Y." Bear in mind that when we say, "X represents Y", it does not mean "X is equivalent to Y". Paintings are typical examples of representation. (But representations need not necessarily be visualisations.) A painting of an apple is a representation of an apple. The painting itself, i.e. the representation itself, represents the apple in the real world. The painting is a two-dimensional object that represents a three-dimensional object. Could a two-dimensional representation carry all the information about that object? It would make sense to think not. A three-dimensional object contains an extra dimension of information, which is beyond what a pure two-dimensional object can contain. Thus, a two-dimensional representation at its best could only carry information that is sufficient to denote the three-dimensional object that it is representing. In general, a visual representation carries limited information about the target it represents.

According to representationalism, all mental states are representations. A mental representation is a representation in our mind processing contents referring to its target, where the target can be an object, an action, a relation or a process. That is to say, all thoughts, beliefs, mental images, feelings, and desires are representations. For example, when I look at the sky, a representationalist would claim that I form a state that represents the sky. And as I have mentioned above, since representation needs not necessarily be visionary, neither does mental representation. The colour of an after-image, the smell of a rose, the taste of wine, the sound of a subject just heard, the texture of a stone, and so on are mental representations. Mental representations have qualities, and we may call those qualities the “first-order senses".

It is a deep problem concerning how exactly mental states manage to denote their targets. However, for the purposes of this chapter focusing on representational theories of consciousness, we will assume that we have some functional accounts of representation. In the following chapter, I will explore the nature of representation in much more depth.

According to First-Order Representationalism (FOR), a conscious state is a representation, and what makes conscious experience arise is determined by the content of the representation available to thought and reasoning, and for the control of action. Phenomenal consciousness is defined as the occurrence of representational content. As Tye puts it, "Phenomenal character (or what it is like) is the same as a certain sort of intentional content" (Tye 1995, P.137). A common rejection of FOR is that there might be some non-representational mental states. There are some conscious states that seem not representing anything, such as moods, pains, emotions, or after-images. Tye responds that pains and other bodily sensations represent parts of the body; while after-images may misrepresent visual qualities, but are still representations. Meanwhile, Tye analyses moods and emotions as representations of bodily responses, "For example, if one feels sudden jealous, one is likely to feel one's stomach sink..." (Tye 2000, P.51).

On the question of what makes a representational state conscious, Tye defends what he calls PANIC theory. PANIC, stands for Poised, Abstract, Non-conceptual, Intentional Content. Tye holds that some representational contents are non-conceptual (N), which means that the subject can lack concepts of what the subject perceives. For example, often, we experience a certain shade of colour that we have never seen before. Conscious states must also have intentional content (IC). Intentional content is here equivalent to representational content, which we have already described. Tye also asserts that intentional content is abstract (A) by which he means that it does not depend on a concrete object for its identity. This feature is needed to handle how hallucinations and veridical experiences could be experientially indistinguishable. Finally, perhaps the most significant notion is "poised" (P). A state is poised if it is ready and available to make a direct impact on beliefs and/or desires. The state in question is a perceptual state, defined by its availability to first-order belief-forming process. In other words, "poised" is a functional notion. Any experience containing all of these qualities is considered phenomenal-conscious. For example, I see a black shadow flying over my head while walking down to the garage. I believe I see a bat. The content is poised because it guides me to believe the shadow was a bat. The content is abstract because it does not require a bat to be present for me to form the believe that it is a bat. The content of the experience is non-conceptual in the case that its content is not f ully constituted by the
correctness conditions of applying the concept 'BAT'.

1: Strengths of FOR theory

Having now outlined the FOR theory, we now turn our focus on what explanatory advantages it provides.

One of the major explanatory advantages of FOR theory is that it provides an account of phenomenal consciousness in animals. According to a FOR theorist, all that an organism needs to be capable of being phenomenally conscious is that it is sophisticated enough to form non-conceptual representations that are poised for the selection of action. It is a direct and clear account of animals' consciousness. Some may argue about how sophisticated is "sophisticated enough". I think as long as the organism is cognitively sophisticated enough to have a certain degree of desires and beliefs, which are capable of having influences upon its behaviours, that will be "sophisticated enough".

Note in particular the important contrast here between FOR and the Higher-order representational approach (to be discussed in section three), which requires the organism to have a stronger capacity to form higher-order thought (where it is suspected that most animals do not have that capability). The result is that according to HOR theory, most animals are not phenomenally conscious, which is a counter-intuitive conclusion. I would be surprised if my dog Bus is not phenomenally conscious, given that he has shown his emotions, desires and feeling through his behaviours. Of course, some may argue that since we cannot "see" his inner states, it is possible that Bus is not phenomenally conscious, i.e. he does not really "feel". But FOR theory provides a more plausible and friendly account for non-humans.

Another advantageous feature of FOR theory is that perceptual states are transparent. The idea of transparency is that experience presents us with the qualities of objects. We do not notice the experiences themselves. For example, when we have a visual experience, we can "hold" it in our mind, pay closer and closer attention to the quality of that experience, and it comes down to pay attention to the quality of the world that it represents. The importance of this feature is that our experience of the world is not distinct from the way it represents the world to us. We can see "through" the experience and directly think of the object that is being represented.

The major challenge to FOR theorists is to explain what makes a perception conscious/nonconsciousness. According to FOR theories, what makes conscious experience arise is determined by the content of the representation available to thought and reasoning and for the control of the action. The content of our mental representations refers to the information and concepts that we hold in our minds, which are derived from our perceptions, memories, beliefs, and other cognitive processes. The specific content of these representations determines the nature of our conscious experiences. For example, if we are thinking about a particular memory, contemplating an idea, or perceiving our surroundings, the content of those mental representations contributes to our conscious experience. However, Carruthers (2000) argues that there is evidence that we have perceptions, and such perceptions indeed affect our selection of action, but we are not conscious of those perceptions/experiences, though they seem to satisfy the conditions for being a FOR. The blindsight cases discussed below provide a concrete example of nonconscious representation.

2: A case of nonconscious representation: Blindsight

The case of blindsight has been studied for decades (Weiskrantz, Warrington, Sanders & Marshall 1974). Patients who have had certain areas of the striate cortex damaged become blind in a particular area of their visual field. They sincerely declare that they are unaware of seeing anything in the blindsight visual region. However, it was later discovered that some of those patients are surprisingly good at "guessing" what is at their "blind" spot. The success rate of the guess is remarkably high, and even surprises the patients (as they thought they were guessing randomly). The data thus shows that they are capable of some nonconscious perceptual discrimination. For example, experiments show that subjects can trace the movement of a light beam, when it is located at their "blind" spot. They are aware of nothing, but still, they can discriminate colours without conscious awareness. (Carruthers 2000, P.155)

In other cases, blindsight patients have shown that they can reach out and grasp objects at their blind-spot, with an accuracy of 80%-90% relative to ordinary sighted people. They can even catch the ball thrown at them. Again, the blindsight subjects claim they are not aware of anything in their blind-spot. (Marcel, 1998, cited in Carruthers 2000, P.156). In another case, a patient with complete bilateral primary visual area (V1) damage was capable of discriminating simulated forms of "optic flow". The subject, without conscious visual experience, could use this ability to walk through a room with obstacles. (Mestre et al. 1992, cited in Carruthers 2000, P.156)

The blindsight cases provide a piece of solid evidence that we can have non-conscious representations, and those non-conscious representations can have impacts on our behaviours. [13]Let us recall Tye's PANIC theory of consciousness, "A state is poised if it is ready and available to make a direct impact on beliefs and/or desires. The state in question is a perceptual state,

[13] Some may argue that non-conscious visual representation is unintuitive. Subliminal priming experiments are a type of study conducted in psychology to investigate the effects of stimuli that are presented below the threshold of conscious awareness, known as subliminal stimuli, on subsequent cognitive processes and behavior.

defined by its availability to first-order belief-forming process." According to Tye's theory, the blindsight patient is unconscious of any visual experience, so the state he is in cannot be poised. Yet, the patient’s state has an impact on his beliefs, desires and actions (he is able to tell what he cannot see, avoid hitting obstacles).

It is plausible that in our daily life, we also undergo some non-conscious representational states and those states have an impact on our behaviours or action-selections. For example, driving while paying no conscious heed to the road; walking while being unaware of what we step on, etc. Those cases convincingly show that we can have perceptions and such perceptions affect our behaviours, but we do not have a conscious experience of those perceptions. Thus, Carruthers argues that FOR theories are unable to give a plausible account for the distinction between conscious and non-conscious experience.

Yet before we move on to Higher-Order Representational theory, can we have a response to the blindsight cases on behalf of the First-Order Representationalist? I think we can. The crucial problem of FOR theory is it cannot distinguish between conscious and nonconscious experience. Perhaps we miss a candidate for transforming nonconscious experiences into conscious experiences. The missing candidate is possibly "Attention".

Section Three: Attention and Consciousness

In the previous section, we have seen that the major problem towards FOR theory is that it cannot identify what makes a representation conscious. After discussing Tye's theory, we turn our focus on

In these experiments, participants are typically exposed to brief and masked stimuli, meaning that the stimuli are presented very quickly and are immediately followed by another visual display that serves to mask or disrupt conscious perception of the initial stimulus. The purpose of this masking is to prevent participants from consciously perceiving the subliminal stimulus.

The subliminal stimulus can be an image, a word, or a symbol that is relevant to the research question being investigated. For example, in a study on the effects of subliminal priming on mood, participants might be exposed to subliminal images of happy or sad faces.

After the presentation of the subliminal stimulus, participants engage in a subsequent task or judgment that is related to the content of the subliminal stimulus. The researchers then examine whether the subliminal stimulus has influenced participants' responses or performance on the task.

The effects observed in subliminal priming experiments are typically subtle and may not be consciously perceived by the participants. However, despite lacking conscious awareness, the subliminal stimulus can still activate certain cognitive processes, influence perception, affect emotional states, or impact decision-making.

considering what First-Order Representational theory misses. In this section, I will present Prinz's theory about Attention and Consciousness. Prinz argues that consciousness depends on attention. Without attending to a certain object or event, we are not conscious of that object or event. He writes,

> “Conscious seems to arise in intermediate-level perceptual subsystems when and only when activity in those systems is modulated by attention. When attention is allocated, perception becomes conscious. Attentional modulation of intermediate level representations is both necessary and sufficient for consciousness. This also makes sense of the Logothetis studies. When presented with conflicting simultaneous stimuli, we may be able to attend to only one. The shifts in attention lead to shifts in consciousness." (2003, p.4)

This is to say, if we are attentive towards something, then we will be conscious of it. In Prinz's words, attention is necessary and sufficient for consciousness. At the end of this section, I will accept the claim that attention is necessary for consciousness; but I will reject the idea that attention is sufficient for consciousness.

1 : Attention

Our key question is “What is the relation between attention and consciousness?”. It is fairly commonsense that attention and consciousness are two phenomena that are closely related. When we pay attention to a certain object or event, we are conscious of it; when we shift our attention away, the object or event fades away from our consciousness. They are so inextricably interwoven that we even find it difficult to identify them distinctively. We want to know whether it is possible to have attention without consciousness, and vice versa.

Attention can be dissected into exogenous (bottom-up) and endogenous (top-down) attention. Exogenous attention, or involuntary attention, is stimuli-driven. For example, an ambulance drives past with its red/blue flashlight on. The flashing light captures our attention even though we do not intend to focus our attention on that light. In other words, this type of attention is perception-related. I need to have a proper functioning visual system in order to be able to see the flashlight, so my attention will be drawn involuntarily by it. Without perceptual faculties, we cannot detect the surroundings or stimuli, exogenous attention will not be effective at all.

In contrast, endogenous attention is voluntary. We focus on something we want to; for example, I am focusing on writing this thesis, and I am voluntarily focusing my attention on this task. Even though I have no exteroceptive faculties, it does not really prevent me from playing attention to my own thoughts. Imagine, if my brain suddenly disconnects from my body, so I instantly lose all my senses. I would immediately think "what has just happened?". More precisely, I can directly concentrate on my mental state.

Prinz (2011) defines attention in terms of reportability to working memory. According to Prinz, attention can be "identified with the processes that allow information to be encoded in working memory. When a stimulus is attended to, it becomes available to working memory, and if it is unattended, it is unavailable." (P.184). "Working memory" is, in Prinz's account, "a short-term storage capacity, but one that allows for "executive control" (Baddely 2007; D'Espisito & Postle, 1999). In general, attention is a process in which we allocate our cognitive resources toward some objects or events, so to gather further information about those objects or events.

Once something is encoded in working memory, it becomes available to language systems for reporting and with systems that allow effortful serial processing. Working memory can play a role in guiding effortful attention (e.g., Cowan, 1995), but it is also where attended perceptual states get temporarily stored (Knudsen, 2007). It is widely recognised that attention is a "gatekeeper" to working memory (Awh, Vogel, & Oh, 2006). Attention determines what information gets in. (Prinz 2011: P.184)

In reality, there is infinite information, how do we choose what to get in? Kentridge states that attention should 'exclude some irrelevant stimuli from consideration' ("Kentridge [2011], p. 229", quoted in Taylor 2018, P.6). Thus, the selection of information is the core part of attention. This is one of the crucial points in the theory of consciousness. If attention functions as a filter of incoming information, then *how does attention select what information should we attend to*? For instance, a mother will be more attentive to her baby's cry than most of us. She responds to her baby’s cry instantly even when she is asleep. It suggests that there are some other factors that determine how attention functions, or how attention prioritizes what information we attend to. This question will be answered in the last section of this chapter.

We further move on to the relation between attention and consciousness. De Brigard and Prinz (2010) claim that "attention is necessary and sufficient for perceptual representations to become conscious" (P.51). Their claim is derived from experimental results. For necessity, they state:

> "In addition, there is a powerful behavioural evidence that consciousness comes and goes with Attention. Working on attentional blink shows that when looking for two stimuli in a rapid series, the first captures our attention, and that results in a brief interval in which we fail to detect the second stimulus. […...] That suggests attention is necessary for consciousness."

In short, this experiment shows that if we lack attention to a certain object, we will not be conscious of that object. Opponents claim that a subject can report the gist of the second image flashed up even when concentrating their attention on other tasks, such that attention is not necessary for consciousness. However, De Brigard and Prinz argue that the fact that the subject that can report the gist of the second image is "predicted by the view that attention is necessary for awareness" (P.58). They claim that “that diminished attention attenuates the amount of detail in a visual representation that can be sent forward to working memory” (P.58). In other words, in the case of gist extraction, attention is not absent, but remains in a minimal state. Thus, the image is already attended to. Hence, the image has already been reported to the working memory, i.e. by De Brigard and Prinz's definition of attention, the subject has attention to that image.

For sufficiency, De Brigard and Prinz again appeal to attentional blink experiments, since the first stimulus captures our attention, and that enables us to report the visual character of the image. In daily life, words and objects that capture attention (such as our own names or a smiley face) are perceived within a split second. That suggests attention is sufficient for consciousness; when attention is captured, invisible stimuli become visible. (2010, P.53-54)

Once an object or event catches our attention, it seems right that we become conscious of that object or event. For instance, a headache may last for 2 hours. In those 2 hours, we may not feel the pain the whole time. Sometimes we forget the pain for a moment when we are attending to something else. We then feel the pain again when we shift our attention back to our body, attend to the headache and be conscious of it. We then suffer from the headache again. This kind of discontinuity of experience happens often.

We usually direct our attention to stimuli or information that we are consciously aware of or find relevant. However, there are cases where information can be processed unconsciously, without conscious awareness, and still influence our subsequent behavior or cognitive processes.

One example is in the realm of subliminal perception, as we discussed earlier. In subliminal priming experiments, participants may be exposed to subliminal stimuli, such as brief flashes of images or words that are presented below the threshold of conscious awareness. Despite not consciously perceiving these stimuli, research has shown that they can still affect subsequent behavior or judgments.

For instance, participants who are subliminally primed with positive words or images may subsequently exhibit more positive attitudes or behaviors without consciously realizing the influence of the subliminal primes.

In these cases, attention is not directed to the subliminal stimuli because they are not consciously perceived. However, the stimuli can still have an impact on cognition or behavior, suggesting that processing can occur without conscious awareness.

While attention generally involves conscious awareness, the example provided highlights that information can sometimes be processed unconsciously, bypassing our conscious attention.

2: The Problem of the Definition of Attention

Prinz's theory about attention and consciousness suggests that attention is the key factor differentiating between conscious and unconscious perception. His theory answers the question "when does consciousness arise?" When we attend to perception, we are conscious of it; when attention is unavailable, we are not conscious of that perception.[14]

Henry Taylor (2013) argues that De Brigard and Prinz's definition of attention cannot be experimentally falsified. He argues,

> "The basic problem is that Prinz does not allow attention, reportability and availability to the working memory to dissociate, and also takes reportability as evidence of phenomenal consciousness. For this reason, it is almost analytic to claim that there is evidence of phenomenal consciousness when there is evidence of attention, and, unsurprisingly, all proposed counterevidence to the claim does not hit the mark." (P.15)

What Taylor points out is that Prinz's definition of attention is functional. As long as a subject fulfils that functional role (i.e. can report the stimuli) then the subject will count as attending to the stimuli (P.14). Since Prinz defines attention in terms of availability to working memory, the threat is that it seems there is no room for the case that a subject is attending in the absence of availability to working memory. If consciousness can only be discerned through report, and report automatically indicates working memory, and working memory automatically indicates attention, then it will never be possible to get consciousness without attention. Any time we try to find a case of attention without consciousness, Prinz can ask "were they able to report?". If they were not, Prinz can deny that they had attention; but if they could report, Prinz can say that they were

[14] Prinz's next step is to locate where consciousness arises. He cites David Marr's work (Marr 1982) on the hierarchies of visual object recognition that there are three stages in processing the perception. The first (low-level) state is about the visual system generating a primal sketch when encountering a stimulus. It can be considered that the primal sketch as a pixel array. Each pixel carries some information about the stimulus, but the pixels "have not been unified with one another to generate a coherent representation of an entire object." (Prinz 2012, P.50). The second (intermediate level) state then generates a two-and-a-half dimensional (2.5D) sketch and unifies the pixels into a coherent representation. The system uses information of the 2.5D to determine the third (high level) state that what three-dimensional forms are currently being perceived. The high-level state abstracts away from textures and other features.

Prinz suggests that consciousness arises at the intermediate level. Here he adapts Jackendoff's (1987) view that consciousness arises "at an intermediate level in sensory hierarchies and not at a low or high level." (Prinz 2012, P.50). The reason is that we simply do not have visual experiences corresponding to the pixels. Nor do we have visual experiences from the high-level state, as in the high-level state, there are abstract forms (i.e. texture, surface features, etc.) of the representation/object. Conscious experiences only arise at this stage where the scene is presented to us as a whole. Thus, conscious experiences arise at the intermediate level.

conscious. Here Prinz is following our criterion 1, trying to follow the science on how we tell whether someone is conscious, but he is not being very informative.

Given Prinz's notion of attention, I admit that attention is necessary for consciousness. One of the reasons is that I cannot imagine a case in which we are conscious of something but do not attend to it. When we are conscious of something, this implies we are aware of it.

However, there should be a deeper sense of when we attend to things. For instance, when we are attending to a red traffic light, the red traffic light draws our attention, but why is that case? Numerous events are happening at the same time around us, but why do we only attend to the red traffic light? This suggests that some other factors determine what we attend to. Possibly there is a more fundamental mechanism in our cognitive system that guides what we should attend. In other words, that cognitive system selects information for us to attend to. We will discuss this kind of information-filtering function in the later chapter.

In Section Four, I will present the Higher-Order Representational (HOR) theory. The HOR claims that a mental state is conscious iff there is a higher-order representation of the first-order state. We will see how this theory explains what a conscious mind is.

Summary on Prinz’s Theory

Does Prinz's theory about attention and consciousness fit into our five criteria for a good theory of consciousness? As I mentioned previously, Prinz's account of attention and consciousness is centered around working memory. As long as there is an availability to working memory, then there is consciousness. Thus, in Prinz's account, to study how consciousness functions is equivalent to study how working memory functions. This offers a chance of using scientific methods to study consciousness, e.g. we can look into the neural network to see how the brain functions when it receives signals and how it stores the signals into working memory. Of course, to discover facts about how brain functions will be the business of neuroscientists. Our task is to use our theory to explain the scientific facts about consciousness. Prinz's theory follows criteria 1 and 2 that he tries to provide a mechanical account for consciousness, explain scientific findings about working memory and use the availability of working memory as the identifier for consciousness.

What's more, Prinz's theory leaves room for the consciousness of other beings. As long as that being can report the information in its working memory, according to Prinz, that being is conscious. Prinz's theory fulfils our expectation that non-human animals can possibly be conscious because, at least, non-human animals seem capable of attention and working memory. Therefore, Prinz's theory also fits criteria 5.

On the other hand, Prinz claims that the character of all conscious experience is that it represents scenes at this intermediary level of organisation. However, this account of consciousness does not fully explain the character of experiences nor the differences amongst different mental states. Therefore, the intermediary level aspect of Prinz's theory only partly satisfies criterion 3.

More importantly, Prinz's theory also does not explain why we have conscious experiences. Speaking of "why", in fact we are hoping a successful theory of consciousness will be able to tell us why we are conscious. Things can exist even if they are not conscious, then why are some beings conscious? One plausible answer may be because that those beings want to maintain their own physical state, which I will discuss this point in chapter five.

Prinz's theory only provides a mechanical account for "how" and "where" consciousness occurs. He does not answer the "why" question, "why is consciousness important to us". Thus, it fails to satisfy criteria 4.

Section Four: Higher-Order Representationalism

Higher-order representationalism (HOR) claims that a mere first-order representation is not sufficient for conscious experience. Higher-order representationalists believe that there is a need for a higher-order mechanism to realise consciousness. The core principle among Higher-order representationalism is the so-called transitivity principle (presented below), which supposes that a conscious state is a state whose subject is aware of being in it. On the other hand, if the subject is unaware of being in the state it would clearly be an unconscious state. In the Higher-order representationalists' terminology, a conscious state is a state whose subject is also part of the representation. For example, having a first-order perceptual state "seeing an apple", even that perceptual state contains certain contents of that apple, if the subject is not aware of who is in that perceptual state, i.e. who does not have a representation of themselves “seeing an apple", then the subject is unconscious of that state.

Thus Rosenthal (1993) writes,

> "Whatever else is true about consciousness, it is clear that a mental state is not a conscious state if one is wholly unaware of being in it. So a necessary condition for a state being conscious is that one be aware, or conscious, of being in that state." (1993, P.2).

That then entails the principle below:

Transitivity Principle (TP): A mental state is conscious, if and only if, the subject is aware of itself is being in that state.

The transitivity principle is the core concept among HOR theories. In general, the transitivity principle applies to all HOR theories. But different HOR theories may interpret the term "aware of" in different ways. For Higher-Order Thought theorists, "aware of" refers to "having a thought about". Since thoughts are representations, in this sense, the term "aware of" can be interpreted as "having a representation of".

For example, when I am looking at the sky, I will have a representation "the sky" in my mind. According to the transitivity principle, the representation "the sky" is conscious, if and only if, I am aware of that I am looking at the sky. Or precisely, for HOR theorists, I have a mental representation "I am looking at the sky”.

The Transitivity Principle has another important implication. Since it claims that a mental state is conscious iff the subject is aware of itself is being in that state, it implies that the content of a conscious state contains the subject itself. We talk about subjects and selves, but what are they? In a minimal sense, a subject can be considered as a system that can maintain its physical boundary from the physical world and can be distinguished individually from the physical world (that includes the system can distinguish itself from others). But under this sense, it seems not enough to distinguish the subject from other objects. Consider ourselves, what makes us distinct individuals is not merely our differences in physical matters, but it also includes the history of each individual and the continuity of thoughts that we have. That is, a "self" has to be a bearer of experiences or thoughts. These two aspects will be sufficient to identify the subject, and so to what a "self" is.

What's more, our motor system also contributes to distinguish ourselves from others. We have a sense of our bodies, which includes our actions, and our emotional reactions. We can tell which belongs to us and which does not. Hence, when I am saying "I", I am confidently able to identify what the content of that "I" is, which is referring to my own existence. And thus, when I am aware that I am looking at the sky, I can confidently identify myself as in a state of looking at the sky.

HOR gives an explanatory advantage over FOR, especially in the case of nonconscious experience. HOR distinguishes what makes a state conscious is to have a higher-order representation of that state. If there is no higher-order representation, then the target state is not conscious.

1: Challenge for the Theory: Flow State

The Transitivity Principle is the core of the HOR theory. It states that I am conscious, if and only if, I am aware of what state I am in. E.g., I am conscious of the state "looking at the sky", if and only if, I am aware of that "I am looking at the sky". In plain words, the Transitivity Principle implies that to claim a subject is in a conscious mental state, the subject needs to know he is in that state. But could there be a conscious mental state that lacks self-awareness? Flow state seems to be one of the cases.

Flow state is an optimal state of mind in which a person is fully immersed in what he is doing. Csikszentmihalyi (1990) describes that optimal experience as "Concentration is so intense that there is no attention left over to think about anything irrelevant, or to worry about problems. Self-consciousness disappears, and the sense of time becomes distorted." (P.71). Such experience happens typically while performing arts or music, playing games, rituals and sports. The way those activities are constructed helps participants and spectators to achieve an ordered state of mind that is highly enjoyable. (P.72)

How does our theory explain the flow state? Is flow state a conscious state? According to the Transitivity Principle, a state is conscious if and only if the subject is aware of itself being in that state. Seemingly, when a person is in a flow state, his attention only focuses on what he is doing. As described by Csikszentmihalyi, the person's self-awareness disappears (if that is the case), which implies that that person will not perform self-monitoring, he will not be aware of what state he is in. i.e. according to the Transitivity Principle, a person who is in a flow state is not conscious of what state he is in.

The claim that one is not conscious of his state when he is optimally enjoying his experience seems counter-intuitive. In particular, I doubt that when one is fully immersed in an activity, one’s self-awareness disappears. Even though one's action and awareness completely merge, that does not imply the self has disappeared. In other words, the claim that the awareness of self is required for consciousness would only be falsified if we consider consciousness is always tied to the same subject-world boundary. In the flow state case, consciousness involves a different sense of the subject, but does not entirely disappear. The subject can still be able to respond to external stimuli, though the stimuli-signal need to be strong enough. Therefore, we can conclude that even in a flow state, self-awareness remains intact, but it is altered.

Section Five: The Theory of Consciousness: FOR, Attention and HOR

Here is my proposal: Let us consider, when I am looking at the sky, I have a first-order state of "looking at the sky" and this first-order state provides me with a bunch of first-order contents, e.g. color of the sky, calmness, relax-feeling, etc. Those contents are formed by my brain automatically. Whenever perception occurs, the brain processes those perceptual data and gets the information ready to be used. That processed information will be stored into short-term working memory; if they are not used, they will be overwritten by the next perceptual data. When we have that ready-to-be-used information, we are not conscious of them yet. My proposal is that there needs to be a trigger for us to request that information.

The trigger factors can be either our desire to self-monitor or the shift of our attention. For the former, we can always be in the desired mode to monitor what is happening to ourselves and what state we are in. Self-monitoring plays a significant role in survival. It will be better to know what state we are in than not knowing it, especially in a wild nature full of dangers. In nature, individuals face various challenges and threats, such as predators, environmental hazards, or the need to acquire food and resources. Having accurate knowledge of our physical state can provide crucial information for making adaptive decisions and taking appropriate actions.

Being aware of our physical state involves factors like hunger, thirst, fatigue, pain, body temperature, and overall health. This awareness allows us to recognize when our bodies require nourishment, hydration, rest, or medical attention. By perceiving these bodily signals, we can respond to our physiological needs in a timely manner, increasing our chances of survival. When we are self-monitoring, we are requesting the information stored in the working memory. When we access the processed information, which is the representation of the sky, we will then be making use of that information. Making use of information is a mental state. Thus, we have a new order representation of that information. In other words, we form a representation of that information, which is a second-order representation of the first-order state. It will be the state "I am looking at the sky".

On the other hand, the request for the information of the current state can also be triggered exogenously. But what draws our attention will depend on our innate knowledge, our past experience and the concepts in our mind that are relevant to our goals and concerns. We have an abundant number of concepts, that help us to understand what is beneficial to our survival and what is harmful to us. Perceptual data are enormous all the time, our cognitive and affective systems would again prioritize the importance of the information. Nonetheless, the process of requesting the information that is stored in the working memory will be the same as how we self-monitor ourselves and how we form a representation of that information.

The difference between the two trigger factors can be further explained. Let us consider, even there is a dinner set available on the table, the consumer still needs to know there is a set of dinner available before she can enjoy it. How would she know there is a set of dinner on the table? (For our metaphor becomes more reasonable, let us set the dinner is prepared spontaneously, it can be there at any time.) In general, there are two ways to let her know.

The first way is that the dinner set can be connected to a signal emitter. Once the dinner is ready, it emits a signal to notify the consumer. The consumer may involuntarily notice the signal. Whether or not the consumer can notice the signal depends on two factors: either she is in a stable/peaceful state such that she has no interference to notify that signal, or the signal is strong enough to override any other inferring signal.

The larger the signal is, the higher the chance draws the subject's attention onto the later-on encoded information. So returning to our "perceiving the sky" scenario, the amplitude of that pulse may be too weak, such that if we are in a not-so-stable state, we would easily omit that information stored in the working memory. Meanwhile, even the signal is vital, if there is another stronger signal entering our brain, we may still omit the original information. We can try an experiment: apply pressure to a pen upon one’s hand, increasing the force gradually until we start feeling some pain. When the pain appears (not so much), we then scream as loud as possible or do something exciting, that would be for a moment we do not feel the pain: we use a stronger pulse to override the original pulse, we physically "relieve" the pain for a second. Similarly, when we have a headache, it usually lasts for hours. But during the period of the headache, if we need to focus on something else, we sometimes neglect the pain. So we physically "relieve" the pain again. (That kind of pain relief is different from taking a pain-killer tablet, which is a chemical way of pain relief).

The second way is that the consumer regularly checks whether or not a dinner set is ready on the table. In the beginning, the consumer may forget to check and miss many dinners, but from time to time she may learn that checking what is available on the table is beneficial for herself (e.g. away from hunger, so can stay alive). Regularly checking the dining table refers to how we monitor our internal state. Certainly, even we do not perform self-monitoring, we can still live. However, for the purpose of survival, if we can always self- monitor what physical state we are in, we would have better chance to survive.

Remember that there may be some signals or pulses that are so weak that they cannot draw our attention. But if we regularly check what physical states we are in, we could notice that and that finding can be critical to our survival. For example, suppose that there is a tiny black spot on my vision that I do not notice during normal activity. If I had monitored my visual state carefully, I could have noticed that black spot. That spot could turn out to be an early sign of retinal detachment.

The theory I propose can be considered a further development of Prinz's theory about attention and consciousness. It adapts Prinz's account of how working memory functions and further develops that we need to *use* the information stored in working memory to claim we are conscious of that information. The concept "to use" plays an essential role in this theory. If we want to use something, that indicates we have a purpose or preference that we want to achieve. For instance, I see a table, I then have a mental representation of that table. If I do not have any purpose that I want to achieve, the mental representation of that table will fade away in my mind. But suppose that I have a purpose that I want to achieve, e.g. I am holding a cup in my hand and I want to put it somewhere. The mental representation of that table (i.e. its information) can be used for helping me to plan some actions (e.g. I can put the cup on the table) in order to achieve the purpose.

The ultimate purpose that an organism has is to survive in nature. Living things have biological tendency to stay alive and avoid being destroyed. These mechanisms can include things like the ability to sense and respond to threats in the environment, the ability to move or hide to avoid danger, the ability to repair or replace damaged tissues, and the ability to adapt to changing conditions. Since living creatures have an internal purpose, it gives them foundation to “use” information. However, not all living creatures are capable of using information. To use information, it requires the living creatures have a certain degree of complexity. Not all living organisms are conscious, for example, a bacterium is a living creature but there is no scientific evidence to suggest that bacteria have consciousness. Consciousness is generally associated with complex neural activity in the brain. Bacteria, on the other hand, are single-celled organisms that lack a nervous system or a brain. While bacteria are capable of sensing and responding to their environment in various ways, such as by moving towards nutrients or away from toxins, these responses are thought to be purely mechanistic and do not involve conscious awareness. A simple life form such as a bacterium, it does not have to be conscious to deal with the challenges from nature. However, complex organisms, such as mammals, face more complicated challenges from their habitat zone. Having consciousness and being aware of the surroundings will be better off for their survival. In short, consciousness is for survival.

One of the strengths of my theory is that it can explain why we are conscious. In general, we are conscious because we have a tendency to be self-monitor what state we are in, which is due to the motivation for survival. In other words, *we are conscious because we want to be conscious (e.g. in order to increase survival chances).* In particular, when we are specifically conscious of a certain thing, it is because we have a preference or interest towards that thing. This idea answers the question we had in Section Three "how does attention select what information should we attend to? "----because we have preference. Our focus is not on what kind of structure will create consciousness. Rather, it is about life and consciousness, the focus is on how being conscious is beneficial for survival.

Does my theory imply that infants and non-human animals lack consciousness? I can answer no. Although infants and non-human animals may not be capable of sophisticated higher-order thoughts, they can still perform self-monitoring. It is the wish to survive that drives them to be aware of their physical states and surroundings, so that they can react correspondingly in order to *maintain themselves*. The motivation to survive leads them to use information (representation) that is required from external stimuli. The only difference between them and us is that the way they use information apparently is different from ours, given that they do not have language. As a result of lacking a language system, infants and non-human animals do not have complicated higher-order thought. Yet, they would still have the ability to “think”, if we consider thinking means “manipulation of representations”. In this sense, infants and non-human animals are not prevented from having higher-order thought. They just may not have the same complex sense of “thinking” that adult humans have. (I will return to the issue of what thinking is in the following chapter.)

Conclusion

In conclusion, I argue that the theory I propose can fulfil the five criteria for a successful theory of consciousness. It provides a mechanical account for consciousness (criteria 1 and 2); It points out that the introspective characteristic is a feature of consciousness. (Criterion 3); it explains why we have phenomenal properties of experience, and it explains why we are conscious (because we want to be conscious) (criterion 4); it accounts for animals and infants would have conscious experience (criterion 5). Thus I propose that this theory is a reasonable basis from which we can proceed to examine whether machines can be conscious.

CHAPTER THREE

The Computational Theory of Mind

This thesis aims to answer the question “How can we build a conscious machine?” After we have understood what consciousness is and how consciousness arises, it is time to turn our focus on how machines can possibly realise thoughts.

In this chapter, I will outline the Computational Theory of Mind, exploring and illustrating its claim that minds are computations. The idea is that minds are representation-manipulation processes while computations are symbol-manipulation processes. Since representations are symbols, minds can be analysed as computations. I will discuss the details of different key terms, e.g. what a symbol is, its content, its properties, how a computation functions, how a Turing machine works, and how this theory explains the nature of mind. At the end of this chapter, I will discuss the objections to the Computational Theory, including the famous Chinese Room argument.

The Computational Theory

“Advances in computing raise the prospect that the mind itself is a computational system—a position known as *the computational theory of mind* (CTM).” (Resorla 2020)[15].

The Computation Theory of Mind is a theory of the mind that suggests that the human mind ***is*** a computational system, processing and manipulating information in a systematic way. According to this theory, mental states and processes, such as beliefs, desires, and perceptions, can be understood as computations that involve the manipulation of symbols and information. Be aware that the computation theory of mind is not metaphorically claiming that the minds is like a computing system, instead this theory is literally holding that the mind is a computational system. This theory can be considered as a machine version of the representational theory of mind. According to the representational theory, intentional states are representations (Cain 2002, P.52). For example, believing that "Apple is sweet" involves being in relation to a mental representation with the

[15] Rescorla Michael (2020) "The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>.

content "This apple is sweet". Representations, in the simplest sense, are symbols that have contents.

Regarding symbols' properties, we will have much more to discuss in the next few sections. However, we pause here and put our effort into understanding what a computational theory of mind is. Let us begin with the characteristics of the computational theory of mind that we are interested in.

The computational theory of mind claims that all mental states are computations. More precisely, the theory states that a computation, i.e. a proper program, is sufficient to realise a mind. We do not know what that program is yet, but the theory guarantees that the project is achievable: a mind truly can be realised computationally. Saying thinking is computation is also saying that mental states such as desiring, believing, feeling, etc, are all computations. Thus, the claim is clear, minds are a complex set of computations.

It is now reasonable to ask what computation is and what features it has.
There are some fundamental tasks that the computational theory has to explain:

1. The computational theory has to define what a formal information process is.
2. If the mind is a computational system, i.e. an information processing system, then what information it processes.
3. If the computational theory is correct, then it has to explain how the mind computes.

For the first task, if the computational theory claims that the mind is a complex set of computations, a further question we would ask is, what is computation? A computational system is a mechanical system that receives an input, processes that input in accordance with its internal mechanical rules, and then yields an output. Since the rules are set, this entails that the same input will always generate the same output. In other words, Computation can be defined as a repeatable input-output process that involves the manipulation of data according to a set of rules or instructions. This process involves taking input data, applying a series of operations or algorithms to it, and producing output data. The input data could be in various forms, such as numbers, text, images, or any other type of data. The output data could be the result of a computation, such as a calculation or a decision. The repeatable aspect of computation means that the same set of input data will always produce the same output data when the same set of instructions or algorithms is applied. This predictability is one of the key features of computation and is essential for the reliable processing of data in a wide range of applications, from scientific simulations to financial analysis to machine learning.

In the book "Introduction to Algorithms" by Cormen, Leiserson, Rivest, and Stein (2009), the authors define an algorithm as “any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output." (P.5). That is to say an algorithm can be thought of as a "mechanical process" that takes inputs, performs a prescribed sequence of computational steps, and produce outputs.

Thus, according to this definition, computation can be seen as a repeatable input-output process that involves following a set of instructions or algorithms to manipulate data and produce a desired result.

The input-output process is a process of manipulating the input. What is involved in the input to a computation? The answer is symbols. A computation is a manipulation of symbols solely based on their internal syntax while disregarding the symbols' semantic contents (Ravenscroft 2005). Here we have three terms that need further clarification: Symbol, syntax and semantic contents. We need to spend some time to explain each term.

Section One: Symbol

In a computer, what it computes is symbols. What is a symbol? And how does a computer handle symbols?

In the simplest sense, a symbol is a sign or mark that represents, indicates or signifies other things or ideas. The content of a symbol is what it represents. For example, the red light at the traffic signifies the meaning "stop at the traffic light". When we see the red light at the traffic, we "go" beyond the red light itself and understand what meaning it really represents, and then we stop at the traffic light. We wait until we see the green light is on---- here we have another symbol-----the green light that represents "Go".

Imagine that we do not use "red light" as a symbol representing the meaning "stop at the traffic light". Instead, we use a description "stop at the traffic light" to let the road users know they need to stop at the traffic light. As a road user, I would be very annoyed if I had to read the whole sentence while driving. Even worse, we will need different descriptions for different purposes, e.g. green light, yellow light, red arrow, etc. That would be practically inefficient and ineffective.

In the traffic light example, the red light is a simple symbol, but it represents a relatively complex meaning "stop at the traffic light" (compare to the Red colour itself). For convenience, we generally choose a simpler symbol to represent a more complex meaning in our daily life.

Another feature of symbols is that a symbol itself does not necessarily resemble the target it represents. Unlike images, which more or less resemble their target, a pure symbol needs no similarity to its target. For example, red traffic lights are, in themselves, nothing like the condition "stop at the traffic light". We cannot understand its denoted meaning solely by looking at the red light. To understand the meaning of a symbol, we need to have some background knowledge about the symbol that is being used. For instance, when we see a red traffic light, we need to know what it means beforehand so that each time we see red traffic lights, we stop. If a person has never learnt what a red traffic light represents, he cannot discern that it means "stop".

So far, we have an idea of what a symbol is, and it will be helpful to distinguish between basic (or simple) symbols and complex symbols. Roughly speaking, a basic symbol is one that has no meaningful parts. I say no "meaningful" parts because it is possible to decompose a basic symbol into a set of further basic components. For example, "Apple" is a basic symbol. It is true that we can further decompose it into "A", "p", "p", "l" and "e". However, decomposing the symbol "Apple" into five letters does not tell us anything more fundamental about the meaning of the symbol "Apple".

One may argue that at least the decomposing process can tell us the components of the symbol

"Apple". But the key point is, even if we can decompose the symbol "Apple" into "A", "p", "p", "l" and "e", those five letters do not tell us any content of the symbol "Apple". The letters do not explain what "Apple" refers to or stands for. Here is the point why I emphasize the term "meaningful": the components of a basic symbol do not contribute to the meaning of the symbol itself.

Meanwhile, a complex symbol is one that consists of two or more basic symbols. For example, "That apple is sweet" is a complex symbol, which contains four basic symbols “That” "apple", "is" and "sweet". Thus, every sentence is a finite arrangement of constituents that are themselves either basic or complex (Fodor 2008). The idea of complex symbols is relatively straightforward, but we have to be aware that the order of the basic symbol within the complex symbol does make a difference. Consider the example "That apple is sweet" again, the order of the basic symbol plays an important role. If we rearrange the order of the basic symbols into, such as "sweet is apple that", it turns out to be a different complex symbol. Why is that a different order makes a different symbol? There are two aspects here that need to be further explained. i.e. the syntactic property and the semantic property of a symbol.

1: Syntactic Properties

In general. the syntactic property of a symbol is the rule of the system that the symbol resides within. For example, in natural language, syntax is the law of language that determines the word order and basic sentence structure. In other words, the syntax of a language, in fact, is its grammar. Each language has its own set of rules (grammar). Thus, the syntax of language A may not be necessarily identical to language B. (e.g. Consider the grammatical difference between French and English). When we communicate in natural language, for instance, during a conversation, the speaker and the listener follow the language rules to determine the order of words and the structure of sentences. The order of words and the structure of sentences, i.e. the order of symbols, at the same time, determine the interpretation of the symbols. Thus, the function of syntax involves determining the symbols' order in a system. In other words, the function of syntax is to disambiguate. Without syntax, sentences will be incomprehensible and incoherent. Of course, ordinarily, syntax determines what (in language) a sentence is. For example, we cannot randomly put five nouns "Cat Car Sun Shoe Key" together and call it a sentence.

Not surprisingly, mathematics and logic are syntactically sensitive. The syntax in those systems involves the rules of reasoning. Incorrect syntax in mathematics and logic results in errors or inconsistencies. Similarly, in computer programming, syntax is a set of rules and principles that defines the combination of symbols and governs the structure of the program. As a result, the role of syntax firstly answers the question of why different orders of a particular set of basic symbols result in different complex symbols: the syntax determines the form of a complex symbol.

2: Semantic Content

Let us take natural language as our example again. Languages have syntax (i.e. grammar); they also have semantics. Semantics refers to the meaning of words and sentences. Moreover, semantics largely determines how we understand each other in communication. For example, when we are holding a conversation verbally, we are exchanging sentences with each other. To have a successful conversation, both parties must adequately understand the meaning/semantic content of each sentence.

The further questions are, what is "meaning" and what does it mean when we say "we understand the meaning of a word (or a sentence)".

The meaning of a symbol involves what it refers to. The referent could be an object, a concept, a feeling or even a non-existent object. Consider the example "Apple" again, if we want to understand the meaning of the word "Apple", the best way we do is to link the word "Apple" with the object it refers to in the real world. We can perceive the object "Apple" through our sensory organs, i.e. eyes, ears, nose, tongue, and skin. And after some processes, we form the concept of "Apple" in our head. (On how we form concepts, this will be the objective of the next chapter, The Theory of Content.) Since then, we have the knowledge of the word "Apple".

The truth value of a symbol, i.e. whether it is true or false, is also a semantic property (Ravenscroft 2005). The basic symbol "Apple" is neither true nor false. But the complex symbol "Sugar is sweet" is true and it is false that "Sugar is salty". In general, when a symbol makes a claim about the world, then it has a truth value. That is why the complex symbol "Sugar is sweet" is true because it makes a claim about an object in the world, and we can examine its truth or falsity. That is also why the symbol "Apple" does not have a truth value, because it does not make any claim about the world.

A quick summary of the relation between syntax and semantics. The function of syntax is to determine the form of a sentence, and to disambiguate its semantic properties. Syntactical incorrectness may result in distorting the meaning of a sentence. However, syntax itself does not determine the semantic properties of a symbol. In general, the meaning (semantics) of a word depends on what it refers to, and the meaning of a sentence generally depends in a systematic way upon the meaning of its component words. Since words are symbols, hence language is a system that manipulates symbols (words).

After understanding what a symbol is and what properties it contains, our next question is how a computer manipulates symbols.

Section Two: Computation and The Turing machine

According to the Computational Theory of Mind, the similarity between a mind and a computer program is that a computer program manipulates data using formal and mechanical rules, while the mind manipulates its mental contents using internal systematic structure. For example, the mind is a program with the rule "If A then B", where A is a form of "Fire is hot", and B is a form of "be careful". If we perceive a signal (symbol, or more precisely here, a representation) "Fire is hot", then the mind will yield an output "be careful".

While traditionally, we say the mind is a computational system, we refer to that "system" as ultimately equivalent to a Turing Machine. A Turing machine is a theoretical model of a computing device that was proposed by mathematician Alan Turing in 1936. It consists of a tape that is divided into cells, each of which can contain a symbol from a finite alphabet, a read/write head that can read or write symbols on the tape, and a finite control unit that can move the head left or right, change the symbol on the tape, and change its state based on a set of rules. The basic idea of the Turing machine is to reduce complex procedures to sequences of elementary steps that can be executed mechanically. Basically, we start a Turing machine by giving it some rules; the machine will follow those rules sequentially and compute the function. The Turing machine can be as simple as a scanner head that can scan the input signal, a writing head that can function as writing and deleting, and a tape paper. The machine operates in a step-by-step manner, where at each step, it reads the symbol currently under the head, consults its table of rules to determine the next action to take based on the current state and symbol, and then updates the state and symbol on the tape accordingly.

Here is an example of how a Turing machine works, calculating "2+2":
The input to the machine will be like "_X X _ X X"

The left sequence of Xs represents the number 2, so does the right sequence of Xs. "_" represents blank.

The Turing machine table of adding two numbers will be like:

1. State 1: If X, then no change, move right, and then go to state 2; If _, then no change, move right, and then go to state 1.

2. State 2: If X, then no change, move right, and then go to state 2; if _, then replace it with X, then move left, and then go to state 3.

3. State 3: If X, then no change, move left and then go to state 3; if _, then no change, move right, and then go to state 4.

4. State 4: Replace X with _, then halt.

By following these rules sequentially, the Turing machine would iterate over the input symbols, performing the necessary calculations and transformations to compute the sum. The machine's ability to manipulate symbols, move along the tape, and change its internal state allows it to emulate a step-by-step computation process.

The procedure will be like:
(Yellow is the beginning of the step, Pink is the end of the step)
Step 0: _XX_XX (the machine is at State 1, read _, no change, move right, go to state 1)
Step 1: _XX_XX (read X, move right, go to state 2)
Step 2: _XX_XX (read X, move right, go to state 2)
Step 3: _XXXXX (read _, replace with X, move left, go to state 3)
Step 4: _XXXXX (read X, no change, move left, go to state 3)
Step 5: _XXXXX (read X, no change, move left, go to state 3)
Step 6: _XXXXX (read _, move right, go to state 4)
Step 7: __XXXX (read X, replace it with _, halt)

Now the sequence of Xs (i.e. __XXXX) represents the number 4.

The Turing machine is an important concept in computational theory because it provides a mathematical model of computation that is powerful enough to capture the essence of all possible algorithms. This means that any algorithm that can be performed by any computer or program can be simulated by a Turing machine. This is known as the Church-Turing thesis, which states that any computation that can be performed by an algorithmic process can be performed by a Turing machine. In theory, a Turing machine can deal with the complexity of algorithms,

Hence, if the computation theory of mind's claim is true that the mind is a computation system and cognitive activities are computation program, then theoretically speaking, the mind can be reduced to a simpler program. This entails that all the properties of mind, i.e. all mind states, can be reduced to a simple step-by-step mechanical processing system.

Are mental processes similar to the computational process that a Turing machine executes? Alan Turing once wrote, “If now some particular machine can be described as a brain we have only to programme our digital computer to imitate it and it will also be a brain. If it is accepted that real brains, as found in animals, and in particular in men, are a sort of machine it will follow that our digital computer suitably programmed, will behave like a brain." (1951, p.2) In other words, this is to say that any symbolic algorithm executed by a human brain can be replicated by a suitable Turing machine. He concludes that the Turing machine formalism, despite its extreme simplicity, is powerful enough to capture all humanly executable mechanical procedures over symbolic configurations. There are several reasons why mental processes can be seen as similar to the computational processes that a Turing machine executes:

First is the use of symbolic representation. Both mental processes and Turing machines rely on symbolic representation to manipulate information. In the case of the mind, information is represented as mental images, concepts, and other mental representations, while in a Turing machine, information is represented as symbols on a tape. And symbols are binary digits (0s and 1s) that are manipulated according to a set of rules.

Second is that both mental processes and Turing machines involve the processing of information through a series of computational steps. Mental processes involve a series of cognitive operations such as perception, attention, memory, and reasoning, while Turing machines involve the execution of algorithms through a series of logical operations.

This entails that all the properties of mind, i.e. all mind states, can be reduced to a simple step-by-step mechanical processing system.

According to this view, thoughts are complex symbols made up of basic symbols. (Ravenscroft, 2005). Thinking is a symbol-manipulation process. Cognitive perception is characterized in terms of symbols and representations because it allows us to understand how the mind processes information and creates meaningful experiences of the world. In order to make sense of the complex and varied sensory input that we receive, our brains need to be able to create representations of that information that can be manipulated and analyzed. Symbols and representations are a way of encoding sensory information in a way that can be easily processed and used to guide behavior. For example, when we see a red apple, our brains create a representation of the apple that includes information about its color, shape, and texture. This representation can be thought of as a kind of symbol that our brains use to categorize the apple as a distinct object in the world. This representation can then be used to guide our behavior, such as reaching out to pick up the apple or avoiding it if it appears rotten or spoiled.

Meanwhile, symbol-manipulation is thought to be a key component of higher-level cognitive processes, such as language and abstract reasoning. When we engage in complex cognitive tasks such as language comprehension or problem solving, we use symbols and representations to encode and manipulate information. These symbols allow us to break down complex concepts into smaller, more manageable parts that can be analyzed and understood.
In short, the use of symbols and representations is an important part of cognitive perception because it allows us to create meaningful experiences of the world and use that information to guide our behavior in a flexible and adaptive way.

Such a view has answered the second task of the computational theory: regarding what kind of information the mind processes or computes, the answer is symbols. The computational theory proposes to explain the nature of minds in terms of computation. Minds compute themselves in terms of principled manipulation of representations (Edelman 2008, P.28). As we have emphasized for a computational system, what it deals with are symbols, and those symbols represent the target objects.

Note also that thoughts must have structure. That is to say, complex symbols are organized in a certain way. As we have discussed previously, complex symbols (i.e. sentences) are built up of simple symbols (i.e. words), but the order of the complex symbols is not random. For example, the complex symbol "the sky is blue" is made up of four simple symbols, and they are organized in that particular order. It is the structure of the complex symbol “the sky is blue” makes itself distinct from the complex symbol "blue the is sky" (or other orders with those four simple symbols).

Overall, the computational theory models the mind as a computational system. Each mental state is a result of the computation.

Here, we have the third task that needs to be explained: what is the nature of the brain's systematic structure? Fodor proposes the explanation in term of internal language.

Section Three: The Language of Thought Hypothesis

If the Computational theory of mind is correct, then the language of thought hypothesis is a more specific version of the Computational theory. The language of thought hypothesis states that there is an innate, pre-set mechanical system that governs our thinking processes. Moreover, since that innate system has rules (syntax) and mental representations have contents (semantic properties), it can be considered as a "mental language", or to use a more famous term “mentalese”.

For Fodor, mental representations are more like linguistic symbols than photographic representations. For example, if P is a photographic representation of X, then a part of P will be physically similar to a part of X. This is called the Picture Principle (Fodor 2008, P.173). However, that may not be the case with linguistic symbols. As mentioned in the section on Symbols, not all symbols can be decomposed. Basic symbols are non-decomposable. Thus, a part of a basic symbol is not part of what it represents.

What is the relation between manipulating mental representations and using languages? In Fodor's account, mental representations are a kind of non-natural language that is known as the language of thought. Similar to natural language, mental representations are arranged in the mind in a systematic way, while the "systematic way" is a set of built-in rules rooted in mind. It will be helpful to compare mentalese with a natural language, let's say English. As we have seen in the previous sections, language has syntax and semantic contents. There are finitely many words (basic symbols) in English, but there are infinitely many sentences (complex symbols). Each sentence is a combination of words in a particular way. And finitely many grammatical (syntactic) rules are employed in combining words into a legitimate sentence. For example, the grammatical rules entail that "I am going to Japan tomorrow" is a legitimate sentence, while "Japan going am tomorrow I to" is not. How grammatical rules work in English is not in the scope of this thesis, nevertheless, it provides an indication of how those words are put together by means of application of those rules.

I have mentioned that there are infinitely many sentences that can be created in English. How does it work? It is because the grammatical rules of English are recursive, which means they can be applied repeatedly. For example, in the sentence "I am going to Japan tomorrow", we can easily apply a certain linguistic form to it, let's say, "It is true that I am going to Japan tomorrow". Certainly, I can re-apply that form again, so we have "It is true that it is true that I am going to Japan tomorrow", and so on so forth.

1: Mentalese

The language of thought hypothesis implies that cognitive processes involve the manipulation and combination of these basic elements to form more complex thoughts. This framework provides a way to analyze and understand cognitive phenomena by treating them as structured linguistic expressions.

In Fodor's account, mentalese is conceptualized as being similar to English. Fodor's language of thought hypothesis proposes that the mental processes underlying cognition can be understood as a language with a compositional structure. This view suggests that thoughts and mental representations can be broken down into basic elements, similar to the way words and sentences are composed in a language. The language of thought hypothesis describes the nature of the mental process as a language-like compositional structure. According to this view, simple concepts can be combined in a systematic way in order to form thoughts. In the simplest description, the mentalese resembles natural language in different respects. Similar to natural language, mentalese has syntax, which is the rules of the internal system. Natural language is compositional, such that complex linguistic expressions (I.e. sentences) are built on simpler linguistic expressions (i.e. words, simpler sentences). Mentalese is compositional, and concepts are the constituents of thoughts. (Fodor, 2008, P.20). Natural language and mentalese are similar in a structural way.

There are syntactic rules of mentalese for combining basic symbols into complex symbols. As such, mentalese is capable of expressing any meaning that English (or any natural language) can express. Thus, for every sentence that a natural language can express, there will also be a sentence that is expressed in mentalese.

How are sentences of mentalese encoded in the brain? Since mental states are computational states, and computational states are functional states of the brain, each mental state is a physical state of the brain. Thus, every mentalese sentence is a state of the brain. In general, a mentalese sentence corresponds to a brain state[16]. Two similar syntactic structure mentalese sentences, not surprisingly, will have two similar internal physical structures.

One may wonder why mental processes do not occur in natural language, but have to occur in mentalese. Here Fodor (1975) replies,

> "Computational models presuppose representational systems. But the representation systems of preverbal and infrahuman organisms surely cannot be natural languages. So either we abandon such preverbal and infrahuman psychology as we have so far pieced together, or we admit that some thinking, at least, isn't done in English" (P.56)

Fodor argues that evidence has shown that non-verbal organisms do have thoughts, suggesting that mental processes do not necessarily occur in natural language. Precisely speaking, although computations presuppose a representational system, it does not presuppose natural language as the vehicle of the representations. There could be a representational system other than natural language.

For example, do we think that human infants, before they can speak our language, cannot have thoughts? It is evident that they can have proper thoughts.[17] [18] For instance, empirical studies have found that newborns can recognize familiar voices, they can tell the difference between their

[16] The term "mental state" refers to the subjective experience or content of one's thoughts, feelings, or perceptions. On the other hand, "brain state" refers to the underlying physiological or neural activity in the brain.
It proposes that every mentalese sentence corresponds to a state of the brain. This implies that there is a one-to-one relationship between specific patterns of neural activity in the brain and the content expressed by a mentalese sentence.

The idea behind this perspective is that the mind's cognitive processes, including thinking and mental representations, can be understood in terms of the physical processes occurring in the brain. It suggests that there is a causal relationship between the neural activity in the brain and the mental states or experiences that we perceive.

[17] Birgit Mampe, Angela D. Friederici, Anne Christophe, Kathleen Wermke, “Newborns' Cry Melody Is Shaped by Their Native Language” , Current Biology, Volume 19, Issue 23, 2009, Pages 1994-1997.

[18] DeCasper AJ and Fifer WP. 1980. Of human bonding: newborns prefer their mothers' voices. Science. 208(4448):1174-6.

mother's native language and a foreign tongue (preferring their mother's language). This discriminatory ability indicates a basic kind of thought. Studies also show that newborns would prefer their mother's voice over other unfamiliar voices. Thus, even without language, one can still have thoughts.

For Fodor, mentalese is innate. He claims,

> "not all the languages one knows are languages one has learned, and that at least one of the languages which one knows without learning is as powerful as any language that one can ever learn." (1975, P.82)

Fodor argues that one can learn a language L only if one already knows some language rich enough to express L's extension of any predicate. Consider, for example, the French word "Pomme". If we do not know what it means in any language (including its decomposed simpler concepts), we cannot know it means "Apple". (Consider if we even do not know concepts like Red, Square, Sweet, etc., how do we understand the French word "Pomme" in a descriptional way?). If learning a language L requires a further language L', then under the same logic, learning the language L' would require a further language L". And this recursive process will keep on going. The plausibility would be that there is a language that is not learnt, but innate to us, as Fodor argues.

After explaining how we have an innate language (mentalese), Fodor further comments on the process of concept learning. Fodor explains what concept learning is in terms of the psychology of perception in his early career. He summarizes as below,

> "Concept learning presupposes a format for representing the experiential data, a source of hypotheses for predicting future data, and metric which determines the level of confirmation that a given body of data bestows upon a given hypothesis." (Fodor 1975, P.42)

And in his later work, he further explains,

> “Roughly, it's that concept learning is a process of inductive inference; in particular, that it's a process of projecting and confirming hypotheses about what the things that the concept applies to have in common.”(Fodor 2008, P. 132) (he calls this the “HF” model of concept learning.)

Suppose we extend the hypothesis formation and testing model. In short, when we learn a new concept, we first set a hypothesis of what it means, and then test it to see if it matches our hypothesis.

However, Fodor argues that there is no such thing as concept learning. He takes the concept GREEN as an example. If a creature can think of something as green, then that creature must have the concept GREEN already. According to Hypothesis Information, the process that one learns GREEN must include the inductive evaluation of some such hypothesis as such " The GREEN thing is the one that is green". But, the inductive process of that hypothesis itself requires bringing the property green. i.e. you cannot represent something as green unless you have the concept GREEN. Fodor further concludes,

> "Conclusion: If concept learning is as HF understands it, there can be no such thing. This conclusion is entirely general; it doesn't matter whether the target concept is primitive (like green) or complex (like green or triangular). And (as I may have mentioned) if we're given the assumption that concept learning is some sort of cognitive process, HF is de facto the only candidate account of what process it might be. But HF is circular when applied to the learning of concepts. So there can't be any such thing as concept learning." (2008, P.139)

Thus, Fodor criticizes that HF of concept learning is circular, i.e. HF in fact does not work on concept learning. Hence, if HF is the only way to learn concepts, then there is no such thing as concept learning. On the other hand, if concepts are not learned, what would they be? In Fodor's account, although concepts cannot be learnt, concepts can be acquired. We acquire concepts through experiencing a prototype (of a certain thing, event, incidence, etc.), and learning that prototype goes through an inductive process. i.e. The process of concepts acquisition involves inductive generalizations. Fodor, in other words, considers that learning a prototype, or acquiring a concept, is a brute physical process. His picture thus looks like this (2008, p. 151):

Initial state → (P1) → stereotype/prototype formation → (P2) → locking (= concept attainment).

P1 is a state of inductive generalization process. Under this process, the stereotype/prototype is formed. In supporting this prototype-forming claim, Fodor appeals to the scientific studies on infant's behaviours, he writes,

> "[........] well-known empirical results suggesting that even very young infants are able to recognize and respond to statistical regularities in their environments. A genetically endowed capacity for statistical induction would make sense if stereotype formation is something that minds are frequently employed to do" (2008, P.153).

The state P2 is where Fodor rejects the idea of concept learning (via HF). He denies that P2 is a inferential process, "......there are only two kinds of inferential process that it could be; and pretty clearly, it isn't either of them."(P.154). Those two kinds are: (1) pairing stereotypes and concepts, which is circular. Our question is about concept acquisition, and we cannot pair a stereotype with a concept you do not have. (2) there is a pattern of inference that, given a stereotype, identifies the corresponding concept. Fodor argues that to identify the relation between a stereotype and a concept is implausible. The problem is that there are so many different kinds of relations in which the concept and the stereotype may stand (P.154). The dog stereotype is related to the concept DOG in one way, the triangle is related to the concept TRIANGLE in another way. There is no way we can tell how to identify the relation between a stereotype and a concept. As he writes, "I don't believe that one can, as it were, just look at a stereotype and figure out what concept it is the stereotype of." (P.155).

Fodor does not describe what the state P2 is. Instead, he writes that "Psychology gets you from the initial state to P2; then neurology takes over and gets you the rest of the way to concept attainment" (P. 152), and "P2 must, of course, correspond to some neurological process or other"(P.155). Nevertheless, he describes the function of P2, "I'm assuming that, in the paradigmatic cases, P2 starts with an (inductively derived) stereotype and eventuates in lock to some property of which the stereotype is a particularly good instance.” (P.155).

Therefore, in Fodor's account, concept learning is not plausible. Yet as we have seen in his account, concepts are acquired as a consequence of experiences with their instances (P.164).

Next, Fodor claims that our brain's internal structure formalizes and manipulates symbols (i.e. mental representations). However, how does this apply to the computational theory? We have discussed that a computer manipulates symbols by means of the application of symbol-manipulating rules, which is to say no more than that there are syntactic generalizations governing its input-output. Consider, that we can build a computer that takes input sentences of a natural language and produces another sentence in the same language as an output. Similarly, according to Fodor's mentalese theory, whenever a mental process occurs, a computational mechanism built-in to the brain takes a mentalese sentence as input, and the machine generates a mentalese sentence as output.

Section Four: Objections to the Computational Theory of Mind

1: Metaphorical reasoning

There are challenges to the computational theory. According to the computational theory, thoughts are realised by a symbol-manipulating process, and that system can be reduced into a Turing machine, i.e. thoughts can be formalised. As Haugeland (1985) states, "an intelligent system must contain some computational subsystem". However, could there be an intelligent system that is not formal? We may agree that some of our intelligence is computational, e.g., when we perform the calculation "1+1", we can manipulate the calculation through mathematical rules to complete the task. However, someone may argue that we also perform some tasks in an informal way. For example, analogical or metaphorical thinking may not be reducible to a formal process. i.e. a step-by-step process may not be able to generate the proper outcome.

It is sometimes thought that metaphors require a special kind of “metaphorical meaning”. However, Davidson (1978) argues that metaphors mean nothing more than the words' literal interpretation.

He claims, “The central mistake against which I shall be inveighing is the idea that a metaphor has, in addition to its literal sense or meaning, another sense or meaning.”. (P.32) Davidson's claim depends on the distinction between what words mean and what they are used to do. He writes “It is something brought off by the imaginative employment of words and sentences and depends entirely on the ordinary meanings of those words and hence on the ordinary meanings of the sentences they comprised. “ (P.33) Thus, to understand metaphor, we need to describe the purpose for which metaphors are put in conversational contexts.

As a first step, it seems natural to say that metaphor is a form of likening or comparing. A metaphor typically states that one thing is another thing, and it equates two things not because they are the same but for the sake of comparison. For example, saying “life is a drama”, taken it literally, is false. However, if we consider it metaphorically, it is meaningful and probably true: life is not identical to drama but instead we mean life is exciting, emotional, or unexpected, which are properties that dramas also possess. Similarly, saying “that exam is a killer” does not really mean that the exam kills somebody, instead we mean that the exam is so painful. So on the one hand, we can see that there is hardly a logical inference between the two relata of the metaphor; but on the other hand, we intuitively (through our experience) understand that they have something in common. As we have experienced, life and drama are common in both having unexpected events; the exam may be exhausting, which we can compare to physical decline and destruction. Hence, a metaphor involves using the features of the new source (e.g. drama), to apply it to the old subject (e.g. life).

A good metaphor requires a certain degree of novelty and surprisingness, so that it can draw the listener's attention. Thus it is common for the subject and the metaphor to be from different domains in order to produce a more striking comparison. On the other hand, the metaphor itself cannot be overly obscure. It has to be understandable by the receiver. For example, saying that “Euclid is the father of geometry”, could mean that he is of importance to the practice of geometry; but if we say that “Euclid is the microwave of geometry”, the listener will have no idea what this metaphor means. Of course, we need to draw a line of how accurate is accurate enough to determine how the metaphor works.

A typical metaphor takes a form like "A is B". We literally say that the subject is something else but in fact we are not making a logical identity statement. When obviously A is not B, but surprisingly someone still says "A is B", this kind of surprise-factor is important as it can draw readers' attention and lead them to wonder why we are saying that. To increase the power of the surprise-factor, the subject and the metaphor had better be two things that are very different (better from different domains). For example, saying “that sprinter is a human”, it is not really surprising. We will not consider it a good metaphor as they have too many similar features. At best, it is an uninteresting metaphor. But saying “that sprinter is a cheetah” (i.e. He runs fast), it has some surprise-effect such that the listener should think “what does it mean?”; if saying “that sprinter is a rocket” (i.e. He runs very fast), it even has a greater surprise-effect that makes us think “what does it mean?!”. The more unusual the metaphor is, the better surprise-effect it has, such that it will be easier to draw listeners' attention, leading them to wonder why "A is B".

Once the metaphor successfully draws listeners' attention, it leads the listeners enter into a process of trying to understand the meaning of the metaphor. On the one hand, we use the difference between the subject and the metaphor to draw the listeners' attention. On the other hand, we use the surprise-effect from the metaphor to invite listeners to think about the similarities between the subject and the metaphor, and by extension, the nature of the subject and its relation to the metaphor. However, if we want to compare two terms conceptually and physically, we need to have solid background knowledge, and the relevant history and culture of those two terms. We also need the ability to connect the new to the old, the novel to the familiar. For instance, we say, "Time is Money". It is literally false because time and money are not the same thing. The first person who created this metaphor surprised his listeners and led them to wonder, "what does he mean?". Apparently, we do not take that "meaning" as the real meaning of that sentence. Instead, we invite people to think about the relation between time and money. We try to understand the metaphor by *comparing* the conceptual and physical similarities between time and money. For example, it can be that both time and money can be "spent"; time is important so is money in our modern daily life; to make more money, we need to spend time; to waste time is the same as to waste money (less time to make money), etc. All those similarities suggest that time and money have a special relation. Hence, even though time is not money, now it is clearer how time "is" money-- we in fact refer to their common abstract properties.

Nevertheless, detecting which features are relevant requires context information, which could be a complex task. How to draw the comparison depends on our interests, which is reliant on our understanding of the subject and the metaphor. Consider our life and drama example again. If we do not have any idea of what a drama is, this metaphorical expression would not work.

So far, we generally have seen what metaphor is and how we identify, understand and accept it. Then how does a computer recognize a metaphor? The task is potentially complex. I am not able to offer any detailed account here. However, I think we can give a broad sketch of the likely computational process involved enough to make us reasonably confident that metaphor will not be a fatal problem for the computational theory of mind. In particular, there are important processes involved in extracting the most relevant or significant similarities between the relata. First, there has to be a recognition that something novel is being done relative to ordinary comparisons. The computer must be able to link the new to old, the novel to the familiar; it must consider its own past and the experiences collected. The general computational idea is that each concept is a domain or set that contains elements. We can program the computer to search for resemblances between domains. However, first of all, in a metaphorical sentence such as "A is B", the computer needs to identify whether or not this is a new matching in its system. If so, this is a "surprise" to the computer, which further draws its attention to investigate whether it is a metaphor or not.

Meanwhile, the computer needs to check the context where the sentence "A is B" fits in and sees whether the use of B is consistent with the content or not. If so, "A is B" is highly possible a metaphor whose meaning should be consistent with the context. Finally, suppose the number of matching elements between domain A and domain B (assume all elements of B are in the computer's system) falls within a certain range. In that case, the computer can identify those matchings that are the most relevant comparisons and consider "A is B" as a metaphor. Then the computer associates B to subject A and forms an output like "Life is a drama because such and such elements are in common". The computer can then point out that the symbol "life" contains such and such elements; the symbol "drama" contains such and such elements. "life" and "drama" both contain elements X1, X2, X3 and so on.

Moreover, we may potentially require the common elements between the relata of the metaphor to reach a pre-set number/line. The computer cannot endlessly search for more and more obscure resemblances between the two relata. Instead, the computer has to settle on those resemblances that go along with the ongoing conversational topic, and constantly update the system database to see if there is any new use of B as per conversations, the meaning of words may vary (slightly) throughout the conversation.

2: Semantic contents and the Chinese Room Argument

As we have discussed in the section on what symbols are, symbols have semantic properties, or contents. What the contents are would be based on how we assign to them. However, contents could be arbitrary. For example, a red traffic light has been assigned the content "stop at the traffic light"; meanwhile, it is perfectly fine (theoretically) to assign it a different meaning, for instance, "Keep going".

Computers operate symbols as the Turing machine does. A computer is a formal information process that runs through the algorithmic manipulation over syntactically structured symbols as input and produces syntactically structured symbols as output by following a set of manipulating rules. In the practical course of computation, a computer does a series of activities, such as reading input symbols, storing symbols in its memory, retrieving symbols, writing symbols, deleting symbols, etc. Computers are mechanical devices. They only operate under what rules we assign to them. Computers do not need to understand the content of a symbol (by understanding, I mean the computer can specify what the symbol represents.). They are only sensitive to the pre-set program's syntax. Thus, the output symbol is entirely determined by the syntactic rules of the program. The contents of the symbol should be consistent throughout the manipulation.

For example, consider a program with a syntactical rule "If A then B". Whenever A is the input, the program will produce B as an output. For instance, if we assign the form as "If it is raining then it is wet" and give the input as "It is raining", then we have the output "It is wet". As long as the input symbol (here is a sentence) matches the form we assign to the program, it will always yield the same output. Again, the computer itself does not need to understand what contents the symbol carries. The term "manipulation" is no more than saying that the computer's operational process is syntactically governed.

One of the major challenges to the computational theory comes from John Searle's Chinese Room argument. Imagine Searle is in a room with an instruction book on how to respond to Chinese symbols, but he does not know anything about Chinese at all. By following the instruction book, Searle can respond to any Chinese message as fluently as a native Chinese speaker. However, Searle argues that even though he can hold a conversation with a native Chinese speaker, he does not understand any Chinese words at all. Thus he does not understand the meaning of the conversation. Searle further argues that his role in the room is no different from that of a computer. Hence, even if a computer can communicate indistinguishably to a human, the computer itself, in fact, does not understand the meaning of the conversation.

In short, Searle's argument runs as follows:

1. Computation is formal (syntactical).
2. Minds have contents (semantic contents)
3. Syntax is not identical with nor sufficient by itself for semantics. Therefore,
4. Computation is not sufficient for nor identical to minds.

In Searle's account, on the one hand, conversations involve syntax (grammar) and semantics (contents); on the other hand, machines are formal systems that can only manipulate symbols according to their programmed rules. Machines have nothing to do with semantics. Furthermore, Searle argues that intelligence requires an understanding of semantic contents, but the Chinese Room argument has shown that a formal computational system does not understand any semantic contents. As a result, computers just merely simulate humans' behaviours; they do not understand anything at all; i.e. mind is not computation.

There are different replies to the Chinese Room Argument. One of the most common replies (and the most related to this chapter) is the so-called The System Reply (the response following Searle 1980, also Block 1995). The System Reply admits that it is true that the man in the room does not understand any Chinese at all, but the man is only a part of a larger system, i.e. the whole room. We should consider the Chinese Room as a whole, including the database, the memory, intermediate states, and the instructions, maybe even the input window, the output facility, etc. That is the complete system that is required for answering chinese questions. So the System Reply is that while the man running the program does not understand Chinese, the system as a whole does.

Searle's response to the Systems Reply is technically simple: let us consider the man in the Chinese Room is a genius. He could internalize all parts of the system on his own. i.e. he can remember the whole instruction book. He can perform all the instructions or complete any calculation in his head. Push it even further. The man does not even need to sit in the room. He can just walk outside and hold a conversation in Chinese with a native Chinese speaker. But what he is replying to is still a straightforward input-instruction-transition-output process. He does not know the meaning of a Chinese word at all. The point here is that that man is the entire system, yet he still does not understand Chinese. The conclusion is, he still cannot get semantic contents from syntax.

The critical point of the Chinese Room argument is that it pinpoints the problem of any formal machine would encounter: they are syntax engines. They follow the pre-set rules and only deal with syntax. Undoubtedly, as we have discussed before, syntax does play a role in clarifying meaning (That is, it can, once we already have an idea of meaning, help to refine or structure it). Still, within a formal/ mechanical system, the system itself has nothing to do with the fundamental semantic properties of a symbol.

Conclusion

The Chinese Room argument has become a prominent and widely discussed thought experiment in the field. It challenges the computationalist approach and raises fundamental questions about understanding, semantics, and consciousness. By presenting the argument, I aimed to address the specific comment and provide a concise explanation of the argument's key ideas.

The Chinese Room argument challenges the computationalist's view that the mind as an information-processing system and suggests that consciousness and mental content can arise from the manipulation of symbols through computation. Searle argues that the manipulation of symbols alone is insufficient to produce true understanding, as it lacks the essential aspect of semantic interpretation.

Critics of the Chinese Room argument propose counterarguments, such as the systems reply, which suggests that while the individual inside the room may not understand Chinese, the room as a whole, including the instructions and the person, could be said to understand Chinese. These counterarguments aim to demonstrate that even though individual components of a system may lack understanding, the system as a whole can still exhibit intelligent behavior.

Ultimately, the Chinese Room argument raises important philosophical questions about the nature of understanding, consciousness, and the limitations of purely computational approaches to explaining mental phenomena. It invites further exploration into the relationship between syntax and semantics, and how meaning is derived from symbol manipulation. This will be the next chapter's task: The theory of content and how a machine would cope with it.

CHAPTER FOUR

Theories of Content

In the previous chapter, we argued that symbols, or representations, have semantic contents. Searle's Chinese Room argument argues that computations are processes of manipulating symbols that are governed by formal rules. Thus, computations have only syntax, but not semantics. This chapter aims to explore the relation between representations (i.e. symbols) and their content. [19]In doing so, I will outline three different theories of content: the causal theory, teleosemantic theory and success semantics. I will conclude that success semantics best fits our needs for a theory of content.

Section One: Causal Theory of Mental Content

1: Content

Mental representations have contents, such as thoughts, desires, beliefs, intentions, emotional feelings, or wishes. The content of a thought is what it is about, the content of a desire is what it is desiring about, the content of a belief is what it is believing in, and so on.

"Content", may be about the world, i.e. my belief that the Sun rises from the east. The content of my belief is about an astronomical object in the universe, and how it appears to the observers on Earth. Whether my belief is true or not can be verified by observations: If the Sun rises from the east, then my belief is true; if the Sun does not rise from the east, then my belief is false.

"Content", may also not be related to the actual world. Consider, for example, the two-dimensional world in the book *Flatland: A Romance of Many Dimensions* (Edwin Abbott 1984).

[19] Experiences have contents. Consciousness and semantic contents are closely related because both involve subjective experiences that are associated with perception and meaning. That's why discussion on contents is so important.

The story describes a two-dimensional world *Flatland* occupied by geometric figures. Men in the Flatland are polygons, women are line segments, the narrator is a square, etc. All characters in the story are purely fictional. None of them exists in the actual world. i.e. the content or essence of each fictional character can only be understood or derived from the book in which they appear. In other words, the book is the primary source of information about the character, and any interpretation or understanding of their traits, motivations, or development should be based solely on the information provided within the text.

Mental states are often understood as having representational contents—that is, they carry information or meaning about something. In the case of desires, their content typically relates to a desired object, goal, or outcome. The content of a desire can be specific or general, concrete or abstract. For instance, a desire could be directed towards obtaining a particular item, such as a new smartphone, or achieving a broader goal, like career advancement. The content of a desire can also encompass emotional or experiential elements, such as desiring to feel loved or to experience joy.

Furthermore, desires can vary in their strength and intensity, influencing the urgency and priority we assign to them. They can also interact with other mental states, such as beliefs, in shaping our overall cognitive landscape and decision-making processes.

The mental content of desires is closely linked to our subjective experience and our intentional states. It is through desires that we strive for certain outcomes or states of affairs that we consider valuable or beneficial.

Here we know that mental contents can be about the world, or the fictional state of affairs, or our subjective experiences. We can further ask, how is a mental content formed? And how does it refer to other objects in the world, or fictional objects?

2: Causal Theory

A theory of mental content is a theory that aims to explain the relation between mental representations and their contents. For instance, having a thought that “the sky is blue”, on a representational theory of thought, involves explaining how the thought “the sky is blue” is formed and why the thought has that content (the sky is blue) instead of different content like “the sky is red”, or “the sky is green”, or even no content at all.

According to the causal theory of mental content, generally, “my thought about Xs is about Xs because (i) my thought was caused by an X and (ii) nothing but Xs cause me to have thoughts about Xs” (Ravenscroft 2005, P.127). For instance, my mental representation “the sky is blue” is about the blue sky because the blue sky causes my mental representation of “the sky is blue”, and only the blue sky causes me to have that exact thought “the sky is blue,”

Straightforwardly, a problem for the causal of content is that one can think “the sky is blue” without looking at the sky. Perhaps a Martian who is living on Mars has only ever experienced a red sky. When someone tells him, “The sky is blue”, even though he has never seen a blue sky, he can still have the thought “the sky is blue”. It seems to be true that we often have a certain thought of something without any direct causal contact with that particular object. For instance, I have never seen a platypus, but I have heard of them and often think I would like to see one. Thus, it seems to be the case that it is not necessary to have a direct causal relation to a particular object in order to have a representation of that object. One plausible reply is to admit that indirect causal links between the thought and the object would be sufficient enough for making the thought about the object (P.128). For example, I read a book about platypus, and presumably the author of the book has a more direct causal relation to a platypus. Thus I have a thought that refers to platypus because the book creates an indirect causal link between my thought of platypus and platypus.

3: Problems for the Causal Theory

3.1: Fictional Objects

One of the problems of the causal theory is about fictional or non-existent objects. According to the causal theory of mental content, if I have thoughts about fictional characters, there must be a causal connection between the thoughts I have and the fictional characters. However, fictional objects, such as the characters in the *Flatland* do not exist in the actual world; so it is hard to see how there could be a causal connection between them and our thoughts about them.

Although fictional characters may not exist in the physical world, it does not follow that they cannot be objects for thought. Representations of *Flatland* characters are products of our own imagination. Can the causal theory perhaps say that our thoughts are caused by non-actual objects? Again, a good theory of mental content must be able to explain what is going on here.

I could respond to this problem by saying that those two-dimensional characters are causally connected to our thoughts via descriptions. My thoughts about those two-dimensional characters are about the descriptions of those characters, instead of ordinary concrete objects. Fictional objects can be considered as a kind of social object. Social objects are network-like objects interacting between people. They gain meaning through reification. We have attitudes about the same propositions as one another, and to communicate about them. Social norms or practices are shared between people. Hence, we can causally connect our thoughts to those social practices. I do not have the space to examine the exact metaphysics of social objects here. However, it does seem feasible to say that we can enter into causal relationships with them, and so the causal theory is not fatally undermined.

3.2: Misrepresentation (1)

It seems possible for the causal theory of content to deal with fictional objects, but it has a more challenging problem to deal with: misrepresentations. Consider the following case of misrepresentation: I see something in the dark night, so that I have a mental representation "DOG". However, that thing, in fact, is not a dog, but a cat. Hence the representation "DOG" incorrectly represents the object that I see. Nonetheless, even though that representation is incorrect, it still has a meaning to me. For instance, it could refer to a furry creature with four legs and one tail creature walking in the darkness. According to the causal theory of mental representation, a representational mental state refers to what causes it. If that is the case, the object should be only causing a corresponding representation, and misrepresentation should not occur. However, in reality, that is not the case. Misrepresentations happen often. Here then is the question: if it is true that there is a causal link between the representation and the object, then why does the object cause a wrong representation? That is to say, if a thought about X can be caused by Y and not X, then the causal theory of mental content does not hold.

Note that if we accept that indirect causal links can cause representations of objects, then the problem of misrepresentation is perhaps even more salient. For example, I may have heard of platypus from someone who saw a platypus in person. My thought about platypuses is merely from a sole description. The person who told me about platypus may give me inaccurate descriptions about platypus (perhaps he said platypus have four claws, but in fact they have five), such that it causes me to have a misrepresentation of platypuses. [20]

One response to the problem of misrepresentation is as follows: if we try to explain the content of a mental representation as states of the world that causes its represented object, as in the dog-and-cat misrepresentation example, perhaps we can concede that what seems to be my representation of a dog in fact represents dogs OR cats in the dark night. Thus, we shift the content of a mental representation "Dogs" to the content of representation 'dogs or cats in the dark night'. Yet this "manipulation" of the mental content seems to eliminate misrepresentation entirely, i.e. for any potential causal link, we can always add a new disjunctive term to X, (hence, in our example, dogs or cats, dogs or goats, even dogs or cats or goats or anything). In this response, misrepresentation becomes impossible. This strategy seems to be too naive, as it does not really explain why misrepresentation happens, but merely eliminates the possibility of misrepresentation.

What is more, if a representation (e.g. "Dogs") could mean anything (i.e. "dogs or anything"), it seems to offer no explanatory power in the case of how we can properly represent dogs and only

[20] We can distinguish two types of misrepresentation: 1) X represents Y but "Y" is not Y (e.g. X represents blue, but that "blue" is not blue; 2) X represents Z but there is no Z (e.g. X represents something that is blue, that "something" does not exist. (Dretske 1995, P.27)

dogs. This is a major problem. If the causal theory cannot rule out the possibility of arbitrary disjunction, it must be rejected (Rupert 2008 P.356, Shea 2018,P.11).

We have just seen how misrepresentations cause problems to the causal theory of mental content. The theory has to explain how misrepresentations occur. Unfortunately, the causal theory (at least in its simple version) cannot provide a solution to the problem, i.e. there are some mental representations that the causal theory cannot explain.

Another major problem related to misrepresentation is that how to distinguish whether a representation is correct or not. The point here is that given that there is a causal connection, we have no basis for distinguishing which causal connections are accurate and which are inaccurate. For instance, I see something X and I have a representation Y. According to the causal theory, Y should represent X, but how do I know the representation Y correctly represents X or not? Unfortunately, the causal theory does not offer any help for this matter.

Section Two: Teleosemantic Theory

In the previous section, I mentioned the problem of misrepresentation and demanded that a good theory of mental content should be able to tackle that problem. Seemingly, the causal theory of mental content is unable to provide a satisfactory response. In this section, I will present the teleosemantic theory of mental content, which is a more sophisticated sub-type of causal theory which attempts to give a plausible response to the problem of misrepresentation. There are different versions of the teleosemantic theory of mental content. In this section, we primarily focus on Dretske's version.

According to teleosemantic theories, the contents of mental representations depend on functions, such as the functions of the systems that use or produce them.

In his book *Naturalizing the Mind*, Dretske writes,

> “ The fundamental idea is that a system, S, represents a property, F, if and only if S has the function of indicating (providing information about) the F of a certain domain of objects.“ (1995, p.2)

Note that not all events that carry information have the function of carrying it (1995, p.4). For example, smoke carries information about the presence of a heat source. But that surely is not the function of smoke. The smoke does not represent or misrepresent a heat source, nevertheless, it carries information that there is a heat source.

To help explain what a function is, Dretske offers the example of a speedometer. “A speedometer (S) represents the speed (F) of a car. Its job, its function, is to indicate, provide information (to the driver) about, how fast the car is moving (F).” (1995, p.2). The function of a speedometer is to provide information about the speed of the car (Let us assume the car is running properly on the road.). Different systematic states of the speedometer carry a different piece of information about the speed, e.g. when the needle on the speedometer is pointing at 50km/h, the speedometer carries the information that the car is running at the speed of 50km/h. The fact that the speedometer has that information-carrying function (i.e. indicating speed), is a representational fact of the instrument. The representational fact of the instrument is “.....what the instrument was designed to do, what it is supposed to do...” (1995, p.2). A system can fail to do what it is supposed to do. For example, if the speedometer indicates 50km/h (carries the information that the car is running at 50km/h), but in fact, the car is running at 60km/h, then the speedometer is not doing its job. i.e. the function fails, which results in misrepresentation.

Dretske further distinguishes between facts about a representational system and a representational fact of a system. For instance, facts about the components of a thermometer, such as how the mercury changes in volume indicates temperature, are facts about a representational device, but are not representational facts of the device. A representational fact of the system is a fact that it is designed to carry. (P.3). Thus, the representational fact of a thermometer is what it is designed to do---indicating temperature.

Dretske applies the distinction between representational facts of a system and facts about a representational system to clarify the relation between the mind and the brain. The difference between representational facts and facts about representations is the difference between the mind and the brain (P.3). Mind is what the brain is designed to do (we will come back to this point about design soon), or the function of the brain. Neuroscientists may know facts *about* the brain, e.g. the bio-structure of the brain, which may lead them knowing facts about mental representations. However, knowing facts about what the brain is, is not the same as knowing what the brain is designed to do, or what function it has. We extend the claim further, each functional state of a system is supposed to carry a piece of information about what it is designed to do (in the example of the speedometer, that piece of information is about the speed of the car). By the same token, each functional state of mind carries a piece of information about what it represents.

What kind of function is a mental representation supposed to have? In Dretske's view, the function of a mental representation is to carry information about the world (1995, p.6). For instance, the function of a sensory faculty, e.g. the visual system, is to provide information of the spatial arrangement of objects in one's environment. The sensory representation itself carries information that indicates something is happening at that part of the body; a mental image of a mountain is a representation that carries information about that mountain, and so on.

1: Natural and Conventional Representations

Recall that the function of a system refers to the job that the system is supposed to do. "Supposed", in Dretske's sense, refers to something that is designed for a specific purpose. In the speedometer example, the function of a speedometer is "supposed" to indicate the speed of a vehicle, which implies that the speedometer is designed to indicate the speed of a vehicle. Knowing what "supposed" means, the next question is what it means the term "design". Dretske distinguished two different kinds of designs. There is a natural one, and there is a conventional (artificial) one. I.e. there are naturally acquired functions, and there are conventionally assigned functions.

Ordinarily, when we use the term "design", we mean something that is deliberately (intentionally) constructed, built, or created, to achieve a specific goal. A simple example is that a watch is deliberately designed to indicate time. We may use a watch to do something else other than check the time, but primarily the purpose of a watch is to provide us with information about what time it is. Conventionally assigned functions are the functions that we design to achieve specific goals. We build them and we assign them jobs to do. When they function, they provide us with the information we need. For instance, we design a speedometer so that it gives us information about the speed of a vehicle. The functional state of the speedometer represents the speed of the vehicle. When a device's function is derived from the intentions of the designers, the representation is conventional.

Representations that are not conventional are natural. (1995, p.7). About naturally acquired functions, Dretske claims “I do not argue for this; I assume it.” (1995, p.7). He accepts that many organs have the function of doing certain jobs, without being designed by anyone to do it. That is, organs (or other natural things) have functions without a designer. Similarly, in the case of perceptual experiences, the function (of providing perceptual experiences) is systemic. i.e. a state is a conscious experience of certain objects if the state has the natural systemic function of providing information about that object. States acquire their systemic functions through the process of evolution, or natural selection. What one is experiencing here and now depends not just causally, but as a matter of logic on what happened yesterday or in the remote past. In other words, what one thinks and feels is likewise hostage to environmental and historical circumstances. (1995, p.126).

2: Misrepresentation (2)

With the teleosemantic theory of mental content in hand now, we are now able to tackle the problem of mental misrepresentation.

We generally admit that our beliefs structure misrepresent the world around us. Yet, according to the simple causal theory we first considered, mental representations always represent what caused them. Hence, there is no room for misrepresentation in this account. In contrast, Dretske's account of misrepresentation considers the information-carrying role of a system fails to perform its function. Consider national flags. There is a strict convention among flags such that each national flag has its unique graphic design. E.g. France's national flag is a combination of Blue, White and Red (and each colour has its meaning) with a specific pattern. We have a good criterion upon which we can identify whether a given flag accurately represents a specific nation (and that is the function of the flag). For instance, if the flag-maker puts the Dutch Flag instead of the France national flag to represent France, then the Dutch flag is giving a piece of wrong information about the country France, I.e. The Dutch flag is not fulfilling the role that it was designed for. Thus, we have a case of misrepresentation because the colour and the pattern of a national flag, according to the common convention, acquires a particular function (i.e. information-carrying role). When it fails to perform that function (here the case is to indicate France the nation), we will misrepresent specific facts about the country.

Dretske believes that information-indicating mechanisms, for example, neural structures in human beings, acquire information-indicating functions through natural selection and learning (1995, p.169 fn 3). If our sensory and cognitive mechanisms are in perfect order, such that they function in the way they were naturally designed for, we will have accurate representational contents. For instance, when our mind is well in order, and we can see clearly the cat in the dark, we can yield a correct representation Cat. On the other hand, if there are environmental distortions; or there is an imperfection in our internal mechanism, such that the mechanism is not correctly performing the function it was designed for, then there is a possibility of yielding misrepresentations.

3: Objections to the Teleosemantic theory of mental content

3.1: Swampman

One of the crucial elements in Dretske's teleosemantic theory of content is that biological functions are the products of natural selection. As we discussed previously, the contents of mental representations depend on functions, such as the functions of the systems that use or produce them. Mental contents are likewise hostage to environmental and historical circumstances. The Swampman argument aims to attack the point that mental contents depend on the historical background of the subject.

The Swampman argument is a thought experiment first proposed by Donald Davidson in his 1987 paper 'Knowing One's Own Mind'. The experiment goes like this: Suppose Davidson went for a walk in a swamp, got hit by a lightning bolt, and died. Co-incidentally, on the other side of the swamp, at the exact moment, a lightning bolt hit the swamp and rearranged some molecules into the same form as Davidson's body, copying every structure completely. That form of being is the Swampman. Not surprisingly, since the Swampman is a perfect copy of Davidson, there is no way we can physically tell the difference between Davidson and the Swampman. So we now ask: Does the Swampman have the same thoughts as Davidson? Does it even think at all?

Let us clarify some facts about the Swampman. First, as it just comes into existence, it does not have any causal history before the lightning bolt struck the swamp. Second, its existence is a pure coincidence. No designer or builder is responsible for the Swampman's existence. On this basis, Davidson argues that Swampman's mental content will not be identical to Davidson's mental content. In Davidson's account, a being that is capable of any cognisance requires a causal history of thought in the first place, which means a being who has no causal history of thought would not be able to have any cognisance. So even though the Swampman can possibly speak and behave as Davidson does, the words and behaviors are contentless to the Swampman.[21]

There are different responses to Davidson's conclusion to the Swampman thought experiment. One of the responses is an Internalist response. According to Internalists' view, the brain is the vehicle of the mind. And the physical structure and activity of the brain are sufficient to realise some mental states. Internalists consider that conscious experience and thoughts are causal

[21] Nonetheless, they could be meaningful to us.

conclusions of electrochemical processes that the brain produces and not their contextual or historical properties. Hence, since the Swampman's brain is molecules-to-molecules identical to Davidson's brain, then the Swampman is Davidson. This response seems to be intuitively reasonable. If the Swampman is physically identical to Davidson, and no one (not even Davidson himself and the Swampman) can distinguish them, then on what ground can we claim that the Swampman is not Davidson?

However, if Internalists are correct, Dretske's teleosemantic theory is questionable. Given that the Swampman appears spontaneously, it is not designed to do anything nor naturally selected. It does not have a causal history of thoughts at all. Under that view, the Swampman's mental content is not identical to Davidson's. In fact, it even lacks internal states at all.

Dretske has his version of thought experiment to respond to the Swampman argument. In short, Dretske basically agrees that Swampman has no content. His main claim is that we have unreliable intuitions about this case (1996:81).

Dretske constructs an analogous case to Swampman based on his car a Toyota Tercel. According to his teleosemantic theory, the function of the Tercel is what it was designed for, e.g. to be driven. One day, there is a bolt of lightning that strikes an automobile junkyard, just like the Swampman appears in the swamp. An object (he calls it Twin Tercel) appears in the junkyard, which is physically similar to the original Tercel (e.g. same engine block, same number plate, same chassis number, etc.). The Tercel and the Twin Tercel look identical. If the Twin Tercel is parked in his garage, he will not doubt that it is his Tercel at all. The only difference between the Tercel and the Twin Tercel is that the fuel gauge in the Twin Tercel is unresponsive to the amount of fuel in the tank. (so, at least, let us assume that when both objects are stationary, we cannot distinguish them).

Dretske argues that the Twin Tercel is not a replica of the Tercel, or even a car. The Twin Tercel came into existence by a purely random coincidence, so it has no history. It just happens to be physically similar to the Tercel. Apparently, from a functionalist's point of view, the Twin Tercel is also a car because it can realize a car's functions, i.e. it does a car's job. However, in Dretske's view, even though the Twin Tercel has all the features the Tercel has, and even though the Twin

Tercel can achieve most of what the Tercel can achieve, we have no reason to say that to be driven is the function of the Twin Tercel. It is because the Twin Tercel is not purposely designed, neither artificially designed nor naturally selected. If it is not purposely designed, it cannot be identified as having such-and-such functions. More precisely, we even cannot assume that the Twin Tercel is a car. We merely can say it is an object that looks like a car, and perhaps it can achieve what a car can achieve. Dretske's claim is clear, we can't just claim an object Z has a certain function simply because Z can achieve a certain task T. For example, the Twin Tercel can also place on top of papers to keep them from blowing away, but we do not consider the Twin Tercel is a paperweight. Thus, what an object can achieve does not 100% define what it is. The way we think what an object is somehow unreliably depends on how we intuitively assume what that thing is. E.g. we see the Twin Tercel looks like a car and can do what a car does, then we assume it is a car. As Dretske writes,

> "The Twin Tercel parable does not directly challenge The Internalist Intuition. It does not show that it is false or even implausible. It was not intended to do this. What it was designed to do, instead, is to reveal how unreliable ordinary intuitions are about miraculous materializations and instantaneous replacements" (1996:81).

Since we have no reason to assume that the Twin Tercel has the same function the Tercel has, thus, to claim that it has a such-and-such function is not appropriate. i.e. to claim that the fuel gauge in the Twin Tercel is not working is inappropriate (because it does not mean to do the such-and-such job). By the same token, to say that the Twin Tercel is working (i.e. can be driven) is not appropriate, too (because it does not mean to do the such-and-such job, too). The Twin Tercel is not a copy of the Tercel. We intuitively think that the Twin Tercel is a copy of the Tercel simply because we think it is physically identical to the Tercel and it can achieve what the Tercel can achieve. Thus we then assume the Twin Tercel has a function (i.e. to be driven). However, we have noticed that the Twin Tercel just exists as it does. Its existence is not purposely designed. It has no history or whatsoever at all. If it is not purposely designed, it is inappropriate to say it has a such-and-such function.

For the same reason, we can question what Internalists claim that the Swampman has content as Davidson has because they have the same internal structure. Since the Swampman came into existence coincidently, it neither has history nor it is purposely designed (artificially constructed or naturally selected). It is inappropriate to say that the Swampman has the same function as Davidson has. If it is inappropriate to say it has any function, then it is inappropriate to say it has content. [22]

The moment that the Swampman came into existence, it did not have any content. However, does it imply that the Swampman cannot have content (or any mental activity) after its existence begins? Intuitively, since the Swampman came into existence, its physical structure (which is identical to Davidson) should allow it to start performing some activities, such as walking around its environment, eating things, perhaps even talking to other people, just as well as Davidson does. However, if we are teleosemanticists, the Swampman should not be able to think like Davidson even after a long period of time. It is because, according to the teleosemantic view, the content of one's mental state is solely in virtue of one's history. Thus, since the Swampman lacks history, it therefore does not have Davidson's beliefs, even though he is physically identical to Davidson. More radically, it follows that the Swampman would not be able to have phenomenal experiences (i.e. qualia): although the Swampman's brain state could be physically identical to Davidson's, for example, the brain state when Davidson is tasting wine, the Swampman would not have a sensation of the taste of wine nor of the smell of it. In Dretske's account, one's sensory experience (i.e. sensory content) is determined by one's biological history. When one is in a certain state, in fact that state is biologically determined (more precisely, genetically determined) to be realised by the environmental factor that one is in. Thus, since the Swampman lacks the above representational capacity, it would have no content.

[22] Davidson (1987, P.443-444) also argues that the Swampman does not have thought nor meaningful language. The Swampman has no thought because it has no causal history to base on them.

“No one can tell the difference. But there is a difference. My replica can't recognize my friends; it can't *re*cognize anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to), it can't remember my house. It can't mean what I do by the word 'house', for example, since the sound 'house' it makes was not learned in a context that would give it the right meaning—or any meaning at all. Indeed, I don't see how my replica can be said to mean anything by the sounds it makes, nor to have any thoughts.” (P.444)

Section Three: Success Semantics

If Dretske's response to the Swampman thought experiment is counter-intuitive, success semantics might provide a more friendly answer. According to success semantics, "the content of a belief is fixed by the success conditions for the performance of an action triggered by this belief. "(Nanay 2013, P.2). In other words, the truth-condition of one's belief is fixed by the success conditions of one's action. J.T. Whyte (1990, p.150) claims,

> "A belief's truth-condition is that which guarantees the fulfilment of any desire by the action which, combined with that desire, it would cause."

There are three elements in the account of success semantics: desire, belief and action. When a desire arises, it plays a role in motivating what we are going to do, and the ultimate goal is to satisfy that desire.

For instance, the desire to not let my car run out of fuel will influence my driving behavior. To satisfy that desire, I will regularly check the fuel gauge to be aware whether the car is running out of fuel or not, or simply go to fill up my car at the nearby petrol station. I would direct my action (i.e. “to fill up the car”) because I believe that my car is running out of fuel. In a success-semanticist's view, the content of my belief that my car is out of fuel is determined by my successful action that I go to a petrol station and fill up my car with fuel.

The advantage of success semantics is that it offers a straightforward way to explain misrepresentation. According to success semantics, misrepresentations can be understood as what happens when a state (i.e. desire) tends to cause actions that would have been successful (i.e. desire fulfilling) if something is the case, but in fact that thing is not the case. Let us recall the dog-and-cat misrepresentation example again. When I see a cat in the dark, I form a false belief (misrepresentation) that it is a dog. The belief tends to combine with my desires (e.g., offering the “dog” some treats it likes, patting it, etc.). My desires cause my behaviors, such as looking for dog food or trying to approach the dog. Those behaviors would fulfil my desires that if it were a dog, but in reality, it is a cat. So we can see that, unlike teleosemantics, which attributes content based on one’s causal history (either biological or learning history, or both), success semantics attributes mental content based on one's immediate goals. If desires can arise spontaneously, then each mental content is an individual business.

Thus, according to the success semantics' account, the content of the Swampman's mental state is determined by the success of its action. For instance, if the Swampman is successfully walking across the swamp, we can conclude that it has a belief, which enables it to satisfy its desire to go to the other side of the swamp. The advantage of this account is that it does not need to consider one's history (just like what we have considered in responding to the Swampman). As long as one successfully performs an action, we can conclude that one has a belief, and the content of the belief can be determined by the actions that the belief allows the creature to successfully achieve.

1: Objections to Success Semantics

1.1: Indeterminate relation between beliefs and actions

One of the major objections to success semantics is that one's action may be caused by more than one belief. How can we determine which belief drives the action? For instance, when I fill up my car, I do not only have the belief that the fuel tank is empty (so I need to fill up the car now). It can be because the fuel gauge is not working, so I am not sure whether the fuel tank is empty or not, or it can be because the fuel price is low, so I better fill up as much as I can now, and so on. In responding to this objection, there is a "systematic" response. Suppose the success of my action is determined by a number of beliefs. In that case, each belief plays a role in determining the action(s). We get a systematic relation between one's beliefs and action(s), such that we still can determine the contents of one's beliefs through the success of performing certain actions. Thus, the belief about the empty fuel tank, the belief about the cheap fuel price, and the belief that the service station is just right on my way, all share a common feature that associates with the desire “fill up the car”., and those beliefs together are a system that together guides the successful action of driving to a service station, pulling over and filling up the car.

1.2: The Truth-condition insufficiency and no-impediment beliefs

Recall Whyte's claim on the truth-condition of one's belief,

> "A belief's truth-condition is that which guarantees the fulfilment of any desire by the action which, combined with that desire, it would cause."

Let us consider, in the case of my desire to fill up my car, my belief that the car needs to be filled up may be true, e.g. the fuel tank is empty, which combines with the desire not to let the fuel tank empty and causes me to drive to a nearby petrol station. Nevertheless, I may not be aware that the fuel tank lid is broken, so I cannot fill up the car. The situation is that my beliefs combine with my desire to produce the behavior of going to a nearby petrol station and trying to fill up the car, which does not fulfil my desire. (Brandom 1994)

Regarding this problem, Whyte (1990,1997) introduces the idea of no-impediment beliefs. For an agent to perform an action successfully, the agent must have the beliefs that there is no obstacle to his action. Whyte's idea is that, when we are preparing to act, we should also believe that we can successfully achieve the action. Put another way, if we know it is impossible to achieve an action, rationally, we would not try to perform that action. In the car example, if I know I cannot open the fuel tank lid, then I will not go to the petrol station. Instead, I will try to fix the lid first. In other words, when someone wants to perform an act, he/she must always have an additional belief that he/she faces no impediments. Whyte's idea is that when an agent fails to perform some actions, it is because that agent has some false beliefs that he/she believes he/she can perform those actions. In the case of filling my car, I fail to fulfil the desire of filling up the fuel tank because in acting on my ignorance about the condition of my car, I am actually acting on a false belief that I can open the fuel tank lid with no impediments (e.g. my false belief leads me unable to fulfil my desire). Whyte demands us that if we want to successfully perform an action, we must have a complete set of true beliefs (which combine with my desire), including beliefs of no impediments, such that my actions are guaranteed to succeed in fulfilling my desires. (Whyte 1990, 1997)

Whyte's no-impediment belief serves two purposes in explaining success semantics (Nanay 2012). First, it motivates one to perform an action. As in the car example, when I have a no-impediment belief (e.g. that my car is mechanically sound[23]) and a relevant desire (not letting the fuel tank goes empty), the no-impediment belief provides a factor for me to go to the nearby petrol station to fill up my car. Primarily, no-impediment beliefs play a "necessary role in motivating us to perform an action." (2012, p.154). Second, the truth of a no-impediment belief guarantees the success of an action. In the car example again, a no-impediment belief "my car is mechanically sound" guarantees that I can successfully drive it to a nearby petrol station and fill it up.

However, Bence Nanay (2012) argues that the first purpose of no-impediment beliefs does not help answering the problem of unknown circumstances; and the second purpose of no-impediment beliefs makes success semantics vacuous. (2012, p.154). In Whyte's account, to perform an action, there cannot be any impediment beliefs. If I know the fuel tank lid is broken, then I will not go to the nearby petrol station. However, Nanay points out that that kind of reasoning does not imply that if I have a belief that the fuel tank lid is fine, I must go to the nearby petrol station. What's more, even though having a no-impediment belief could motivate us to perform an action, it does not imply that motivation can lead to success in action. I may have a belief that my car is mechanically sound, it does not necessarily mean I can reach the nearby petrol station. For instance, I may have an accident while driving to the nearby petrol station. Having a no-impediment belief merely provides a piece of information that there is a good chance I can successfully perform an action. But it still does not guarantee the success of the action. What's worse, to *guarantee* the success of an action, in theory, we will need to have all sorts of no-impediment beliefs: so if I want to fill up my car, I will need to know the conditions of the traffic, the weather, the condition of the car, etc. But still, even though I may have known every necessary information, the action “fill up my car” may still not be able to fulfill, for example, the petrol station has power outage upon my arrival. Thus, an action may fail

[23] And it also can be more than one no-impediment belief, for instance, I have a true belief that the traffic is good)

without any falsity in the agent’s beliefs, contrary to Whyte's claim on the truth-condition of one's belief.

If Nanay is correct, the idea of no-impediment belief does not help in responding to the unpredictable circumstance problem. Since adding no-impediment beliefs still cannot guarantee the successful achievement of an action, Whyte's success semantics should be rejected.

Do we need to abandon Whyte's account?

The term "guarantee" is a very strong word, which implies something is necessarily a certain case, and strictly excludes anything that is not entirely fit into the domain. Do we have to be that restrictive in the sense of performing an action successfully? Perhaps we can refine the claim into a modest version, such that a no-impediment belief does not need 100% guarantee the success of an action, but reliably supports the success of an action. For instance, my car is mechanically sound would reliably indicate that I can drive to the nearby petrol station under most circumstances. Still, there is a chance that I will fail to reach the petrol station. So, in most circumstances, the content of my beliefs is reliable. Under this view, at least we can modestly save success semantics. Alternatively, we can refine Whyte’s principle: A belief's truth-condition is that which reliably supports the fulfilment of any desire by the action which, combined with that desire, it would cause.

Section Four: The Theory of Content and AI

The previous chapter discussed how the Chinese Room argument attacks AI, claiming that AI (or as described before a Turing-style computer) is only a syntactic machine; it has nothing to do with semantic content. While consciousness refers to the subjective experience of awareness, semantic content refers to the meaning of language, symbols, and other representations.

One way to understand the relationship between consciousness and semantic content is through the concept of qualia. Qualia are the subjective, conscious experiences that are associated with sensory perception, such as the experience of seeing the color red or feeling the sensation of pain.

Experiences have contents. Thus, consciousness and semantic content are closely related because both involve subjective experiences that are associated with perception and meaning. As we discussed in the previous sections, the Chinese Room argument argues that a mere syntactical machine without contents (semantics) is not sufficient for identical to minds (Minds have contents). For example, the word "Apple" has its content that refers to a fruit that has red-ness, sweetness, size, texture, etc different kinds of content. The contents of "Apple" hence mean something to us, and the contents may cause us having different responses in behaviors or thoughts. In other words, contents trigger our conscious activities, lead us to be aware of the external world and be responsive to the surroundings, without contents, symbols have no meaning to us.

In order to guarantee that a machine is properly aware of the world, we need a good theory of mental content. In this chapter, we have seen different theories explaining the relation between representations and their contents. The causal theory of content provides a straightforward explanation of how a representation gains its content through causal relations. However, the causal theory faces the problem of misrepresentation; it cannot provide a satisfactory answer to that problem.

Meanwhile, the teleosemantic theory of content provides an answer to the problem of misrepresentation regarding mechanical/biological malfunctioning. The downside of this theory is that its counter-intuitive view on the Swampman argument. In reply to the problem of misrepresentation, success semantic provides a solution that misrepresentation is realized by one's action. For instance, the content of my "dog" belief (i.e. misrepresentation) is realized by my action of treating that "dog" really as a dog, e.g. staying away from it to avoid being bitten. The advantage of success semantics is that it does not concern the history of the subject. As long as it can successfully perform actions, then we can conclude it has contents.

Consider the similarity between a freshly built computer and the Swampman. Assume the computer is fully capable of simulating human minds. When the computer connects to a power source for the first time, it starts running its program and simulation begins. The process will be similar to the Swampman that the moment it just comes out from the swamp. The crucial difference is that the computer was designed for whatever purposes, but the Swampman just came out from random-ness, purely no purpose. i.e. the computer is conventionally designed so it has intrinsic functions to achieve specific goals. While the Swampman is not designed so it has no intrinsic functions at all (but potentially can achieve some functions). To have the computer being conscious, according to Teleosemantics, what one is experiencing is a matter of logic on what happened yesterday or in the remote past. What one thinks and feels is likewise hostage to environmental and historical circumstances. At this respect, certainly a computer has a function, but it lacks a history of its own. Under this teleosemantic view, a computer can hardly be considered that it has contents. At this point, the success semantics theory seems to be our best candidate for a theory of mental content. We can accordingly consider how it could apply to an AI system. In brief, the computer can represent so long as it has preferences, or goals (which could be programable) and take action to fulfill its preferences, such that it can gain content through successful action. For example, we can build a machine with the goal of "stay running", which is literally equivalent to an agent's "desire" to do something. The machine will be built together with the ability to represent and interact with its surroundings, and take action to fulfill its goal. Under this view, the machine has gained content. However, it is far from a trivial matter to create a machine that has its own goals. In the final chapter, I will explore how this requirement depends upon being alive.

CHAPTER FIVE

Life and Consciousness

In the previous chapters, I have presented a theory of how consciousness is realized; and since we are conscious of content, I then have proposed a theory of mental content, i.e. how the relationship between a mental representation and its object is established. With these tools in hand, I shall discuss the relation between life and consciousness. My claim is simple, in this chapter, I argue that being alive is a necessary condition for mental content, and thereby, for consciousness. Thus, to build a conscious machine, we need to build a living machine first. In the meantime, the term “life” will be discussed.

Section One: Being alive is a necessary condition for mental content

The initial idea for the claim that being alive is a necessary condition for mental content comes from an observation about nature. When we look around the world, the only things we can be reasonably sure are conscious are also living beings. We might thereby start to wonder if there is an important connection between being alive and being conscious.

Scriven (1953) has a similar thought that life is a necessary condition for being conscious. He distinguishes between "unconscious" and "incapable of being conscious". The former implies consciousness, while the latter does not. He points out that "It is absurd to ask of a stone or a stop-watch "Is it conscious?" because it is absurd to talk of it being dead, asleep, drugged or stunned, i.e. unconscious." (P.232). In this sense, consciousness is a kind of capability and unconsciousness is a state of not performing or showing that capability. E.g. A unconscious perosn is a person that it is normally conscious but for whatever reason happens not to be conscious at this moment.

Thus, Scriven says that the question we should ask is, "Is a stone capable of being conscious?" In his view, being alive is one of the essential elements for being conscious. The doubt about whether an object is conscious or not arises only when that object is alive. Scriven argues that asking a machine whether it is conscious or not is absurd as it is not alive, simply because it lacks the essential element (i.e. living) for being conscious. Of course, that an object is alive does not necessarily mean that it is conscious. There are many other cases in which we doubt the capability of being conscious: plants, bacteria, earthworms, etc.

Regarding life and consciousness, Scriven's idea was that if an object is not alive, there is no point to ask “is that object conscious”.   (1953, p.232). Scriven's point is clear, being alive is necessary for being conscious. But he didn't go further to illustrate why being alive is necessary for being conscious. Meanwhile, he added a comment on the relation between consciousness and behaviours.

“A series of behavioural observations is not equivalent to the observation of consciousness and so it is proper to doubt if it is ever a proof of Consciousness. There is an essential connexion between the capacity for complex behaviour and Consciousness; the one is a -necessary condition of the other. But it is not a sufficient condition; and though we may decide which living things are Conscious from their behaviour, we cannot decide if everything is Conscious from its behaviour. Life is itself a necessary condition of Consciousness, and though behaviour is a factor which sometimes decides the question whether a certain system is alive, it is again not' the only one. " (p/234)

Here he argued that we cannot decide whether or not a thing is conscious from its behaviours, no matter that thing is alive or not. And certainly, in his view, if it is not alive, it cannot be conscious. He commented that machine cannot be alive, as he wrote,

" Robots, too, are machines; they are composed only of mechanical and electrical parts, and cannot be alive” (p.234)

If an individual is alive, its behaviours can be used as a criterion to determine whether it is conscious or not; however, if an individual is not alive, then its behaviours alone cannot be used to decide whether it is conscious or not.

Scriven further notes that perhaps machines are capable of having greater complexity of behaviour than living creatures. Still, he rejects the idea that such behaviours have essential connectivity with consciousness, even if the behaviours are as complex as human behaviours. In short, life is a necessary condition for being conscious.

Scriven's view on life as a necessary condition for being conscious is more like an assumption than an argument. This assumption is generated from an observation of the world, which at its best is merely an inductive argument. It is comparable to ancient times in which we could only see white swans, and thereby concluded that all swans are white. But we now know the claim that all swans are white is false, that there are black swans in Australia. Similarly, Scriven's claim, or assumption, that machines or any non-biological form of objects are not conscious (in the second sense we just discussed above) is merely an assumption based on the unfortunate fact that we have never seen any exception.

In contrast to Scriven, I have no difficulty imagining that when properly constructing a mechanical system, it may become conscious. If Scriven wants to exclude the possibility that machines or any non-biological form of objects cannot be conscious, he needs to explain *why* only living things can be conscious. We will discuss this point in the following sections, but in general, the reason is that living things have a preference to maintain their conditions.

At this stage, we could have two options in evaluating the relation between life and consciousness. We can either reject the idea that only living things can be conscious (So Scriven is wrong), or we can look further into the definition of being alive and see if machines can potentially satisfy it. I will choose the second option to examine the definition of life to see whether machines can be "alive" or not.

1.1 : Defining Life

If it is true that being alive is necessary for being conscious, then it means that if we want to build a conscious machine, we must build a living machine first. Just as in previous chapters we put efforts into explaining what consciousness is, we now have to discuss what life is.

How do we define “life”? There is no easy answer. I will not be able to review all the different definitions here. Instead, I will focus on some of the most respected definitions. One prominent definition appeals to the notion of autopoiesis (Maturana and Varela 1998). Autopoiesis refers to the self-reproduction and self-maintenance (autonomy) of a system. Self-reproduction refers to the ability of an organism to produce offspring that are genetically similar to itself. Many may argue that self-reproduction is a fundamental characteristic of all living organisms, from single-celled bacteria to complex multicellular organisms. In general, the process of self-reproduction involves the replication of genetic material (DNA) and the division of a single cell into two or more daughter cells.

Self-maintenance refers to the ability of an organism to maintain its internal state and physical structure in the face of external changes and challenges. This includes maintaining homeostasis, the process by which an organism regulates its internal environment to keep it stable and consistent. It also includes repair and maintenance of tissues and organs, the removal of waste products, and the acquisition of energy and nutrients. Both self-reproduction and self-maintenance are essential for the survival and success of an organism. Without the ability to reproduce, a species would not be able to continue and evolve over time. Without the ability to maintain its internal state, an organism would be unable to survive and thrive in its environment.

An autopoietic system is “A component production network that produces its own physical border” (Moreno and Eteberria 2005, P.164). In this view, a living thing is an individual system that can be distinguished from the surroundings. As Signorelli (2018) states, a unitary system is a “network of processes which interacts with the environment to keep their autonomy and increase their capability to reproduce” (Signorelli 2018, P.3). Where “a unitary system” could be understood as an individual system that cannot be divided, which is able to regenerate through its interactions and continuous transformation. (ibid)

Meanwhile, in the astrobiology community, life is defined as a “self-sustaining chemical system capable of Darwinian evolution.” (Benner 2010). “Self-sustaining” implies that a living system should not need any intervention by a higher entity to continue its existence, which further implies that “a living system” is driven to avoid being destroyed or end its existence. “Darwinian evolution” here refers to a “process that involves a molecular genetic system that can be replicated imperfectly, where mistakes arising from imperfect replication can themselves be replicated, and where various replicates have different 'fitnesses.' ”(Benner 2010).

Another key feature of a life is that living things react to external stimuli as a means of survival and adaptation to their environment. External stimuli, such as changes in temperature, light, or the presence of predators, can pose a threat to an organism's well-being, so the ability to sense and respond to these stimuli is essential for its survival. Reactions to stimuli are often controlled by the nervous system, which coordinates the organism's response to the stimulus. For example, if a predator is detected, an animal may respond by running away, hiding, or fighting. These responses are automatic and are triggered by the detection of the predator.

In addition to survival, reactions to external stimuli can also play a role in an organism's reproduction and social behavior. For example, mating behaviors and territorial displays are often triggered by external stimuli such as pheromones or visual cues. Overall, the ability to react to external stimuli is a fundamental characteristic of living things and is essential for their survival and adaptation to changing environments.

Living things are also sensitive to their internal environment, which is known as homeostasis. Homeostasis is the maintenance of a relatively stable internal environment despite external changes. Living things have various mechanisms in place to monitor and regulate their internal conditions, such as temperature, pH, and nutrient levels, to ensure that they remain within a narrow range that is suitable for their survival and optimal function. For example, the human body has mechanisms to regulate body temperature, such as sweating or shivering, to maintain a constant temperature. Similarly, the body has mechanisms to regulate blood sugar levels, pH balance, and electrolyte levels to maintain internal stability. These internal monitoring and regulatory mechanisms are essential for the proper functioning and survival of living organisms.

The approaches I have outlined here give us a general idea of what “living” means: an individual system that is capable of self-reproduction and self-maintenance, being sensitive to both internal states and external stimuli, and responsive to both internal and external changes that reacts to the environment, and is subject to evolutionary pressures. [24]

Yet we must note that the above ideas are generalized from our observation of life forms on this planet. Certainly, the above definitions of life are not totally undeniable. Still, one may argue that being reproductive is not necessary for life. There are numerous animals that cannot reproduce. For example, a mule is of an infertile sub-species. If we agree that self-reproduction is necessary for life, then we have to conclude that the mule is not alive, which is a highly unappealing consequence. Similarly, we may not need to accept the process of evolution as a necessary condition for life either, simply because we do not observe evolution happening in individuals. Thus, for the sake of our discussion, I propose that we focus the condition of life on the immediate case and consider *self-maintenance and reactance to the environment as our essential conditions of life*.

Self-maintenance in our world relies upon a certain kind of biological structure. One might intuitively generalize that living organisms have “bio-agency” (Meincke 2018, P.2). For instance, Moreno and Etxeberria (2005, P.162-163) claim that living organization depends on special materiality (i.e. bio-material, metabolism). Moreno and Etxeberria later conclude that “living organization is crucially dependent on the materials with which it is built. Hence, to reproduce this organization, we would have to use molecular—and even biomolecular—components.” (2005, p.171). Let us see how Moreno and Etxeberria reach this conclusion.

[24]

While there may be debates and disagreements surrounding the definitions of life, cognition, and consciousness, it doesn't mean that these concepts are completely elusive or devoid of meaning. Rather, they are complex and multifaceted, which makes them challenging to define precisely. However, researchers and scholars across different disciplines have made significant progress in studying and understanding these phenomena.

By exploring the relationships between these concepts, we can draw upon existing knowledge and theories to shed light on their interdependencies. While some aspects may remain unclear or contested, it is through these ongoing discussions and investigations that our understanding can evolve and become more refined over time.

Therefore, while it is important to acknowledge the challenges and debates surrounding the notions of life, cognition, and consciousness, it is equally valuable to engage in interdisciplinary dialogue and research to deepen our understanding of these complex phenomena. Such an approach allows us to explore the interconnected nature of these concepts and make meaningful progress in unraveling their mysteries."

First of all, consider what an agency is. According to Moreno and Etxeberria, to define agency, we need to "consider the task of self-construction while interacting with a given environment (that is to say, the relation with the environment can be considered to serve self-construction)" (2005, P.162). In Moreno and Etxeberria's account, self-construction is a process that has to do with maintaining the system. In the most common form, an organism's self-construction refers to the maintenance of its chemical structure. Organisms are chemical systems that conduct metabolic processes. For instance, organisms selectively intake substances (e.g. bacteria may intake carbon) and transfer them into their body structure to maintain their systemic functionality.

Most of the structures that make up organisms are made from the basic molecules, i.e. amino acids, carbohydrates, proteins, etc. As these molecules are vital for an autopoietic system to persist, metabolic processes focus on making these molecules or breaking them down as a source to supply energy to the organism. Metabolism plays an important role in autonomy. Metabolism drives the system to search for resources. This mechanism in nature gives rise to needs, "most importantly to the need for the supply of matter and energy to keep the metabolizing system stable in a state far from thermodynamic equilibrium." (Meincke 2018, P.28). If one's metabolism is disrupted, in the worst scenario, it may result in the collapse of the autopoietic system. i.e. for a living creature, it results in death.

Moreno and Etxeberria believe that the early life form's simple autonomy and its ability to interact with the environment cannot be distinguished, because both are based upon the same bio-mechanism. To maintain the system (life) or to deal with the environment is mostly a conceptual distinction. (P.162). For instance, a bacterium's life form is simple. Its self-maintenance system either derives energy from light through photosynthesis or breaks down chemical compounds through oxidation. The mechanism of how it maintains itself is also a matter of how it deals with the environment. E.g. bacteria use photosynthesis to gain energy to maintain their bio-structure, while conducting photosynthesis relies on it being properly positioned within the surrounding environment. When the environment changes, simple life forms have to adapt to the change. When simple life forms cannot cope with the rapid change in the environment, they quickly become extinct. Through evolution, life forms become more complex in order to cope with different environments and compete with different organisms. Multicellular life forms appear, then plants and animals.

And with the appearance of nervous systems, “organisms capable of sensorimotor coordination emerges” (P.162). The conceptual distinction between “living or biological processes (to keep the system alive, and reproduce and evolve) and adaptive or cognitive ones (act in the environment or plan behaviours in some other way)” (P.162) ceases to be merely conceptual or arbitrary, but acquires a structural correlate.

Consider now in contrast to current machines. Moreno and Etxeberria claim that the continuity between interacting with the environment and focusing on maintaining the machinery system is missing. (P.162). A machine can interact with the environment, but no process corresponding to self-maintenance can be found. Metabolic processes are vital for organisms. Existing machines lack metabolism. Without metabolism, the machine lacks the basic autonomy to maintain itself. Thus artificial agents, such as robots, cannot instantiate living agency without metabolism (Moreno and Etxeberria, 2005). Hence, under Moreno and Etxeberria's view, since machines cannot be autonomous thus they cannot be alive, therefore they cannot be conscious.

In order to respond to Moreno and Etxeberria's that inorganic machines cannot be autonomous, there are two options. The first option is to build a machine functionally and materially identical to metabolic self-maintenance. Alternatively, we can deny the claim that a living system requires metabolism to be self-maintain.

1.2 : Building Blocks

Moreno and Etxeberria deny option one. In their view, “the material structure of robots is very different from that of living systems, and this imposes qualitative differences on the form of organisation.” (2005, P.174). As we have seen, the difference between a living system and a machine is that a living system is an organisation of organic matters that can conduct metabolic processes. In contrast, a machine is a structure of inorganic matter. That means if we want an autonomous machine, it has to use organic matter instead of pure inorganic-mechanical components. Thus, to build a “living” system, we need to use organic matters as building blocks.

Overall, Moreno and Etxeberria’s account of living systems requires organic matters as the building blocks. In fact, it is a claim that denies that life can be defined purely in terms of functions. In chapter one, I discussed the way that something can be defined in terms of its function, for instance, mouse traps. A device is a mousetrap as long as it can trap a mouse, it does not matter what that device is made of, e.g. it can be made of plastic, wood, metal, etc. On the other hand, there are things that only can be defined in terms of what they are made of/ how they are constituted. For example, water has to be H2O. Even there is a liquid, which behaves exactly like water, e.g. colourless, liquid, wet, freezing at 0C, if it is not H2O, then it is not water. The water example is similar to Moreno and Etxeberria’s claim on life that life has to be constructed from organic matters. They claim that if a thing is not organic, then it is not a living thing. The reason is simple, it is because inorganic things are not metabolic. If a thing is not metabolic, then it is not driven to self-maintain. i.e. it has no autonomy.

Yet, why can't something functionally comparable to metabolism be realised by inorganic systems? Metabolism is ultimately a chemical process that makes necessary molecules for and provides energy to the organic system. If we can program a machine to realise the same functional characteristics as a metabolism has, theoretically, the machine has an artificial metabolism. Metabolism is a chemical process that takes place in cells. Cells are like factories that convert one substance into another substance (or energy). For each type of cell, what job they do is genetically designed. Perhaps we can program a machine that prioritises tasks like “searching for power source” and “self-repair”, which resembles the function of metabolism, and provide a ground for autonomy. If that is programable, we have a good chance to overcome the problem that inorganic machines lack autonomy.

1.3: Self-maintenance and Responsiveness to Surrounding Environments

A living system implies that the system itself has to interact with its surrounding environment to maintain itself. This is one crucial element of being alive. In this world we live in, *all living* *things have a tendency to maintain themselves, which ultimately is to stay alive.* Structurally, a living creature would do its best to seek the best environment for its survival. E.g. A tree has a mechanism for growing toward sunlight for photosynthesis. If it does not have that mechanism, it will hardly survive. Thus, “staying alive” is an intrinsic activity to all living systems. I.e. If a system S is a living system, then it inclines to perform that intrinsic activity, which is to stay alive. Consequently, if a system has a natural inclination to perform its intrinsic activity, the system itself should mechanically favour achieving that natural inclination.

The natural inclination (i.e. to stay alive) of the living system S grants itself an inclination to stay away from destruction. And because S needs to interact with the environment, it would be best for it to have sensors (in a natural living creature, “sensors” would be its faculties and organs) to detect the surroundings, in order to be aware of hazardousness and so to escape from danger, such that S can stay alive. Since S is required to have sensors, which implies that S must have a certain degree of complexity, it has a sufficient causal power to manipulate representations. In short, a living system has three crucial elements:

1. It has a motivation to maintain its own condition, and the motivation is intrinsic to the system.

2. To maintain its conditions, it must be sensitive to its physical conditions.

3. It needs to be able to detect its surroundings and be responsive to those surroundings in order to maintain its conditions.

4. It must have a certain degree of complexity.

Element 1 ensures that the system has a certain degree of autonomy. As we discussed earlier with regards to metabolism, Element 2 ensures that the system monitors itself so it can maintain the system as a whole and prevent itself from falling apart, or it can scan its internal status and seek energy supplies when it detects its energy reserves to be falling too low. Element 3 ensures that the system can respond to stimuli from the surroundings; for example, if the surrounding temperature is too hot, the system can detect the hazard and stay away from the surroundings to maintain itself from being destroyed by the heat. Element 4 ensures that the system is capable of manipulating data from sensory perceptions. Note, to detect the surroundings is a feature that requires sensors or special devices, that already implies the complexity of a system.

Generally, if an artificial system can be programmed[25] in an open-ended way to perform the above-mentioned features, such an open-ended program would have very minimal pre-set factors. If the artificial system can decide its own goals and strategies, act towards these goals, fit into the surroundings and develop a strategy to gain energy supply by using the surrounding resources.[26] Then behaviorally, the system is indistinguishable from an organic-autopoietic living system. Functionally that system would be a living system.

1.4: Mind and Life

I will return to the issue of whether machines can achieve life functions later on in this chapter. For now, let us turn our focus on the question “Why is being alive necessary for being conscious?”. We can review the meaning of “living” to answer the question.

"Living" means an individual system capable of autonomy, reacting to the environment, and subject to evolutionary pressures. Evan Thompson (2007, 2011) makes a similar claim regarding the concepts of autopoiesis, autonomy, life and sense-making (where sense-making is the basic mark of the cognitive (2011, p.36)). In his account, autopoiesis is "the paradigm case of autonomy, in the sense that it is the best-understood case and the minimal case of an autonomous organization" (2011, p.36). Autopoiesis and adaptivity are jointly sufficient for immanent purposiveness (self-organizing) and sense-making ("behaviour or conduct in relation to significance, valence, and norms that the system itself brings forth or enacts on the basis of its autonomy" (2011, p.36).). Since any living system is an adaptive autopoietic system (in Thompson's account), then life is an adaptive *autonomous* system, hence a sense-making system,

[25] The machine has to be programable. I cannot imagine how purely mechanical machines (e.g. those made entirely of wheels and sprockets) could be conscious.

[26] If it is a closed-end program specified for a particular task, e.g. detecting temperature, it can only be pre-set for that particular purpose.

and thus a cognitive system. As a result, cognition is necessary for life (2011, p.37). In order words, mind is necessary for life.

The above claim implies that when there is life, there is mind. This is to say, if we can build an artificial life, it automatically has a mind. However, we want to know whether life is necessary for mind, i.e. is autopoiesis necessary for autonomy (sense-making)? On the one hand, Thompson admits that "it seems conceivable that there could be an adaptive self-constituting system that was not based on autopoietic constituents." (2011, p.40). However, on the other hand, he analyzes the concept of genuine autonomy and quotes Barandiaran (et al. 2009):

> "It would need to (i) be an individual, in the sense of continually enacting or bringing forth its own existence in challenging thermodynamic conditions (…....); (ii) be the active source of its interactions, in the sense of modulating the parameters of its coupling with the environment on the basis of its internal (self-organized) activity ('interactional asymmetry'); and (iii) generate the norms for those interactions on the basis of its activity ('normativity')."

Thompson then points out that no existing robot meets these criteria, and it is even hard to see how this requirement could be met without something like a metabolism (Thompson 2011, p.41). In his view, autonomy must be realized by autopoiesis. Hence, if Thompson is right that autonomy and sense-making depend on autopoiesis, then it implies that life is necessary for mind.

Thompson's argument on the connection between mind and life is developed from an enactivist's point of view, which basically reduces mind (in its simplest forms) to biodynamics. He may be right on the claim that mind is necessary for life, but his claim that life is necessary for mind, in which the "mind" he refers to is minimal sense-making, is not sufficient for our use. In his terms, to be a sense-making system is to be a cognitive system, in a wide or broad sense of the term 'cognitive' (Thompson 2011, P.20). Thompson writes, "living is sense-making and that cognition is a kind of sense-making. A wave or a soap bubble is an individuating process but not a sense-making one, because it does not modulate its coupling with the environment in relation to virtual conditions and norms." (P.32). Thompson accepts that all living things are sense-making systems, thus cognitive. He writes, "Newen's declaration that bacteria, amoebae, and plants are not cognitive systems is question-begging, for I maintain that these organisms are sense-making systems, and thus cognitive in a broad (but well-motivated) sense of the term." (P.21). Apparently, his definition of sense-making is too broad for us. We want a theory that can explain more sophisticated mental activities, e.g. consciousness.

Nevertheless, Thompson's enactivist's notion of sense-making is similar to our view on how meaning is gained by subjects interacting with the surroundings, and how the subject's action is driven internally by their self-maintenance. Yet my thesis has been developed from a representationalist's angle. In the next section, I will present a representationalist theory of what it is to have a mind and how consciousness is realized.

Section Two: The Use of Information

1: The representationalist argument

We have seen that mental content enables living creatures to engage in sophisticated survival maintenance. Thompson, from an enactivist point of view, argued how life is necessary for sense-making. But, from the representationalists' point of view, how can we show that being alive is necessary for being conscious?

First of all, *mental states are physical states*. My account of the mind is clear that consciousness arises from purely physical processes. (Thus, consciousness can be described in terms of physical mechanisms). Meanwhile, representational states are states that represent something, e.g. properties, objects, or relations. Why does a system X have mental states that represent something? A reasonable reply is that mental states are for the sake of physical interactions between X itself and its surroundings. As we have discussed in the previous chapter regarding the success semantic theory of mental content, the content of a belief is "fixed by the success conditions for the performance of an action triggered by this belief." (Nanay 2013, P.2). And as Whyte claims, "A belief's truth-condition is that which guarantees the fulfilment of any desire by the action which, combined with that desire, it would cause." (1990, P.150). So, for instance, the content of my belief (i.e. representation) that my car is out of fuel is determined by the fact that this belief allows me to manage the fuel level in my car successfully, e.g. go to the nearest fuel station and fill up my car. As such, my mental content is determined by how it can be successfully acted upon.

Practically speaking, when X is interacting with the environment, mental representations arise involuntarily (through perceptions). E.g. When X perceives a visual stimulus, an image will be formed involuntarily in its mind. Since representations arise from interacting with the environment, representations indicate something about the environment. In other words, each representation carries information that points to its target. Consider ourselves as the example of how an agent handles its mental representations. When we have a piece of representation in mind, we may either ignore it or use it for other purposes. Indeed, it is so common that we ignore most of the mental representations as countless representations are being formed in our mind daily. We may "use" our mental representations. *"To use" implies that there are purposes that are needed to be achieved.* For instance, I am holding a cup. While I am holding a cup, I also see a table. I then have a mental representation of that table. Whether or not I use the mental representation of the table, depends on (not necessarily, but most likely) my preference of whether to hold the cup or not.

If I want to keep holding the cup in my hand, I simply ignore that mental representation of the table. But if I do not want to hold the cup any longer, then I will use that mental representation in my mind to adjust where I should place my cup. Holding a cup or not are different physical states. Thus, if I do not want to hold the cup in my hand anymore, I in fact have a preference to change myself into a physical state of not holding the cup. The preference to change myself into a particular physical state serves as a physical factor of why I use the mental representation. (And why I would have that preference would depend on my history and my physical state. In the cup example, for anyone who is holding a cup, naturally, they are likely to put the cup down somewhere.) *Thus, if I want to achieve certain purposes, then it implies I want to achieve, or reach, particular physical states.*

So far, I have connected the possession of mental states with preferences to change from one physical state to another. However, what exactly is meant by preference here? Preferences are proportionally responsible for a system's behaviors. In different situations, a system could have different preferences that it prioritizes to satisfy. Generally, *preference is either driven internally or externally*. If the preference is driven externally, then for instance, system Y dictates that preference Z into system X. So it is reasonable to ask the same question to Y, whether its preference W (implanting preference Z to X) is driven internally or externally, and so on, ad infinitum.

If the preference Z is driven externally, it means that an external force, in this case, System Y, is dictating or influencing the preference Z into System X. This implies that System Y has control over the preferences and is implanting them into System X.

It then poses a question: If System Y has this control and influence over preference Z, it is reasonable to question whether the preference W, which is the preference Z being implanted into System X, is driven internally or externally within System Y. In other words, does System Y have its own internal source of preferences that dictate its actions, or is it also being influenced by another external force?

This line of reasoning continues ad infinitum, suggesting that if the preference W in System Y is externally driven, then we can ask the same question about the origin of that external influence. This leads to an infinite loop of questioning the source of preferences and whether they are internally or externally driven.

It suggests that if preferences are driven externally, it raises the question of the origin and nature of those external influences, ultimately leading to an infinite chain of inquiry.

Ultimately, we shall need to determine a system for whom the preference is internally driven. *If the preference is driven internally, it must arise within the system.* The preference must either be a primitive, built-in preference (the preference is there since the system exists), or be originated within the system. *A preference that arises within the system indicates that the system has a goal that it prioritizes to achieve.* Actions that can be successful or not depend upon a preference that a system wants to achieve. The preference plays a role in motivating what actions the system is going to perform, such that the preference can be satisfied. In other words, the system's behaviors are determined by its preference. Thus, the content of a thought is determined by the fact that the thought allows the system to manage to satisfy its preference successfully. We can make the same point using success semantics. According to the success semantic theory of content, “the content of a belief is fixed by the success conditions for the performance of an action triggered by this belief.” (Nanay 2013, P.2). For instance, if a system has a preference to maintain its conditions, the content of its belief “there is an energy source” is then determined by the fact that the belief “there is an energy source” allows the system to manage its preference connect to a power socket successfully.

The process of manipulating mental representations is, in fact, "thinking". Given that, representational states are physical states. It follows that *manipulating mental representations are a process of manipulating physical states.* Thus, all kinds of thinking activities are in fact processes of manipulating physical states. As I have pointed out, every time I use representations, it is because I have a preference to change my current physical state into a different physical state. Therefore, every time I think or undergo any kind of thinking activity, it indicates that I have a preference to change myself in a particular state. Moreover, the reason why I think (I.e. manipulate representations, therefore manipulate physical states), is because I have internal preferences that are needed to be achieved. Generally speaking, *the use of a representation implies that the system must prefer to be in some states. That is, information is used to change a certain state of the user.*

It is worth noting that, our preferences that involve changing physical states are about responding to environmental circumstances. Representation is always about something. It carries information about that something. A photo of an apple is a representation of an apple. It carries information such as the colour and shape of that apple. A mental representation of X is a representation of X. It carries information about that X. Hence, when we consider thinking as a manipulation of representations, it is also a manipulation of information. Mental representations are most likely raised from sensory inputs. In other words, they are most likely raised from the external world. For instance, when I see an apple, I form a mental representation of that apple, and whether or not to use the representation of that apple is up to my preference. If I am hungry, I may have a preference to eat that apple. i.e. in my mind, I may use the mental representation of that apple, form a new mental representation “eat the apple”, and realize the new mental representation by actually eating the apple. Here I complete the process of thinking (manipulating a mental representation and forming a new one). My mental state, i.e. my representational state, has changed from a state of hunger to a state of less hunger.

Further clarification is needed that though mental states are physical states, it does not mean mental states are the actual physical states of a body. The body is in a physical state of needing energy, so it emits a signal of needing energy. We receive the signal of needing energy and form a representation of needing energy, i.e. we are in a mental state of needing energy. While the body may be showing signs of needing energy, e.g. the stomach feeling empty and grumbly. For some reason, possibly we are concentrating on other things, we are not aware of the signal of needing energy, as we are not in a mental state of needing energy. Vice versa, we may have a thought of “needing energy” while we actually do not need it. For example, we may have a thought of “needing breakfast” in the morning even though our body does not show any sign of needing energy.

Meanwhile, when I introduce the idea of change, I bring the time-factor into the nature of representations. Often when we discuss the nature of representations, we consider them, just like photos or drawings, in a static manner. When we discuss the nature, or the characteristic, of a photo, the content of the photo is taken as static; we could hold the photo in hand, examine it from different angles. Similarly, when we discuss that mental representation, we indeed assume that that representation is a kind of non-changing, static form of entity. However, our perceptions of a certain object are constantly changing, co-variant with our physical conditions and the surroundings. Analogously, a picture of a river cannot fully capture the running entity of a river. We need to inclusively consider the time-factor in order to fully capture the nature of a river. Similar to mental representations, to capture their nature, we need to consider time-factor too, because they are constantly changing. Thus I emphasize that there is no single frame of mental representation, but only a constant, continuous stream of mental representations.

Why does a system have a preference to change its current state into a different state? Consider the cup example again. My current physical state is holding a cup in hand, the reason why I want to change my current physical state can be simply because I prefer a different physical state, for instance, to get my hand free from holding the cup. After I change to a hand-free state, I would maintain myself in that physical state until I have another preference, e.g. I would keep my hand free until I want to hold something else. However, the maintaining time would not last long. Since I am constantly interacting with the surroundings, numerous representations arise both voluntarily and involuntarily, which implies that I have to constantly use or ignore numerous representations. Either way, *it indicates that I have further preferences to either maintain myself in the current physical state (i.e. Hand-free) or change to a different physical state* (e.g. holding something else). As long as my brain (and my faculties) functions normally, representations arise constantly and continuously.

To change myself in a particular physical state implies that I have to maintain myself (my own existence) such that I can change myself into that particular state. More specifically, I am in fact maintaining a continuous sensitivity to the environment and myself. The activity of mental representation is essentially aimed at maintaining that sensitivity. For example, consider deliberately maintaining a vehicle running at a certain speed. To do so, we need to maintain the physical structure of the vehicle so that it can run at that speed. Similarly, suppose a system has a preference to change itself into a particular state. In that case, it has to maintain itself (its own existence) before it can change itself into that particular state. For this reason, thinking implies the maintenance of self, which implies life. Let us reconstruct the argument in a clarified form:

1. If X is an agent and it forms a representation of Y, then X can use the representation of Y.

2. If X uses the representation of Y, then the representation of Y allows X to bring about some outcome Z that X prefers. If the representation of Y allows X to bring about some outcome Z that X prefers, then X prefers to be in a condition in which it represents outcome Z.

3. If X prefers to be in a condition in which it represents outcome Z, then X prefers to represent outcome Z.

4. To represent outcome Z is a new physical state of X.

5. If X prefers to represent outcome Z, then X prefers a new physical state of X (from 4, 5). Then if X is an agent and it forms a representation of Y , X prefers a new physical state of X) (from 1-5)

6. If X continuously represents Y, then X continuously prefers new physical states of X.

7. If X continuously prefers new physical states of X, then X must maintain its condition of sensitivity to its own physical states.
8. To maintain X's condition of sensitivity to its own physical states requires X maintains its physical structure such that X can continue being sensitive to its owner physical states.

9. If X maintains its condition of sensitivity to its own physical states, and if X continuously represents Y, then X is alive.

Conclusion:

10. Thus, if X is an agent and it forms a representation of Y, then X is alive.

X is the user of the representation of Y. The user of a representation is the subject that manipulates the representation, which is different from the representation itself. For instance, when I see an apple, I form a representation of that apple, such that I have a representation of that apple in my mind. I can use that representation if needed. If I have a preference "to eat something", then I will manipulate (use) the representation of that apple to form a new representation "eat that apple". If I have no preference of eating anything, I will not further manipulate the representation of apple in my mind.

Premise 1 establishes the relation between the user of representation and the representation itself. draws on the success semantics theory that X uses the representation of Y (if I see a table and form a representation of that table; then I use that representation). Premise 2 identifies that if an agent uses a representation, then there will be an outcome that that agent prefers, i.e. whether an agent uses a representation or not depends upon preferences (To use the representation of a table, it allows me to bring about an outcome that I prefer, I.e "place my cup on a table"). Premise 3 follows directly from premise 2 and the idea that the subject can only complete an action by registering whether or not the goal has been satisfied. Premise 4 follows from a materialist theory of representation which states that a representation state is a physical state. Premise 5 and Premise 6 indicate that X uses representations to change into a physical state that it prefers. Premise 7 is where I connect representational preferences with the key conditions of life for which I argued in Section One: Self-maintenance and responsiveness to surrounding environments. X has a preferred physical state, so to reach that physical state, it will constantly maintain a sensitivity to its own physical state. (I prefer a hand-free state, so I constantly maintain my sensitivity to my own physical state, see if I can reach my preferred state.). Premises 8 and 9 explain that if X maintains its condition to be sensitive to itself, X has to maintain itself in order to be sensitive to itself, which is the key condition for being alive. In P.131, I state three crucial elements for a living system, being sensitive to its physical conditions is one of the elements. allows a system to identify and address its own needs. In addition, earlier in P.139 I use the example that prior to maintaining a vehicle running at a certain speed, we need to maintain the physical structure of that vehicle such that it can maintain its speed instead of falling apart. That example explains if a system is maintaining itself continuously running, it implies that the system has to maintain its physical structure (i.e. its existence) prior to its performance. Hence, being sensitive to the system's own physical conditions implies that the system has to maintain its physical structure first, thus it is alive. Therefore, we conclude that if X represents Y, then X is alive.

Overall, since X is a conscious (and physical) system, it has mental states, and it uses the information that the mental states carry to achieve its preference, i.e. to maintain itself in some particular physical states. In other words, X has preferences, and the preferences are fulfilled by reaching some particular physical states. As mentioned previously, X’s preferences depend on its history and physical states. For example, if X is in a state of needing energy, it would have a preference to achieve that physical state. Let us consider that X has a mental representation of energy supply (e.g. X sees something it can consume). The representation carries information about energy supplies which can get X into its preferred physical state. How to get X into its preferred physical state? We have discussed the idea of impediment beliefs in chapter 4. Certainly if we want to guarantee a success in performing certain actions, we will need to know all non-impediment beliefs such that we can decide how to achieve the outcome we want.

However, “guarantee” is a strong word, which implies something must be achieved, and strictly excludes anything that does not entirely fit into the domain. Perhaps we can refine the claim into a modest version, such that a no-impediment belief does not need 100%guarantee the success of an action, but reliably supports the success of an action.

For a living thing, a preferred physical state is a motivation for it to perform behaviors. When the living thing is in a physical state of needing energy, i.e. it has a representation of needing energy. The representation plays as a goal to motivate the living thing to fulfill the need, such as looking for food. In reality, there is no guarantee that the need can be fulfilled. However, for a simple living thing, e.g. an ant, will still try its best to fulfill its need. Thus, the representation indeed has a genuine content for X to fulfil its preference of getting energy supply. Therefore, X would use that information in order to reach the preferred physical state (e.g. being fed).

At this stage, we know that X has an internal preference. A pure internal preference implies the preference of the system is not driven by external factors. That further implies that the system has a minimal degree of autonomy. The system's preference is guided by its own needs. The needs for the system guide its behaviours. As we have just seen, X has a representation of needing energy supplies and that representation comes from the continuous self-monitoring, so X has a preference of gaining energy. The representation of needing energy supplies can be orientated from its physical structure. For instance, in a bio-system, the metabolic structure provides the system with a natural need for intake of substance and energy. Every particular kind of cell would have a natural need for a particular substance, e.g. some may need more oxygen, others may need more glucose. The physical structure (or bio-structure) of a cell determines what it needs. Similarly, we can program a machine to monitor its physical state, such that the machine can monitor its condition, in order to determine what it needs to maintain its structure and functions. For instance, if it needs power, it can look for a power supply so that it can stay running. If the system X has a preference, it means it has a preference to use certain information to reach its preferred physical state. And if X has a preference, which further implies it has a particular internal need, and the pure internal need indeed implies X has a minimal degree of autonomy, i.e. it has a minimal degree of living.

2.2: Living Machines?

Throughout the discussion on being alive, I emphasize the importance of self-maintenance. An opponent of this argument may propose the following counterexample to suggest that a preference to maintain itself in a particular physical state does not imply that the system has a minimal degree of life. Consider a metal spring. When we compress or stretch a spring, it has a physical tendency to return to its original shape. According to our description earlier, a spring tends to maintain itself in a particular physical state, which should be considered a form of life.

We can compare the spring to a conscious system X. X would constantly maintain its sensitivity to the environment, and constantly and proactively make use of the representations in order to achieve a physical state that that system prefers to be. No representation is created by the spring, and it does not use any information to maintain its physical state. *It is important to note that system X 's preference can be different in different situations and at different times*. For instance, when X is in a hot environment, it may have a preference to be in a cooler environment. X's preference may co-vary with the external environment. However, consider a spring that sits in its natural position, unlike the conscious system X, the spring does not represent anything while it sits there. The spring has one and only one physical preference, independent of what environment it is in: to maintain its most stable position. Unlike the system X which will have different preferred physical states while it is in different environments. The spring will sit there forever if there is no external force acting on it.

In other words, there is no representation or information that the spring uses to maintain itself in a particular physical state. The spring brutally and directly bounces back to its original shape. It is the constant use of information to interact with the environment and the responsiveness to different physical states that distinguishes a system whether it is alive or not.

A metal spring is a simple mechanical device, and we can be certain that it does not represent. What about a more complex device, such as a thermostat, does it represent? A thermostat is a device sensitive to temperature. A typical one is used in engines to control water flow inlet towards the engine block. When its surrounding temperature reaches a certain degree, the thermostat will open up its gate and let water flow towards the engine block to cool down the engine. A thermostat does maintain its sensitivity to itself, and if it is heated to a certain degree, it will be in a physical state “open”. We might even claim that it has an internal preference to maintain itself in a physical state “close”. Yet similar to the metal spring case, a thermostat does not represent, because it does not use any information to reach its preferred physical state, i.e. a thermostat just mechanically responds to its target temperature. Thus, a thermostat cannot be counted as a living thing.

What if we upgrade a thermostat into a digital unit, such that the timing for the open-gate is controlled by a control panel, does it count as a living thing? For instance, the control panel will monitor the water temperature. Once it reaches a certain degree, it sends a signal to the thermostat to open its gate and let water flow into the engine block. This digital thermostat seems to more closely fit in our model: (1) it constantly maintains sensitivity to its physical state, (2) it represents temperature and use the representation (i.e. information) to send signal to the gate in order to change its physical state, (3) it has internal preferences to maintain itself into a certain physical state. Therefore, it should be considered as a living thing. I would accept (1) and (2), but not (3). The crucial point of the failure of (3) is that the digital thermostat's preference is not internally driven. Consider how the digital thermostat gains its preference (i.e. at what temperature to do what). This preference is mechanically pre-set by an external controller. The thermostat does not have any self-driven preference, e.g. it will not pre-set a temperature for its own sake. Moreover, the thermostat does not maintain itself at all. It does not have any sensitivity in maintaining its physical structure. As a result, the digital thermostat still does not fit in our model, i.e. it is not an artificial life-form.

So, what kind of machine could fit in our model? To represent and maintain sensitivity to its physical state, these two functions can be achieved by using different types of sensors, e.g. thermo-sensor, vibration-sensor, pressure-sensor, Etc. The machine also must have a certain degree of complexity, so it has a sufficient causal power to manipulate representations. The core problem is how to create a machine with internal preferences. From the digital thermostat example, we know that the preference of the machine cannot be pre-set. If so, the machine's preference, in fact, is externally driven.

An internally driven preference is one where the behavioural commands come from the physical tendency of the parts. Here, behavioural instruction directly originates from the need of the system to maintain itself. Note that the system should be vulnerable and complex[27], which constantly needs to be maintained. Vulnerability in living organisms refers to their susceptibility to harm or damage from various factors such as environmental changes, diseases and predators. Living organisms can be vulnerable at different stages of their life cycle, and their vulnerability can be influenced by a variety of factors such as genetics, behavior, and environmental conditions. Being vulnerable can stimulate the development of adaptive mechanisms that enable organisms to cope with stressors and recover from damage. For example, exposure to moderate levels of stressors can induce the expression of stress-response genes and trigger physiological and behavioral responses that help organisms to withstand subsequent stressors. As such, a living organism needs to constantly monitor itself in order to maintain and respond to different harm. *The constant need for maintenance gives rise to preferences.* Not surprisingly, the vulnerability of a system more or less depends on its substrates and structures. For example, a system that is constructed of steel may be too rigid to be changed. Hence, naturally, it does not have a constant need to be maintained, or the need is minimal. On the other hand, in nature, organic systems that generate metabolism are vulnerable systems such that they have a constant need for self-maintenance and their preferences arise within the system itself. So, in this sense, internal preferences could be driven by the substrates and structures of the system.

By using suitable substrates and structures, we might build a system with internal preferences that constantly require specific responses to be maintained.[28] The downside of such bio-structure is that it is vulnerable to extreme environments. For example, a regular mechanical thermostat can function well in high temperatures but a bio-thermostat will not last long.

Some may argue that inorganic systems can also be vulnerable. For example, a clock will break if we drop it. A machine can be vulnerable too, so what is the difference between a vulnerable machine and a vulnerable living system? The answer can be referred to thermodynamics.

[27] A complex system would have more components and structures, which certainly needs more frequent maintenance.

[28] Note that organic matters alone or a system constructed by organic matters do not guarantee the rise of internal preferences. For instance, a just dead fish is an organic system, but it does not have preferences either. A system that is constructed of organic matters just merely has a potential to be a system with internal preference.

Thermodynamics deals with the relationships between heat, energy, and work. It is concerned with the behaviour of macroscopic systems, such as gases, liquids, and solids, and how they respond to changes in their surroundings. The Second Law of thermodynamics states that the total entropy (disorder) of an isolated system always increases over time, unless work is done to decrease it. Thermodynamic equilibrium is a state of a thermodynamic system in which the macroscopic variables that characterize the system, such as temperature, pressure, and density, do not change over time. In other words, the system is in a state of balance where there is no net transfer of energy or matter between different parts of the system.

At thermodynamic equilibrium, the system's entropy is at its maximum, which means that it is in a state of maximum disorder. This state is reached when the system has reached its maximum level of energy distribution, and any further exchange of energy or matter does not lead to any further changes in the system's macroscopic properties.

Consider living systems. Vulnerability arises from the complex and dynamic nature of living systems, which are composed of multiple interacting components that are interconnected and interdependent. A living system is vulnerable because any changes in one component of the system can affect the stability and function of the whole system, and disruptions or imbalances can cause dysfunction or failure. Though living systems are systems that they are constantly exchanging energy and matter with their environment. They rely on the continuous input of energy and the removal of waste products to maintain their structure and function. This continuous flow of energy and matter allows living systems to maintain the systems in a state high organization and complexity, despite the second law of thermodynamics, which states that entropy of an isolated system always increases over time. A living system is far from thermodynamic equilibrium.

Meanwhile, consider a clock. It is vulnerable as we can easily break it by dropping it. In a mechanical clock, for example, the energy needed to keep the clock running is provided by a wound-up spring, which gradually unwinds and loses energy due to friction and other forms of energy dissipation. The gears and other components of the clock are subject to wear and tear, which can cause them to become less efficient over time. When the clock is not able to access an external source of energy to replenish what it loses, it will eventually reach a state where it is no longer able to function, the system is in a state of balance where there is no net transfer of energy or matter between different parts of the system. The clock will eventually reach a point that same temperature as the surroundings and it may be considered to be closer to thermodynamic equilibrium at that point.

Thus, the difference between a vulnerable machine and a vulnerable living system is that the vulnerable machine is towards thermodynamic equilibrium while a vulnerable living system is not.

In sum, internal preferences come from the needs of the vulnerable system. A vulnerable body, e.g. an organic body, because of its vulnerability, naturally requires constant attention to maintain its physical structure.

2.3: Elmer and Elsie

Non-organic machines are not as vulnerable as organic systems, so can internal preferences arise in them? For example, a steel-made machine may be rigid, but it still progressively gets weaker if it lacks maintenance. In other words, machines are vulnerable, but they are just on a different scale when compared with organic systems. Furthermore, the vulnerability of machines may not be noticeable to human eyes but a machine can be given sensors to detect its specific maintenance needs. What matters is that preferences that constantly require specific responses to be maintained can still arise. In this sense, substrates are not a barrier to building conscious machines.

Meanwhile, in section 1.3, I have pointed out that one of the features of an autonomous machine is that the program has to be open-ended. For instance, an open-ended program such as "maintain the system running" or "maintain the system physically as a whole” requires the machine to decide the best strategy to survive.

To be self-maintaining, the machine needs to be able to constantly and continuously monitor its own status, to detect what parts of the system need to be fixed. Such internally driven preferences determine whether the content is meaningful to the system or not. For instance, my (internally driven) preference to be hands-free will make the representation of placing my cup on a table meaningful to me.

So, can that machine be built and programmed? As early as the 1940s, robots equipped with similar features had been built. Elmer and Elsie (ELectroMEchanical Robot, Light-Sensitive) [29] were two electronic 'tortoises' built by neurobiologist Grey Walter in the late 1940s. The robots were equipped with light and motion sensors, such that they were capable of phototaxis which is the movement that occurs in response to light stimuli. They would move towards light sources and respond to motion/pressure so that they would move away around obstacles. The robots were allowed to wander around the floor without any pre-set pattern. As Walter writes, "These machines are perhaps the simplest that can be said to resemble animals. Crude though they are, they give an eerie impression of purposefulness, independence, and spontaneity." (Cited in Holland 2003, p. 16).

Restricted by then-technology, the tortoises could not be built with better functions. For example, they could have been built with a battery level monitoring sensor and positioning sensor, such that the robots can monitor their battery level. Once it goes low, they can go to the charger to recharge its battery. And perhaps the robots can keep checking how far it is from the charger, ensuring it is in range to regain energy. Given these features, the systems appear to have a preference to maintain themselves running. This feature potentially satisfies the characteristic of a living system, or at least it is going towards looking more like a characteristic of a living thing. Certainly, other features can be added to the tortoises so that they can maintain themselves more effectively, e.g. sensors for their structure rigidity, the ability to replace parts on their own, etc.

For now, the tortoises imitate a very minimal form of life, but are they intelligent? If we consider intelligence as the capacity of any system to take advantage of its environment to achieve a goal (Signorelli 2018, P.7), then the intelligence of a living system would be the ability to take advantage of its environment to stay alive. Under this view, the tortoises may be considered intelligent as they can sense the intensity of light around them and decide where to go. The more

[29] Elmer and Elsie were two electronic tortoise-like machines built in the late 1940s by neurobiologist Grey Walter. The tortoises equipped with sensors and they were allowed to wander around the floor with no fixed pattern. https://gizmodo.com/the-very-first-robot-brains-were-made-of-old-alarm-cl-5890771

intelligent a living system is, the more significant advantage of the environment it can take. A simple living system, such as a bacterium, may not need very sophisticated exploitation of its environment to maintain its living conditions. Thus, it does not need to be highly intelligent.

On the other hand, a more complex living system would undoubtedly need to take greater advantage of the environment to maintain its complex form, which entails it needs to be relatively more intelligent. For example, an insect needs to deal with a more complex environment in order to survive. Thus bio-mechanically, it will be more intelligent than a bacterium.

Conclusion

Organic matters could be the building blocks for conscious machines. However, that is not the objective of this thesis. From a computationalist's account, this thesis aims to explore the possibility of building conscious machines in a computational manner. Using organic matters to build a living system is not our interest. Nevertheless, suppose metabolism is functionally required for giving rise to consciousness. Metabolism is essential to life as it is the set of chemical reactions that occur within living organisms to maintain life. It is involved in a wide range of biological processes, including energy production, growth, reproduction, and the regulation of bodily functions. Energy production is one of the most important functions of metabolism, as it allows cells and organisms to convert food into energy that can be used to power biological processes. This energy is produced through the process of cellular respiration, which involves the breakdown of glucose and other molecules to release energy. Overall, metabolism plays a critical role in the proper functioning of the body. A malfunctioning metabolism may result in the collapse of the body. In general, any process that involves the conversion of energy from one form to another, or the synthesis of new molecules (i.e. growth) can be considered functionally equivalent to metabolism. Hence, what we require is that, at most, the program that runs on a computer must have features that are functionally equivalent to a metabolism, which is the system has the function to convert resources into energy supplies to the system that can be used to power physical processes, such that it can self-sustain So the point is, even though organic matters could generate metabolism such that the system could give rise to internal preferences, that does not imply that non-organic matters could not generate internal preferences with a similar function that has the same causal power as metabolism has.

CONCLUSION TO THIS THESIS

To end this thesis, I would like to re-emphasise the crucial point I have made that being alive is necessary for being conscious. To build a conscious machine, we need to build a living machine first. Our discussion began with exploring what consciousness is. Block distinguishes between phenomenal consciousness and access consciousness, and we identify that phenomenal consciousness is our target to be explained. We took a functionalist approach to try to explain and understand what consciousness is. According to functionalism, what consciousness is does not depend on what substance it is made of; rather, it depends on its causal role in relation to the circumstances and other states of mind. In other words, what defines consciousness can be considered in terms of what function it is. For example, if a system can realise such and such consciousness-functions, we consider that system is conscious. Thus functionalism provides us a ground that it is theoretically possible to build a conscious machine without using organic substances or bio-systems.

In Chapter Two, I begin with the representationalist position of consciousness by considering that all mental activities are representational states. Thoughts, desires, feelings, emotions, etc., are all representations. The key point is where and how consciousness arises. Next, I examined First-Order Representationalism and realised its main problem: First-Order Representationalism cannot differentiate the nonconscious state from the conscious state. Blindsight cases are strong evidence showing that a nonconscious mental state can affect our behaviours, which contrasts with first-order representationalists' claim that what makes conscious experience arise is determined by the content of the representation available to thought and reasoning, and for the control of actions.

I later turned our focus on Prinz's theory about attention and consciousness suggests that attention is the key factor differentiating between conscious and unconscious perception. Prinz defines attention in terms of availability to working memory. That is, we are attending to an object X if and only if the perceptual information of the X is available to working memory. Yet if consciousness can only be discerned through the report, and the report automatically indicates working memory, and working memory automatically indicates attention, it will never be possible to get consciousness without attention. In other words, Prinz's theory is unfalsifiable.

At the end of Chapter Two, I presented the Higher-Order Representation Theory of Consciousness. The key idea of HOR is the Transitivity Principle, which states: A mental state is conscious, if and only if, the subject is aware of itself being in that state. For instance, if I am looking at the sky, I have a mental state "The sky" in my mind (the mind here I adapt Prinz's notion of working memory). "The sky" is a conscious state if and only if I am aware of "I am looking at the sky". The term "aware of" is interpreted as "having a representation of". Hence, "The Sky" is a conscious state if and only if I have a higher order representation of "I am looking at the sky".

The theory I proposed adapts Prinz's account of how working memory functions and further develops that we need to *use* the information stored in working memory to claim we are conscious of that information. The concept "to use" plays an essential role in this theory. If we want to use something, that indicates we have a purpose or preference that we want to achieve. One of the strengths of my theory is that it can explain why we are conscious. In general, we are conscious because we have a tendency to self-monitor what state we are in, which is due to the motivation for survival. In other words, *we are conscious because we want to be conscious in order to increase survival chances*. Another advantage of my theory is that it can explain that human infants and other non-verbal or non-reportable non-human animals all could have conscious experiences, as long as they can self-monitor and use the information being stored in working memory.

In Chapter Three, I shifted our focus to the Computational Theory of Mind. The claim is simple: Minds are Computations. The idea is that minds are representation-manipulation processes while computations are symbol-manipulation processes. Since representations are symbols, minds can be analyzed as computations. In this chapter, I discussed the details of different key terms, e.g. what a symbol is, syntax, semantics, its content, how a computation functions, how a Turing machine works, how this theory explains the nature of the mind, and how the famous Chinese Room Argument argues against this theory. In short, symbols are representations. They contain syntax and semantics values. This is the point that Searle uses to attack the Computational Theory of Mind: since computation only deals with syntax, but not semantics (or contents), and syntax is not identical with nor sufficient by itself for semantics, but minds have contents. Thus computation is not sufficient for nor identical to minds. Hence, machines do not understand contents and computation cannot give rise to minds, i.e. consciousness cannot be realized through computation.

The critical point of the Chinese Room argument is that it pinpoints the problem any formal machine would encounter: they are syntax engines. The entire process of symbol manipulation does not involve any semantic contents, which are the chief factor for mental contents. In response to the Chinese Room argument, we need a theory of content to show how machines can gain content. This is the aim of Chapter Four.

Chapter Four aims to explore the relation between representations (i.e. symbols) and their content. It shows that machines can gain content, so to respond to the challenge from the Chinese Room Argument that computation has nothing to do with contents. I discussed the causal theory of content, teleosemantic and success semantics.

After reviewing these theories, I ultimately supported success semantics, according to which, "the content of a belief is fixed by the success conditions for the performance of an action triggered by this belief. "(Nanay 2013, P.2). In other words, the truth condition of one's belief is fixed by the success conditions of one's action. There are three elements in the account of success semantics: desire, belief and action. When a desire arises, it plays a role in motivating what we will do, and the ultimate goal is to satisfy that desire. Under this view, the Swampman's mental state's content is determined by its action's success. The advantage of this account is that it does not need to consider one's history. As long as one successfully performs an action, we can conclude one has a belief, and the content of the belief can be determined by the actions that the belief allows the creature to achieve successfully. Notably, based on success semantics, a machine can gain content through the success of its action.

In Chapter Five, I argued that being alive is necessary for mental contents. In other words, if we want to build a conscious machine, we must build a living machine first. Recall the three elements in the account of success semantics: desire, belief and action. Desire plays a role in motivating what the subject is going to do, and by all means, all action performed by the subject is to satisfy that desire. The content of a belief is determined by the success of its action.

Therefore, desire has to arise in the subject, such that the action of satisfying the desire is meaningful to the subject. Being alive ensures that the system has a preference (desire) to stay alive. Thus, all its actions are to satisfy that preference. Therefore, actions performed by the system are meaningful (have contents) to the system. That is one of the reasons why being alive is necessary for mental contents and thus for consciousness.

Meanwhile, I argued from a representationalist's view that if a system can represent and has a preference to use that representation, then that system is alive. To build a living machine, we need to know what "living" is. In this chapter, I also discussed the term "alive". A living system has to have an internal preference to perform self-maintenance, be sensitive to its own conditions and have sensors to detect its surroundings. I reject the idea that a living system has to be built by organic substances and insist that being alive can be realized in terms of functional notions.

At last, let me summarize the flow of our strategy. We want to build a conscious machine. Once we understand what consciousness is and its functional notion, we can confidently claim that it is possible to build an artificial consciousness system if the system can realize the functional characteristic of consciousness. Second, we grant that mind is a computation, and conscious states are computational states. What we need to do next is that we have to explain how a machine can gain contents. Success semantic theory of content shows us how a system can achieve content. The final task is identifying the necessary condition for constructing a conscious system. Once we identify the necessary condition and show that it is possible to realize that necessary condition, i.e. build a living machine first, then building a conscious machine will be merely a technological problem.

This project has ended here. Affirmatively, it is possible to build a conscious machine.

REFERENCES

Alvaro Moreno, Arantza Etxeberria; Agency in Natural and Artificial Systems. Artif Life 2005; 11 (1-2): 161–175. doi: https://doi.org/10.1162/1064546053278919

Baars, B. (1995). Evidence that phenomenal consciousness is the same as access consciousness. Behavioral and Brain Sciences, 18(2), 249-249. doi:10.1017/S0140525X00038218

Baars, Bernard. (1997). In the Theater of Consciousness: The Workspace of the Mind. 10.1093/acprof:oso/9780195102659.001.1.

Benner S. A. (2010). Defining life. Astrobiology, 10(10), 1021–1030. https://doi.org/10.1089/ast.2010.0524

Birgit Mampe, Angela D. Friederici, Anne Christophe, Kathleen Wermke, "Newborns' Cry Melody Is Shaped by Their Native Language" , Current Biology, Volume 19, Issue 23, 2009, Pages 1994-1997.

Block, Ned. (1980). Troubles with functionalism. Readings in philosophy of psychology, 1, 268-305.

Block, Ned. (1995) "On a confusion about a function of consciousness." Behavioral and brain sciences 18.2 (1995): 227-247.

Block, Ned (1995), "The Mind as the Software of the Brain", edited by D. Osherson, L.Gleitman, S. Kosslyn, E. Smith and S. Sternberg, MIT Press, 1995). online paper: https://www.nyu.edu/gsas/dept/philo/faculty/block/papers/msb.html#4.

Brandom, Robert B. (1994). Unsuccessful Semantics. Analysis 54 (3):175-178

Brooks, Rodney (2001)“Steps Towards Living Machines” in *Evolutionary Robotics, 2001,* Takashi Gomi ed,

Cain M.J. (2002), “Fodor Language, Mind and Philosophy”, Polity Press in association with Blackwell Publishers Ltd

Carruthers, P. (2000).*Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Carruthers, Peter & Veillet, Benedicte. (2007). The Phenomenal Concept Strategy. Journal of Consciousness Studies. 14. 212-236.

Chalmers, David J. The conscious mind: In search of a fundamental theory. Oxford Paperbacks, 1996.

Chalmers, D. (2007). The hard problem of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 225–235). Blackwell Publishing. https://doi.org/10.1002/9780470751466.ch18

Church, Jennifer. "Two sorts of consciousness?." Communication and Cognition: An Interdisciplinary Quarterly Journal 31.1 (1998).

Cochrane, Tom. (2019). The Emotional Mind. In “*The Emotional Mind: A Control Theory of Affective States”* (pp. I-Ii). Cambridge: Cambridge University Press.

Csikszentmihalyi, Mihaly (1990). Flow: The Psychology of Optimal Experience. Harper Perennial Modern Classics. p.80.

Davidson, D. (1978). What Metaphors Mean. *Critical Inquiry*, *5*(1), 31–47. http://www.jstor.org/stable/1342976

Dennett, D., (1987), ‘Fast Thinking’, in *The Intentional Stance*, Cambridge, MA: MIT Press, 324–337.

Dennett, Daniel C. (1996). *Kinds of Minds*. Basic Books.

Dennett, D., 1997, 'Consciousness in Humans and Robot Minds,' in M. Ito, Y. Miyashita and E.T.Rolls (eds.), *Cognition, computation, and consciousness*, New York: Oxford University Press, pp.17– 29.

Dretske, F. (1981). Knowledge and the Flow of Information, Cambridge, Mass.: The MIT Press.

Dretske, F. (1995), Naturalizing the Mind, Cambridge, Mass.: The MIT Press.

Dreyfus, H. , What Computers Can't Do: The Limits of Artificial Intelligence (Revised Edition), Harper and Row, New York, 1979.

Edelman, Shimon (2008), "*Computing the Mind*", p.28 Oxford University Press

Fodor, Jerry (2001) "Language, Thought and Compositionality" Mind & Language, Vol. 16 No. 1 February 2001, pp. 1–15. Blackwell Publishers Ltd.

Fodor, Jerry (1968). "Psychological Explanation", New York: Random House.

Fodor, Jerry (1975). "The Language of Thought", New York: Crowell.

Fodor, J.A. (1998), Concepts, Oxford: Oxford University Press.

Frank Jackson, (1982). Epiphenomenal Qualia. *The Philosophical Quarterly (1950-)*, *32*(127), 127–136. https://doi.org/10.2307/2960077

Gallagher, Shaun (2006), *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition*. Co-edited with W. Banks and S. Pockett (Cambridge, MA: MIT Press, 2006)

Gray, Richard (2014). Pain, perception and the sensory modalities: revisiting the intensive theory. Review of Philosophy and Psychology 5 (1) , pp. 87-101. 10.1007/s13164-014-0177-4 file

Haugeland, John (1985) "Artificial Intelligence: The Very Idea", second printing 1986, The MIT Press.

Holland, Owen 2003 Exploration and high adventure: the legacy of Grey Walter*Phil. Trans. R. Soc. A*.**361**2085–2121

Hutto, Daniel and Myin, Erik (2013) *Radicalizing Enactivism: Basic Minds without Content* (with Myin, E.). Cambridge, MA: The MIT Press, Bradford Books. (2013)

Jane A. Smith, Ph.D., A Question of Pain in Invertebrates,ILAR Journal, Volume 33, Issue 1-2, 1991, Pages 25–31, https://doi.org/10.1093/ilar.33.1-2.25

Kim, Jaegwon (1989). “Mechanism, Purpose, and Explanatory Exclusion”, in J. Kim, Supervenience and Mind, Cambridge, Cambridge University Press.

Kim, Jaegwon (1998) “Mind in a Physical World”, Cambridge, MA: Bradford.

Kriegel, U (2013), Two Notions of Mental Representation: Current Controversies in Philosophy of Mind (pp. 161-179), https://philpapers.org/archive/kritno-3.pdf

Lagercrantz, H., Changeux, JP. The Emergence of Human Consciousness: From Fetal to Neonatal Life. Pediatr Res 65, 255–260 (2009). https://doi.org/10.1203/PDR.0b013e3181973b0d

Lewis, David (1980). Mad pain and Martian pain. In Ned Block (ed.), *Readings in the Philosophy of Psychology*. Harvard University Press. pp. 216-222.

Ludwig, Kirk (1992), Brains in a Vat, Subjectivity, and the Causal Theory of Reference: The Journal of Philosophical Research XVII (1992): 313-345, https://socrates.sitehost.iu.edu/papers/biv.pdf

Maturana, H and Varela, F. (1998) cited in Signorelli CM (2018) Can Computers Become Conscious and Overcome Humans? Front. Robot. AI 5:121.doi: 10.3389/frobt.2018.00121

Meincke, Anne Sophie. (2018). Bio-Agency and the Possibility of Artificial Agents. 10.1007/978-3-319-72577-2_5.

McMurbay, Gordon A. (1953) “Congenital Insensitivity to pain and its implications for Motivational theory”, The University of Saskatchewan, CANAD. PSYCHOL., 1955, 9 (2)

Mowber, O. H. (1950) “Learning theory and personality dynamics”. New York: Ronald, 1950, was cited by GORDON A. McMURBAY (1955) “Congenital insensitivity to pain and its implications for motivational theory”, University of Saskatchewan

Nagel, T. (1 974) . \'vl1at is it like to be a bat? Philosophica.l Review 83: 435-450.

Nanay, Bence (2012) Success semantics: the sequel, Philo Stud (2013) 165:151-165

Newell, Allen; Simon, H.A. (1963), "GPS: A Program that Simulates Human Thought", in

Feigenbaum, E.A.; Feldman, J. (eds.), *Computers and Thought*, New York: McGraw-Hill

Newell, Allen; Simon, H. A. (1976), "Computer Science as Empirical Inquiry: Symbols and Search", Communications of the ACM, 19 (3): 113–126, doi:10.1145/360018.360022

O'Brien, Gerard (1999) “Connectionism, Analogicity and Mental Content”, Acta Analytica Vol. 22 (1999): pp.111-31

O'Brien, Gerard and Opie, Jon (2008) “The role of representation in computation” , Published online: 18 September 2008

Putnam, Hilary. (1975).” The Meaning of "Meaning" “.University of Minnesota Press, Minneapolis. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/185225.

Putnam, H. (1980). The nature of mental states. Readings in philosophy of psychology, 1, 223-231.

Putnam, Hilary (1988) “Representation and Reality”, Cambridge, MA: MIT Press.

Prinz, Jesse (2012). *The Conscious Brain: How Attention Engenders Experience*. Oup Usa.

Rapaport, William, J (2007), “Searle on Brains as Computers”, State University of New York at

Buffalo, Buffalo, NY 14260-2000, http://www.cse.buffalo.edu/~rapaport/

Ravenscroft, I. (2005). *Philosophy of mind: A beginner's guide*. Oxford University Press, USA.

ROBERT KIRK, Sentience and Behaviour, ***Mind***, Volume LXXXIII, Issue 329, January 1974, Pages 43–60, https://doi.org/10.1093/mind/LXXXIII.329.43

Rosenthal, David (1993) "Higher-Order Thoughts and the Appendage Theory of Consciousness", *Philosophical Psychology* VI, 2 (June 1993): 155-167.

Rupert, Rob. (2008). Causal Theories of Mental Content. Philosophy Compass. 3. 353 - 380.

10.1111/j.1747-9991.2008.00130.x.

Shapiro, Lawrence and Shannon Spaulding, "Embodied Cognition",*The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>

Searle, John (1980). “Minds, Brains and Programs”,Behavioral and Brain Sciences, 3: 417–457

Searle, John (1990) , “Is the Brain a Digital Computer?”,Proceedings and Addresses of the American Philosophical Association, 64: 21–37.

Searle, John (1992) “The Rediscovery of Mind”, Cambridge, MA: MIT Press.

Searle, John (2002) “Why I am not Property Dualist” Journal of Consciousness Studies, 9(12), 57–64.

Shea, Nicholas (2018. Representation in Cognitive Science, Oxford University Press

T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, “Introduction to Algorithms,” 3rd Edition, The MIT Press, Cambridge, 2009.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein: Introduction to Algorithms, 3rd Edition.MIT Press 2009, ISBN 978-0-262-03384-8, pp. I-XIX, 1-1292

Turing, Alan(1936) “On Computable numbers, with an application to the Entscheidungs problem”

Turing, Alan (1937) “Computability and λ-Definability”, The Journal of Symbolic Logic, Vol. 2, No. 4 (Dec., 1937), pp. 153-163, Published by: Association for Symbolic Logic

Turing, Alan (1950) ”Computing Machinery and Intelligence”, *Mind*, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind.* The MIT Press.

Tye, Michael (2000). *Consciousness, Color, and Content*. MIT Press.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. Brain, 97, 709–728.

Wilson, Robert A (2014) " What Computations (Still, Still) Can't Do: Jerry Fodor on Computation and Modularity"New Essays in Philosophy of Language and Mind. Supplementary issue 30 of the Canadian Journal of Philosophy.

Whyte, J. T. 1990. Success Semantics. Analysis 50(3): 149-157.

Whyte, J. T. 1997. Success Again: Replies to Brandom and Godfrey-Smith. Analysis 57(1): 84-88