Quantifying Suboptimal Automation Use

By

Megan L. Bartlett

Bachelor of Psychology (Honours)

Thesis

Submitted to Flinders University

for the degree of

Doctor of Philosophy

College of Education, Psychology & Social Work

16th July 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Advances in sensors and information processing algorithms offer the prospect of powerful automated decision aids to assist human operators in fields such as transportation security, military operations, and medical diagnosis. A decision aid can improve human performance, however, only if the user acts appropriately on its advice. The current thesis investigated elements of automation interface design, individual differences, decision strategies, and training protocols that shape human-automation interaction. The thesis comprises four papers (two published, one under review, and one in preparation).

The first study investigated the differential effects of automation misses and false alarms on operator behavior, asking whether they reflect a tendency for operators to prefer automation whose response bias matches their own. Participants performed a simulated baggage screening task either unassisted or with assistance from a 95%-reliable automated decision aid. The response bias of the automation and the participant were manipulated orthogonally. Contrary to earlier findings, data gave no evidence that false alarms from the aid compromised human performance more than misses did, suggesting that this effect might be less robust than earlier evidence suggested.

The second study measured the efficiency of operators' automation use, comparing automation-aided performance to the predictions of various statistical models of collaborative decision making. Across three experiments, participants performed a binary signal detection task either unassisted, or with assistance from a 93%-reliable automated decision aid. As anticipated, assistance from the aid improved discrimination performance. However, aided performance was consistently and highly suboptimal, hewing closest to the predictions of some of the least efficient collaborative models. Performance was similar whether the aid provided binary cues or more informative graded judgments.

The third study investigated whether manipulating the format in which the aid's cues were rendered would improve automation-aided performance. Across three experiments, participants performed the same binary signal detection task as in the second study, with judgments from the aid

rendered either as raw signal levels, confidence ratings, likelihood ratios, verbal descriptors, or verbal-spatial descriptors. While assistance from the aid again improved participants' discrimination performance, aided performance remained highly inefficient regardless of the format of the aid's judgments.

The final study sought to generalize the above effects across varying levels of aid reliability. Participants performed the same task as in the second and third studies, but with an aid of either 60%, 85%, or 96% reliability. Assistance from an 85% or 96%-reliable decision aid, relative to assistance from a 60%- reliable aid improved automation-aided sensitivity. Interestingly, automation-aided sensitivity approached optimal levels as aid reliability level decreased, but only because the asymptotically optimal strategy as the aid's reliability goes to zero is to ignore the aid's judgments.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signed Megan Bartlett

Date 16-07-2018

LIST OF MANUSCRIPTS AND PUBLICATIONS

Research from this thesis has been published in the following articles:

Bartlett, M. L., & McCarley, J. S. (2018). No tendency for human operators to agree with automation whose response bias matches their own. *International Journal of Human Factors and Ergonomics, 5(2),* 111-128. Copyright © [2018] (Inderscience). doi: https://doi.org/10.1504/IJHFE.2018.092227

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors, 59(6),* 881-900. Copyright © [2017] (Human Factors). Reprinted by permission of SAGE Publications. doi: http://dx.doi.org/10.1177/0018720817700258

Research from this thesis is currently under review in the following article:

Bartlett, M. L., & McCarley, J. S. (2017). *No effect of cue format on automation dependence in an aided signal detection task.* Manuscript submitted for publication.

Research from this thesis is also currently under preparation in the following article:

Bartlett, M. L., & McCarley, J. S. (2017). *Ironic efficiency.* Manuscript in preparation.

LIST OF CONFERENCE ABSTRACTS AND PRESENTATIONS

In addition, research from this thesis has been presented at:

Bartlett, M. L., & McCarley, J. S. (2017. November). *Quantifying suboptimal automation use*. Paper presented at the Defence Human Sciences Symposium, Adelaide, Australia.

Bartlett, M. L., & McCarley, J. S. (2017. November). *No effect of information format on performance in an aided signal detection task*. Poster presented at the Australasian Cognitive Neuroscience Society Conference, Adelaide, Australia.

Bartlett, M. L., & McCarley, J. S. (2014. November). *Investigating ideal operator and aid response bias in a simulated baggage screening task*. Paper presented at the Defence Human Sciences Symposium, Adelaide, Australia.

Bartlett, M. L., & McCarley, J. S. (2014. September). *Human interactions with automated decision aids*. Colloquium presented at the Psychology Seminar Series, Flinders University, Adelaide, Australia.

McCarley, J. S., & Bartlett, M. L. (2015. November). *Benchmarking automation-aided decisions.* Colloquium presented at the Applied Cognitive Science and Human Factors Forum, Michigan Tech, Houghton, United States.

DEDICATION

This thesis is dedicated to my beautiful parents, Barbara and Michael Bartlett.
You are my strength and inspiration.

ACKNOWLEDGMENTS

Professor Jason McCarley. Thank you for your time, commitment and support of me throughout my PhD. It's been an honour getting to know you and working with you over the past six years. It's a privilege to be able to call you my supervisor, mentor and friend.

Professor Mike Nicholls and Dr Michelle Arnold. Thank you for being part of my Proposal Committee and for providing me with excellent feedback and ideas. I enjoyed sharing my research with you.

My Parents, Barbara and Michael Bartlett. Thank you for your unfailing belief in me and your unconditional love and support. You mean the world to me. I am so proud to be your daughter.

To **Kingsley Fletcher and Susan Cockshell**, it has been a pleasure and honour getting to know you and work with you over the last six years.

To my fellow PhD cohorts, office mates (past and present), School of Psychology staff and Perceptual & Cognitive Performance Laboratory members (past and present), it's been a pleasure sharing this journey with you.

I am also grateful for the financial assistance that I have received that assisted me to undertake my thesis. This support has been provided to me in the form of a Flinders University Research Scholarship (FURS), Defence Science and Technology Group (DSTG) Top-Up Scholarship and Psychology Department Write-up Scholarship. Thank you also to the School of Psychology for financial support that enabled me to conduct my research and to disseminate my findings at conferences.

Finally, thank you to all who participated in my studies. Without you, this thesis would not have been possible.

CHAPTER 1:
LITERATURE REVIEW

Automation has been described as, “the execution by a machine agent (usually a computer) of a function that was previously carried out by a human” (Parasuraman & Riley, 1997, p. 231). It increasingly pervades our day-to-day life, as our shoppers (Qiu & Benbasat, 2010), assistants (Walter et al., 2014), or even companions (Breazeal, 2002). Automated decision aids, more particularly, assist in gathering, transforming, or interpreting information (Lee & See, 2004), helping the human operator to perform tasks that might otherwise be complex, time-consuming, or hazardous (Wickens, Lee, Liu, & Becker, 2004). They are typically better equipped than humans to multitask, respond to events rapidly, monitor and store important information, and ignore irrelevant information (Fitts, 1951; Fuld, 2000; Sheridan, 2002).

Familiar decision aids, such as smoke detectors, alert the human operator to potentially hazardous conditions (Meyer & Bitan, 2002). Less familiar aids may help the operator identify threatening objects in passenger luggage (Nercessian, Panetta, & Agaian, 2008), spot enemies on the battlefield (Bhanu, 1986), detect the presence of a tumour in a mammogram (Nishikawa, 2007), or recognize an impending automobile collision (Zhang, Antonsson, & Grote, 2006). But even with assistance of an aid, the human still plays a pivotal role in decision making and should not be viewed as merely a passive observer (McDaniel, 1988). Humans are adaptable and flexible, and better able to cope with unpredictability than are computers (Parasuraman & Riley, 1997). They are also better able to detect and interpret many sensory stimuli, and are able to use deduction, reasoning, and judgment (Fitts, 1951; Fuld, 2000; Sheridan, 2002). To be effective, automation should therefore be ‘human-centred’, supporting and cooperating with the human operator (Billings, 1996, 1997; Wickens, Maver, Parasuraman, & McGee, 1998; Woods, 1996). Wickens et al. (2004) argue that ideally, automation will inform operators (e.g., providing a pilot with all the information

required to fly straight and on course), train them (e.g., preparing a pilot for the changing demands of the automation; Lee & Sanquist, 2000; Zuboff, 1988), offer them flexibility (e.g., allowing the driver to choose whether or not to employ the cruise control feature in a car), and respond adaptively to their circumstances (e.g., increasing automation in response to high operator cognitive or psychophysical workload; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2000).

Automation Taxonomy

Sheridan (1980) ranked automated systems along a continuum of ten levels, shown in Figure 1-1. At the low end of the continuum, a task is fully non-automated: the human operator retains full control and all decision-making responsibility. At the high end, the aid assumes full control from the human and takes responsibility for all actions, functions and decisions (Sheridan, 1992; Wickens et al., 1998).

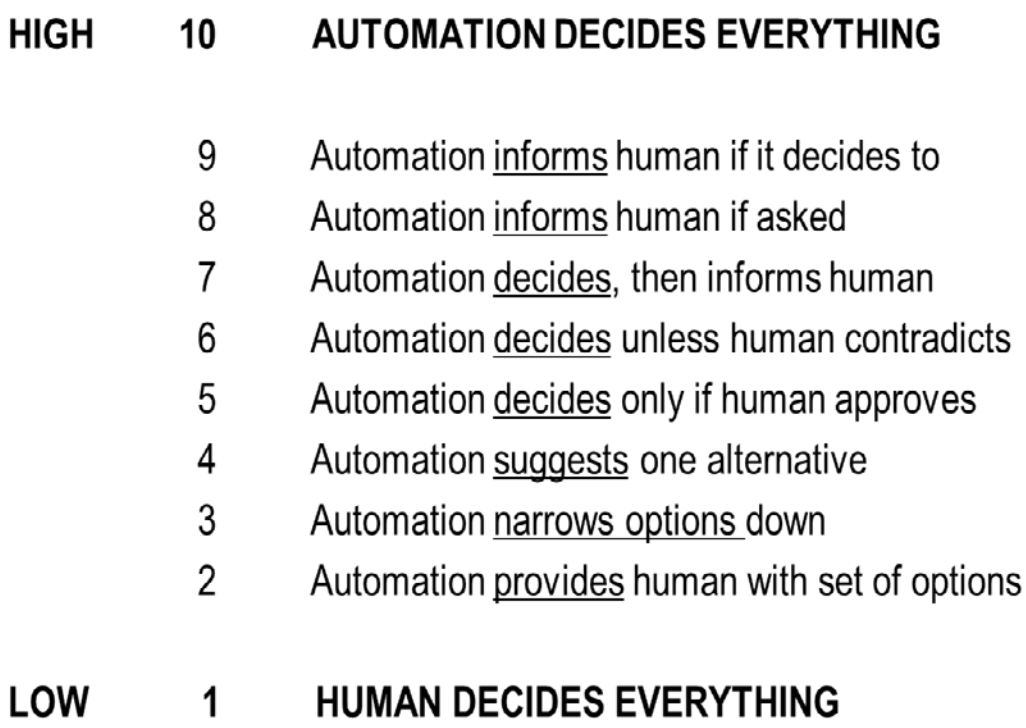

Figure 1-1. The continuum proposed by Sheridan (1980).

For example, at level 4, the automation suggests a response to the human operator, but has no further involvement in the response selection and execution (Parasuraman, Sheridan, & Wickens, 2000). Practically, this might involve a conflict detection aid suggesting that a pilot adopt a particular course of action to avoid an impending collision. At level 6, however, the automation would take action if the human operator has not made a decision within a certain amount of time (Parasuraman et al., 2000). Practically, this might involve the conflict detection aid altering course on its own to avoid the collision.

Parasuraman and colleagues (2000) provide a similar taxonomy of the level of automation (*None*, *Low*, *Medium*, *High*, *Full*), but also classify automation on an orthogonal dimension of function. Their taxonomy distinguishes four stages at which an aid might function:

Stage 1: Information Acquisition

Information acquisition by an automated aid entails the collection of raw data, akin to the basic stage of attentional selection in human information processing (Parasuraman, 2000; Parasuraman et al., 2000).

Stage 2: Information Analysis

At the stage of information analysis, an automated aid configures information to make it easy for the human operator to interpret.

Stage 3: Decision Selection

At the stage of decision selection, the costs and benefits of different decisions and their outcomes are weighed up by the automation and transformed to a discrete choice.

Stage 4: Action Implementation

Finally, at the stage of action implementation, the automated decision aid initiates or executes the chosen response option (Parasuraman et al., 2000).

Signal Detection Theory

Many automated decision aids can be thought of as performing a standard signal detection task, discriminating between two or more distinct categories or states of the world, generally termed *signal* and *noise* (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). For example, in a simple yes/no task where participants are required to determine the presence or absence of a threat in passenger luggage, a threat will be present on some trials (signal trials) and absent on others (noise trials).

A signal detection aid can thus produce one of four possible decision outcomes on a given trial: a hit (correctly reporting the presence of a signal, e.g., knife is present in the luggage and screener says it is present), a correct rejection (CR: correctly reporting that no signal is present, e.g., knife is absent in the luggage and screener says it is absent), a false alarm (FA: incorrectly reporting the presence of a signal, e.g., knife is absent in the luggage but screener says it is present), or a miss (incorrectly reporting that no signal is present, e.g., knife is present in the luggage but screener says it is absent) (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999).

Ideally, an automated signal detection aid would produce only correct judgments, committing no false alarms or misses. Unfortunately, limitations on the quality of the data with which the aid operates often make this impossible. Because an aid, like a human decision maker, makes decisions based on probabilistic data, it will unavoidably make mistakes. The balance that the aid strikes between the two potential types of errors is determined by the aid's response criterion placement: *unbiased*, *liberal* or *conservative*. Signal detection theory (SDT) assumes that the evidence can be represented as a scalar *decision variable*, and that the decision variable is stochastic and distributed continuously (Pastore, Crawley, Berens, & Skelly, 2003). Figure 1-2 presents a depiction of the evidence distributions in signal detection theory. The figure assumes a standard equal-variance Gaussian model. The left curve represents noise trials, and the right curve represents

signal trials. The risk of misjudgement arises when the evidence distributions corresponding to signal and noise states overlap. Because values in the regions of overlap do not distinguish unambiguously between signal and noise, a decision rule is necessary to transform a sampled value of the decision variable into a discrete choice. The model assumes that the decision maker reaches a discrete decision by comparing the decision variable to a pre-established response criterion. A positive judgment (i.e., 'signal') results when the decision variable exceeds the criterion value, and a negative judgment results when the decision variable falls below the criterion value.

A decision maker—human or electronic— operating under an unbiased criterion (see criterion placement in Figure 1-2) will have a false alarm rate that is matched to their miss rate. An individual or automated aid operating under a liberal criterion, however, (moved to the left in Figure 1-2) is biased towards 'yes' responding, and will show an FA rate greater than their miss rate (Stanislaw & Todorov, 1999). Conversely, an individual or automated aid operating under a conservative criterion (moved to the right in Figure 1-2) is biased towards responding 'no', and will show an FA rate less than their miss rate (Stanislaw & Todorov, 1999).

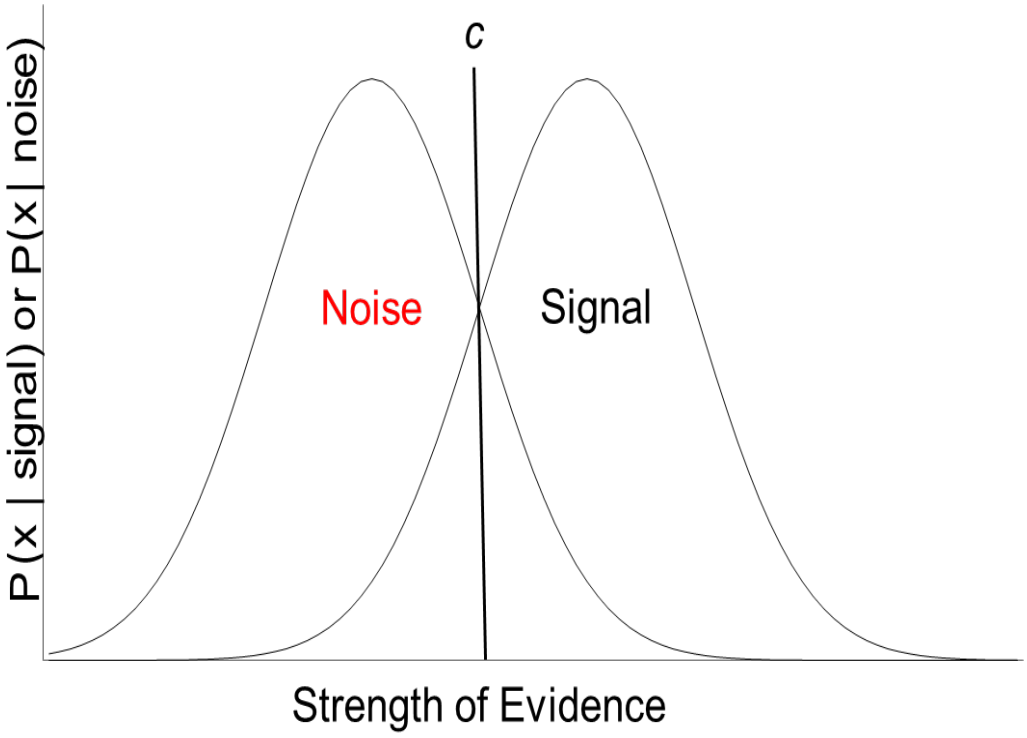

Figure 1-2. Depiction of the evidence distributions in a standard equal-variance Gaussian signal detection model. The vertical black line represents the criterion setting (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999).

Automation designers, and sometimes system operators (Botzer, Meyer, Bak, & Parmet, 2010; Meyer & Sheridan, 2017), are able to set the criterion of the automated aid, seeking an optimal trade-off between misses and false alarms (Dixon & Wickens, 2006; Rice & McCarley, 2011). The ideal setting of the aid's bias depends upon the frequency of signal and noise events as well as the benefits or payoffs attached to various decision outcomes (i.e., hits, correct rejections, false alarms, misses). SDT transforms a respondent's target detection rate and false alarm rate data into measures of sensitivity, the respondent's ability to discriminate signal from noise, and response bias, the respondent's general willingness to respond 'signal present' or 'signal absent' as determined by criterion placement.

Sensitivity

Sensitivity refers to the ability to successfully discriminate signal from noise (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). Sensitivity is conventionally measured by *d'*, which reflects the distance, or degree of overlap between the signal and noise curves in an equal-variance Gaussian model (Stanislaw & Todorov, 1999). *d'* is measured in standard deviations of the curves. A *d'* of 5 or greater signifies near-perfect discrimination, whereas a value of 0 represents performance at chance level (Stanislaw & Todorov, 1999). The formula for *d'* is as follows,

$$d' = z(HR) - z(FAR),$$

where z indicates the inverse normal transformation, *HR* indicates the human operator's hit rate, or the probability of responding signal present on a signal trial, and *FAR* indicates the human operator's false alarm rate, or the probability of responding signal present on a signal absent trial. *d'* assumes equal-variance Gaussian evidence curves. SDT also provides statistically optimal benchmarks of sensitivity against which the human respondent's data can be benchmarked (Green & Swets, 1966; Tanner & Birdsall, 1958).

Response Bias

Response bias, or criterion placement, reflects the amount of evidence needed in order for an individual to respond 'yes' (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). As noted, the decision maker can operate with an unbiased, liberal, or conservative criterion (Rice & McCarley, 2011). Though there are multiple indices of bias, the statistic *c*, which measures the position of the criterion relative to the point of unbiasedness, is perhaps most robust (Macmillan & Creelman, 1990; See, Warm, Dember, & Howe, 1997). The formula for *c* is as follows,

$$c = -0.5 \times [z(HR) - z(FAR)].$$

A value of 0 denotes unbiasedness, negative values denote liberal bias, and positive values denote conservative bias. Like *d'*, *c* assumes equal-variance Gaussian evidence curves, and is measured in units of the standard deviation of the evidence curves. SDT also provides statistically optimal benchmarks of response bias against which the human respondent's data can be benchmarked (Macmillan & Creelman, 2005; Wickens, 2002; Wickens & Hollands, 2000).

Suboptimal Automation Use

While automated aids can improve human performance (Maltz & Meyer, 2001; Meyer, 2001; Wickens & Dixon, 2007), an aid is only useful if the human operator trusts it and acts appropriately on its advice (Wickens & Dixon, 2007). Unfortunately, operators are prone to either of two forms of error when interacting with automated aids, either relying too heavily on the aid, *misuse*, or relying too little on it, *disuse* (Parasuraman & Riley, 1997). Misuse and disuse compromise the benefits of automated assistance, and in extreme cases can even increase the risk of human error relative to unaided performance (e.g., Alberdi, Povyakalo, Strigini, & Ayton, 2004).

Misuse

Misuse occurs when an individual chooses to depend on an imperfectly reliable automated decision aid more than is optimal (Lee, 2008; Lee & See, 2004). An individual misusing an automated aid may monitor the system inadequately, leading to a phenomenon known as *automation-induced complacency* (Bahner, Hüper, & Manzey, 2008; Molloy & Parasuraman, 1996; Parasuraman & Manzey, 2010; Parasuraman, Molloy, & Singh, 1993; Skitka, Mosier, & Burdick, 2000). Complacency has been linked to real-world incidents and accidents in domains such as aviation (Hurst & Hurst, 1982; Wiener, 1981) and shipping (Degani, 2001; Parasuraman & Riley, 1997).

Unfortunately, the consequences of misuse can be severe (Parasuraman et al., 1993; Parasuraman, Mouloua, Molloy, & Hilburn, 1996). A pilot, for example, who misuses a flight

management aid may cease to be vigilant for automation failures (Bainbridge, 1983) and thus be slower to detect a failure when it arises. Riley (1994), found evidence for such misuse of automation in a dual-task study of pilots. In this case, instead of recognizing that an automated system had begun to fail and abandoning its use, the pilots continued to rely on the system, even when their performance degraded as a result.

More insidiously, misuse may lead to deskilling (Wiener, 1988), causing even stronger reliance on the automation (Lee & Moray, 1992), and may compromise situation awareness (Chen, Visser, Huf, & Loft, 2017; Endsley & Kiris, 1995).

Disuse

Conversely, disuse occurs when an individual chooses to depend on an imperfectly reliable automated decision aid too little (Lee, 2008). An individual may commit disuse because he or she lacks trust in the aid (Lee, 2006; Lee & Moray, 1992; Liu, Fuld, & Wickens, 1993; Muir, 1987; Rice & McCarley, 2011), is overconfident in his or her own ability (Dzindolet, Pierce, Beck, & Dawe, 2002), or simply has an aversion to relying on a mechanical or statistical decision process (Dietvorst, Simmons, & Massey, 2015).

Disuse results in suboptimal performance through multiple mechanisms, such as ignoring highly reliable automated advice, and operator fatigue. Take the example of a conflict detection aid, warning pilots of impending in-air collisions, mentioned previously. What might cause the pilot in the example to commit disuse? Disuse of the aid in the example could either result from the pilot under-weighting the aid's advice and/or ignoring perfectly reliable advice from the aid, or could instead result from the pilot becoming fatigued and performing slower or less accurately, than with assistance from the aid (Parasuraman et al., 2000).

Compliance vs. Reliance

In an elaboration on the misuse/disuse dichotomy, Meyer (2001) identified qualitative differences in operators' responses to positive and negative judgments from an aid. *Compliance* refers to the operator's tendency to act on a 'signal present' judgment from the aid, whereas *reliance* instead refers to the operator's tendency to act on a 'signal absent' judgment from the aid (Meyer, 2001). Evidence for the compliance/reliance distinction come from the finding that automation misses and false alarms have different effects on operator behavior. Misses from an aid adversely affect reliance, whereas false alarms from an aid adversely affect reliance in addition to compliance (Meyer, 2001).

Diagnosing Suboptimal Automation Use

Dependence on automation is diagnostic of whether a human operator is prone to misuse or disuse (Wang, Jamieson, & Hollands, 2008, 2009) and can be measured in various ways, such as by investigating the *consistency* or *correlation* between the operator's and the automation's judgments (e.g., Biros, Daly, & Gunsch, 2004; Bisantz & Pritchett, 2003; Brunswik, 1956; Murrell, 1977), the *performance* of the human operator when the automation provides them with correct and incorrect feedback (e.g., Maltz & Shinar, 2003; Parasuraman et al., 1993), the *behavioral patterns* of the human operator (e.g., Ezer, Fisk, & Rogers, 2007; Moray & Inagaki, 2000), or the *misuse and disuse rates* of the human operator (e.g., Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Parasuraman & Riley, 1997).

Wang et al. (2008) believe that the methods listed above fail to provide explicit benchmarks of appropriate automation dependence, and conclude that the best way to measure optimal dependence is to instead compare empirical bias with optimal bias under instances where the aid provides a binary judgment, (Maltz & Meyer, 2001; Meyer, 2001). Assuming a symmetrical payoff matrix, where the costs associated with incorrect judgments (i.e., misses and FAs), and the payoffs

associated with correct judgments (i.e., hits and CRs), are matched, optimal bias, as measured by the statistic beta (β^*), can be calculated as follows,

$$\beta^* = p(\text{no signal} \mid \text{diagnosis}) / p(\text{signal} \mid \text{diagnosis}),$$

where p(no signal | diagnosis) indicates the probability that a target is absent given the aid's diagnosis, and p(signal | diagnosis) is the complementary probability. In a simulated combat identification task, Wang et al. (2009) demonstrated that participants' empirical bias was less than optimal.

An alternative but related method to measuring optimal reliance is to compare empirical sensitivity to optimal sensitivity (Sorkin, Hays, & West, 2001). Optimal sensitivity, d'_{optimal}, can be calculated as follows,

$$d'_{\text{optimal}} = (d'^{2}_{\text{operator}} + d'^{2}_{\text{aid}})^{1/2},$$

where d'_{operator} indicates the sensitivity of the unaided human operator, and d'_{aid} indicates the sensitivity of the automated decision aid. Note, however, that optimal sensitivity, as defined by the equation above, assumes the aid and operator each make independent judgments under an equal-variance Gaussian model, and that the aid shares continuous evidence values directly with the operator (Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001). Under different constraints—for example, when the aid provides the operator a binary decision rather than directly sharing a continuous evidence value (Robinson & Sorkin, 1985)—optimal sensitivity will be lower.

Automation Usage

The degree to which an operator uses an automated aid is determined by the interaction of multiple variables, such as reliability and trust (Lee & Moray, 1992; Muir, 1988), attitudes (McClumpha & James, 1994), framing (Lacson, Wiegmann, & Madhavan, 2005; Madhavan & Wiegmann, 2007; Rice & McCarley, 2011), manipulation of information format (Botzer et al., 2010), workload (Riley, 1989), accountability (Skitka et al., 2000), and social processes (Dzindolet,

et al., 2002; Mosier & Skitka, 1996). Understanding these factors is critical to understanding how and why individuals use automated decision aids suboptimally.

Reliability

Humans tend to expect that technology will be accurate and rarely prone to errors, a belief described as the 'perfect automation' schema (Dzindolet et al., 2002). When asked to predict the number of errors a human aid or an automated decision aid would make in 200 trials of a visual detection task, for example, participants assumed the automated aid (24.79 errors) would be far superior to the human (51.26 errors), even though the descriptions they had been given of the aids, aside from the label 'human' versus 'computer', were identical (Dzindolet et al., 2002).

But despite the *a priori* confidence that users put in it, after exposure to an aid users often discover that it is imperfect. Automation often fails simply because the world is highly uncertain and some tasks, such as weather forecasting, are impossible to perform with perfect accuracy (Wickens, Thomas, & Young, 2000; Wickens et al., 2004). According to Wickens et al. (2004), an aid may also perform imperfectly if its components fail or if its design is flawed (Leveson, 1995), if it is incorrectly used (Lin, Vicente, & Doyle, 2001), or if it performs poorly under certain circumstances but is otherwise reliable. As a rough guide, an aid needs to be at least 70% reliable in its diagnoses in order to improve human performance (Rice & McCarley, 2011; Rovira, McGarry, & Parasuraman, 2007; Skitka, Mosier, & Burdick, 1999; Wickens & Dixon, 2007).

Trust

Trust is a subjective judgment of the extent to which a system can be relied upon to enhance performance under uncertain conditions (Lee & See, 2004). Trust is determined by the characteristics of both the automation (i.e., competence, responsibility, predictability, and dependability; Muir, 1987, 1994) and the human operator (i.e., personality, Merritt & Ilgen, 2008; age, Ho, Wheatley, & Scialfa, 2005; and mood, Merritt, 2011). Researchers believe that trust is

multi-faceted, subsuming a number of constructs. Sheridan (1988), for instance, states that trust can be defined in terms of reliability, robustness, familiarity, understandability, explication of intention, usefulness and dependence, while Muir and Moray (1996) believe trust can be defined by predictability, dependability, faith, competence, responsibility, and reliability.

The human operator's level of trust is directly related to the perceived reliability of an aid (de Vries, Midden, & Bowhuis, 2003; Lee & Moray, 1992; Merritt, 2011; Wang et al., 2009). Trust in an automated decision aid, furthermore largely determines whether the human operator uses the aid (Lee & Moray, 1994). In other words, humans are more likely to trust an aid that is perceived to perform well than one that is perceived to perform poorly, and they are therefore more likely to use it (Merritt & Ilgen, 2008). In a simulated aided pasteurization task, for example, Lee and Moray (1992, 1994) and Muir and Moray (1996) measured operators' trust in the automation when either the pump systems or automated controller erred. As expected, operator trust diminished following automation errors, and though it recovered when the automation returned to functioning well after an error, it failed to return to earlier levels. Use of automation was directly related to the difference between the operator's level of trust in the system and his or her own self-confidence. That is, operators who perceived themselves to be less reliable than the aid trusted the aid more and relied on its advice more strongly than operators who perceived themselves to be more reliable than the aid. Ideally, trust will be proportional to, or calibrated with, the reliability of the aid (Wickens, Gordon, & Liu, 1997). Unfortunately, in practice, trust is often not well calibrated with reliability.

Finally, the degree to which automation can be considered *transparent* has been shown to affect trust (e.g., Dadashi, Stedmon, & Pridmore, 2012; Seong & Bisantz, 2008; Wang et al., 2009). Transparent automation provides the operator with clear and valid information about its functioning (Seong & Bisantz, 2008). Trust can be bolstered, for example, if an automated system provides the human operator with information regarding its reliability (e.g., Wang et al., 2009), or provides

explanations as to why and how it might err (e.g., Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003).

Attitudes

Automation use may also be influenced by attitudes beyond trust. Attitudes, such as liking of automation, consist of cognitive (i.e., beliefs) and affective assessments (Zanna & Rempel, 1988). Attitudes toward automation vary across individuals (Helmreich, 1984; McClumpha & James, 1994), but unsurprisingly, are directly related to the extent to which the automation is perceived to be reliable (Parasuraman & Riley, 1997). Individuals may differ in their attitudes towards automation in general, or toward specific types of automation, such as collision avoidance systems (Parasuraman & Riley, 1997).

Research has also distinguished between explicit (e.g., conscious) and implicit attitudes (e.g., sub-conscious). Explicit attitudes towards automation are consciously accessible, and are formed when an individual actively tests the truth and/or falsity of the attitudes he or she holds (Merritt, Heimbaugh, LaChapell, & Lee, 2013). Thus, only explicit attitudes that are congruent with the beliefs of the individual are retained. Self-report measures are commonly used to investigate such attitudes (Merritt et al., 2013).

Implicit attitudes, conversely, are automatically formed in conjunction with an individual's concept of automation (Merritt et al., 2013). Implicit attitudes can be inferred from the speed with which individuals associate the subject (i.e., automation) with the corresponding attitudinal judgment (i.e., good or bad) in a paradigm known as an *implicit association test* (IAT) (Greenwald, McGhee, & Schwartz, 1998). Individuals who hold a positive attitude toward automation will typically be faster to associate the word 'automation' with a positive descriptor such as 'good' than with a negative descriptor such as 'bad' (Merritt et al., 2013). The formation of implicit attitudes largely occurs through conditioning or association, whereby an item is paired with either positive or

negative stimuli (Gawronski & Bodenhausen, 2006; Merritt et al., 2013; Olson & Fazio, 2001; Wilson, Lindsey, & Schooler, 2000). This conditioning largely occurs through the accumulation of personal or cultural experience or through exposure to media (Merritt et al., 2013).

Merritt and colleagues (2013) investigated the link between implicit and explicit attitudes on trust in a simulated baggage screening task, where participants were assisted by either a perfectly reliable automated decision aid (the clearly good condition), or an imperfectly reliable aid that committed either obvious (the clearly poor condition) or non-obvious errors (the ambiguous condition). Explicit attitudes towards automation were assessed via a self-report measure investigating participants' propensity to trust machines, while implicit attitudes towards automation were instead assessed via the IAT. Results showed that when participants were assisted by an aid that committed non-obvious errors, implicit attitudes combined additively with explicit propensity to trust to predict actual trust. In contrast, when participants were assisted by an aid that committed obvious errors, an interaction between implicit attitudes and explicit propensity to trust was found, such that only participants who held both positive implicit and explicit attitudes also reported greater trust. No such relationships, however, were found when participants were assisted by a perfectly reliable decision aid. Findings in total suggest that trust in automation may be determined in part by unconscious mechanisms (Merritt et al., 2013).

Framing

A manipulation of *framing* changes the description of a problem without changing its information content (Tversky & Kahneman, 1986). Multiple framings of the same problem are, by definition, logically equivalent, but they may nonetheless influence "the decision-maker's conception of the acts, outcomes and contingencies associated with a particular choice" (Tversky & Kahneman, 1981, p.453), changing the decisions maker's judgments. Very often, the effects of framing are demonstrated by presenting decision makers with mathematically equivalent problems

described in terms of either losses or gains (Tversky & Kahneman, 1981). Previous work has shown that the framing of an automated aid's cues can affect automation dependence and usage. Automation use, for example, may vary depending on whether an operator is told they will be assisted by an inexperienced automated aid or inexperienced human assistant (e.g., Madhavan & Wiegmann, 2007), or by an aid described as either 20% unreliable rather than 80% reliable (e.g., Lacson et al., 2005).

Rice and McCarley (2011) investigated the framing of trial-by-trial cues from an aid on performance. Participants performing a simulated baggage x-ray screening task were assisted by a 65%-reliable automated decision aid prone to either misses or false alarms. Additionally, some participants received signal present/absent diagnoses each trial regardless of the aid's certainty (explicit error framing), or conversely only received correct signal present/absent diagnoses each trial (neutral error framing). Results in general indicated that neutrally framed automation diagnoses, as compared to explicitly framed ones appeared to benefit performance. This effect was especially true for participants assisted by an aid prone to false alarms (Rice & McCarley, 2011)

Manipulation of Information Format

Manipulating the format in which information is presented to a decision maker has been shown to influence participants' reasoning and decision making (e.g., Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000), and can likewise alter automation usage. For example, Botzer et al. (2010) had participants perform a simulated fault detection task with assistance from an automated aid. Participants were allowed to make adjustments to the aid's response threshold, but the probability of a signal and the payoffs associated with different responses varied between blocks of trials, changing the optimal criterion setting. To help them make optimal threshold adjustments, participants received information regarding either the aid's *predictive value* (i.e., the probability of a fault conditioned on the aid's diagnosis of a fault) or its

diagnostic value (i.e., the probability of the aid diagnosing a fault, conditioned on the presence of a fault). In accordance with their predictions, Botzer et al. (2010) found that predictive values made participants more responsive to variations of signal probability, while diagnostic values made them more responsive to variations of decision payoff.

Workload

The cognitive workload of the human operator, that is, the mental exertion required to carry out a task (Bailey, Scerbo, Freeman, Mikulka, & Scott, 2006; Edwards, 1977; Kirlik, 1993; Parasuraman & Riley, 1997; Wong & Seet, 2017; Woods, 1995), is another factor affecting automation use. The potential for automated decision aids to decrease the costs associated with high cognitive workload, may lead to increased use of such aids (Parasuraman et al., 1993). However, the use of automated aids can also unnecessarily increase the human operator's workload, or create varying levels of workload over time. This is largely due to the fact that the need to initiate and engage aids can increase the level of task complexity (Kirlik, 1993), both mental and physical (Wickens, 1992; Wiener, 1985). Automated systems that actually increase operator workload are commonly described as *clumsy* (Wiener, 1988). Increased cognitive workload, in an automated context, has been shown to negatively impact task performance by increasing the time taken to complete a task and increasing the number of errors made (Adams, 2009).

Accountability

Accountability may lead to more optimal human-automation collaboration, as research has shown that making decision makers accountable for their decisions can reduce many cognitive biases (e.g., Simonson & Nye, 1992; Tetlock & Kim, 1987). In a simulated automation-aided flight tracking task, Skitka et al. (2000) investigated whether making participants accountable for their performance would reduce the number of errors participants made. In contrast to non-accountable participants, accountable participants were advised that their performance would be evaluated, that

they would be required to justify their decisions post-experiment, and that they should aim to maximise either their overall or specific task performance (Skitka et al., 2000). Results demonstrated that participants who were made accountable for their overall performance displayed more optimal decision making as compared to either non-accountable participants, or participants who were only made accountable for specific task performance. Participants who were made accountable for their performance were more attentive and vigilant in seeking out information to verify the automated aid's judgment before carrying out a recommended action, and therefore made fewer errors.

Social Processes

Finally, social processes can also influence automation use. Individuals working in human teams, for instance, often fall victim to *social loafing*, a phenomenon where individuals are less inclined to contribute in team, as compared to individual contexts (Latane, Williams, & Harkins, 1979). Individuals working with assistance from automation can likewise be thought of as working in a team, whereby one member of the team is a computer (Bowers, Oser, Salas, & Cannon-Bowers, 1996; Scerbo, 1996; Woods, 1996). Such automated teams may also fall victim to a similar tendency, termed *diffusion of responsibility*, whereby the responsibility for actions and decisions in an automated team is distributed between team members (Dzindolet, et al., 2002; Mosier & Skitka, 1996). Diffusion of responsibility in an automation context may result in misuse, especially when a decision aid is more reliable than the human operator. For instance, if a human operator feels dispensable, and that their individual contribution is not valued, they will put in less individual effort, which may cause them to rely more heavily on the automation (Dzindolet et al., 2002).

Current Aims

Research on the psychology of human-automation interaction is necessary to ensure that aids are useful to the human operator and are designed to encourage optimal behaviour (Parasuraman, 2000). The current thesis investigated human interaction with an automated aid in a naturalistic visual search task and a two-alternative forced choice (2AFC) task. A series of four studies investigated elements of automation interface design, individual user differences, and user training protocols that shape human-automation interaction.

The current thesis had four main aims: firstly, to explore and seek to explain earlier findings that different forms of automation error (i.e., false negative and false positive detections) produce asymmetrical effects on the user's willingness to trust the automation in a simulated baggage screening task, secondly, to better understand operators' inefficient use of decision aids in a 2AFC task by comparing participants' automation-aided performance levels to the predictions of a number of statistical models representing a variety of potential decision-making strategies, spanning a range of performance levels from highly efficient to inefficient, thirdly, to determine whether automation-aided performance can be bolstered by manipulating the format of the aid's cues, making the aid's binary judgments easier for users to interpret, and finally, to determine whether these results can be generalized across aids of varying reliability.

CHAPTER 2:
STUDY 1

The following manuscript entitled, *No tendency for human operators to agree with automation whose response bias matches their own*, was published in the International Journal of Human Factors and Ergonomics (IJHFE). Inderscience retains copyright of the original article. The version of the manuscript presented here is the final, peer reviewed version, prior to the publisher applying their formatting. This manuscript has been published as:

Bartlett, M. L., & McCarley, J. S. (2018). No tendency for human operators to agree with automation whose response bias matches their own. *International Journal of Human Factors and Ergonomics, 5(2),* 111-128. Copyright © [2018] (Inderscience).

Link to authoritative document (doi): https://doi.org/10.1504/IJHFE.2018.092227

All authors were involved in the formulation of the study concept and design, and data analysis. Megan Bartlett collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

In addition, the findings from this experiment have been presented at:

Bartlett, M. L., & McCarley, J. S. (2014. November). *Investigating ideal operator and aid response bias in a simulated baggage screening task.* Paper presented at the Defence Human Sciences Symposium, Adelaide, Australia.

Introduction

Automated decision aids assist human operators in fields including transportation security screening (e.g., detecting explosives in passenger baggage; Hudson et al., 2012; Wells & Bradley, 2012), military operations (e.g., identifying enemies on the battlefield; Wang, Jamieson, & Hollands, 2009), and medical diagnosis (e.g., detecting cancerous masses in a mammograph; Gilbert et al., 2008). Unfortunately, because they operate on probabilistic data, such aids rarely make judgments with perfect accuracy (Wickens, Thomas, & Young, 2000). Even highly sophisticated algorithms for detecting abnormalities in mammographic images, for example, may overlook a significant number of masses while producing non-negligible numbers of false positive responses (e.g., Jen & Yu, 2015; Liu & Zeng, 2015; Pereira, Ramos, & do Nascimento, 2014). Likewise, automated systems for explosives detection in aviation security, even when designed to rigorous technical standards, are unlikely to achieve perfect performance in operational environments (Hudson et al., 2012; Wells & Bradley, 2012).

The most direct and obvious consequence of imperfect automation performance is that the aid will occasionally provide the human operator with inaccurate assessments, potentially misleading the operator into his or her own misjudgment. More insidiously, errors from an aid may cause the operator to engage with or depend on the aid in a suboptimal way (Parasuraman, 2000). This generally takes one of two forms, *misuse* or *disuse* (Parasuraman & Riley, 1997). Misuse occurs when the operator relies too heavily on the aid, acting on its judgments uncritically, without attempting to corroborate them. Disuse occurs when the operator relies on the aid too little, ignoring or underweighting its judgments (Parasuraman & Riley, 1997). Either tendency may compromise the benefits of automated assistance, and in the worst case can even increase the risk of error relative to unaided human performance (e.g., Alberdi, Povyakalo, Strigini, & Ayton, 2004). Optimal interaction thus depends on the ability of the human operator to calibrate his or her own use of the aid to the reliability of the aid (Lee & See, 2004; Wiegmann, Rich, & Zhang, 2001). This should

result in collaborative performance that surpasses the level obtained by either the unassisted human operator or the automated aid alone (Bartlett & McCarley, 2017; Robinson & Sorkin, 1985).

A goal in the design of automated decision aids is therefore to encourage appropriate automation use from the human operator, discouraging either misuse or disuse. One apparent way to encourage a closer-to-ideal pattern of human operator use is to change the aids' response bias (Dixon, Wickens, & McCarley, 2007; Murrell, 1977; Rice & McCarley, 2011). Automated decision aids can often be thought of as performing a standard signal detection task, discriminating between two or more distinct states of the world (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). A transportation security agent screening x-rays of carry-on baggage for threat items, for example, will encounter some bags that contain a threat (*signal events*) and some bags that do not (*noise events*). An automated aid that scans each bag and provides a target-present/target-absent judgment can produce four possible decision response outcomes: *hits* (correctly detecting the presence of a signal), *correct rejections* (CRs: correctly rejecting the presence of a signal), *false alarms* (FAs: incorrectly reporting the presence of a signal), and *misses* (incorrectly rejecting the presence of a signal) (Botzer, Meyer, Bak, & Parmet, 2010; Stanislaw & Todorov, 1999).

The aid's response threshold setting determines which of the categories of responses the aid is inclined toward. An aid with a conservative response threshold (i.e., miss-prone) will produce a miss rate higher than its false alarm rate, and complementarily, a hit rate lower than its correct rejection rate. An aid with a liberal response threshold (i.e., false alarm-prone) will produce a miss rate lower than its false alarm rate, and a hit rate higher than its correct rejection rate. An aid with a neutral response threshold (i.e., unbiased) will produce a miss rate equal to its false alarm rate, and a hit rate equal to its correct rejection rate. Sensitivity, the ability to distinguish between states of the world, can be measured by the statistic *d'*,

$$d' = z\,(HR) - z\,(FAR),$$

and response bias can be measured by the statistic *c*,

$$c = -0.5 \times [z\,(HR) - z\,(FAR)],$$

Assuming the signal and noise distributions have the same standard deviations, and are normally distributed (Macmillan & Creelman, 2005), signal detection theory, most notably, allows for the separation of sensitivity and bias, such that sensitivity does not vary with changes in response bias.

For a signal detector operating in isolation, the optimal response threshold is fully determined by the relative likelihood of signal and noise events and the payoffs attached to the decision outcomes (Macmillan & Creelman, 2005). Because the performance of a human-automation team is a joint function of the aid's behavior and the human operator's usage strategy, however, the threshold that optimizes performance of a detector in isolation may not be the ideal threshold for the same detector used as an aid to a human operator (Lehto, Papastavrou, Ranney, & Simmons, 2000; Papastavrou & Lehto, 1996; Sorkin & Woods, 1985). For example, the base rate of critical signal events in many applied contexts is extremely low (Parasuraman, Hancock, & Olofinboba, 1997; Sanquist, Minsk, & Parasuraman, 2008). In such cases, a conservative criterion, even if statistically optimal, may produce an unacceptably low hit rate, effectively leaving the operator to perform the task unaided, and unprepared to respond to an alert from the aid on the very rare occasion that one occurs. A more liberal criterion may be optimal when the costs of missed events are significant, but will reduce the posterior predictive probability of the aid's positive responses, engendering less responsive behavior from the operator (Getty, Swets, Pickett, & Gonthier, 1995).

But even when the base rates of signal and noise events are matched, operators may respond differently to liberal and conservative aids (e.g., Rice & McCarley, 2011). Most importantly, different forms of error from an aid—false alarm or miss—appear to produce qualitatively different

effects on human operator behavior (Cotté, Meyer, & Coughlin, 2001; Lehto et al., 2000). False alarms produce a 'cry wolf' effect, making the human operator less willing to comply with the aid's future signal-present judgments (Bliss, Dunn, & Fuller, 1995; Bliss, 1997; Dixon & Wickens, 2006; Maltz & Shinar, 2003). More surprisingly, false alarms can also make the human operator unwilling to act on a correct signal-absent judgment from the aid, even when the aid's signal-absent judgments are highly reliable (Dixon & Wickens, 2006; Dixon et al., 2007; Rice & McCarley, 2011). In other words, automation false alarms compromise both *compliance*, the human operator's willingness to act on a signal-present assessment from the aid, and *reliance*, the operator's willingness to act on a signal-absent assessment (Meyer, 2001, 2004; Wickens & McCarley, 2008). Misses from the aid, in contrast, tend to degrade reliance but have little effect on compliance (Dixon et al., 2007; Rice & McCarley, 2011).

One potential reason for the asymmetrical influence of automation false alarms and misses on human operator behavior is that false alarms are frequently accompanied by a salient perceptual event (e.g., an auditory alarm) and are thus easy to detect, whereas automation misses frequently go unnoticed (Dixon et al., 2007). This alone, though, does not appear to fully account for the human operators' tendency to depend on false alarm-prone aids less than miss-prone aids, as the bias against false alarm-prone aids persists even when the two forms of error are matched in perceptual salience (Rice & McCarley, 2011). This suggests a difference in cognitive salience between false positive and false negative responses from an automated aid (i.e., Rice & McCarley, 2011). Automation false alarms, that is, may be noticed, remembered, or weighted more in decision making than automation misses, even when the two forms of error are similar in gross perceptual characteristics.

The present study aimed to replicate this asymmetry, and to explore its potential cause. Rice and McCarley (2011), noting that unaided participants in their study showed a conservative

response bias, speculated that an asymmetry in the effects of automation misses and false alarms might reflect a tendency for operators to agree with automation whose response bias matches their own. Operators with a conservative bias, that is, would find themselves agreeing more often with a conservative aid than with a liberal aid, and might therefore be less inclined to disuse the conservative aid. Alternatively, operators may simply have an inherent tendency to disuse false alarm-prone aids more than miss-prone aids (McCarley, Rubinstein, Steelman, & Swanson, 2011; McCarley, Steelman, & Rubinstein, 2013).

The present experiment tested these possibilities by independently manipulating the response bias of both the human operator and their automated assistant in an aided signal detection task. Participants performed a simulated baggage screening task (McCarley, Kramer, Wickens, Vidoni, & Boot, 2004) similar to that used in several earlier studies of human-automation interaction (e.g., Merritt, 2011; Merritt, Heimbaugh, LaChapell, & Lee, 2013; Rice & McCarley, 2011; Wiegmann, McCarley, Kramer, & Wickens, 2006). Some performed the search task alone, whereas others performed the task with assistance from a 95%-reliable automated decision aid prone to either misses or false alarms. Concurrently, a point system assigning different payoffs to correct and incorrect responses encouraged the participants themselves to use either a conservative, neutral, or liberal response bias.

Hypotheses

Hypothesis 1: On average, sensitivity (*d'*) will be higher and response time (RTs) will be shorter for participants assisted by an aid than for those who are unassisted (e.g., Rice & McCarley, 2011). This effect will serve as a manipulation check that participants could and did use the automated aid to improve their performance.

*Hypothesis 2: S*ensitivity will be lower and mean RT will be longer for participants assisted by the liberal (false alarm-prone) aid than for those assisted by the conservative (miss-prone) aid (e.g., Rice & McCarley, 2011).

Hypothesis 3: A payoff matrix that penalizes misses more than FAs will induce participants to use a more liberal response bias than does an unbiased matrix. A payoff matrix that penalizes FAs more than misses will induce participants to use a more conservative response bias than does an unbiased matrix. This effect will serve as a manipulation check that differences in the payoff matrix influenced participants' response bias as intended.

Hypothesis 4: Automation false alarms will compromise compliance, the human operator's willingness to act on a signal-present assessment from the aid, whereas automation misses will degrade only reliance, the human operator's willingness to act on a signal-absent assessment (Dixon & Wickens, 2006; Meyer, 2001, 2004; Rice & McCarley, 2011).

Hypothesis 5: A tendency for participants to agree with an aid whose bias matches their own will emerge as an interaction between payoff matrix and aid response bias, on agreement rates (i.e., compliance and reliance).

Method

Participants. Participants were 140 adults (mean age = 22.46 years, *SD* = 5.02, range = 17-40; 102 females, 38 males) recruited from Flinders University. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 60 min. Participants were fluent in English, had normal color vision, and had normal or corrected-to-normal visual acuity. This research complied with the tenets of the Declaration of Helsinki and was approved by the Social and Behavioural Research Ethics Committee at Flinders University. Informed consent was obtained from all participants.

Apparatus and Stimuli. The experimental task was controlled by E-prime (Psychology Software Tools, Inc., Pittsburgh, PA) and stimuli were presented on a 23-inch Samsung monitor with a resolution of 1,920 x 1,080 pixels and a 120 Hz refresh rate. Participants were seated approximately 60 cm from the monitor, with viewing distance unconstrained.

Stimuli were 200 pairs of colored x-ray images (1,024 x 768 pixels) of passenger baggage (e.g., suitcases, backpacks, briefcases) lightly-to-heavily cluttered with a variety of common objects (e.g., clothing, keys, shoes, glasses), presented on a white background. These stimuli were identical to those created and used by Rice and McCarley (2011). Paired images were identical except for the fact that one of the images contained a knife (target-present), while the other did not (target-absent). Figure 2-1 shows a sample target-present stimulus image. The target of interest was a knife, 10 pixels wide x 100 pixels long, viewed on its flat side. The knife was randomly located in each target-present image, at a random orientation of 0º, 45º, 90º, 135º, 180º, 225º, or 315º within the frontoparallel plane.

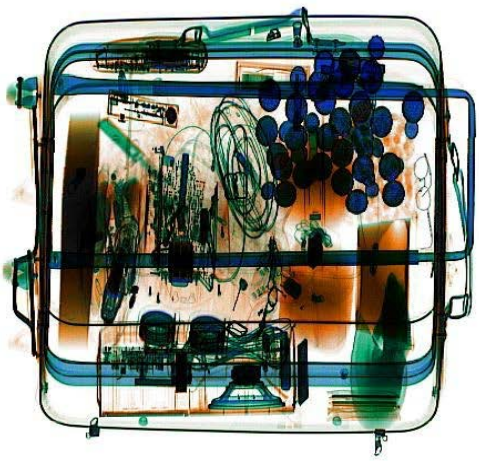

Figure 2-1. A sample target-present stimulus image.

Procedure. Upon arrival, participants first read and completed a consent form before starting the task. Demographic information (i.e., age and gender) was also collected from participants. Participants began the session by performing a block of 20 unaided practice trials to familiarize themselves with the search task. The instructions informed them, "For this study, you will pretend to be an airport security worker looking for knives hidden in x-ray images of passenger luggage. Each trial you will see an image of a bag, and your job will be to decide whether or not it contains a knife." Participants were asked to press the letter F on the keyboard to indicate that the knife was absent, and to press the letter J to indicate that the knife was present. Instructions informed the participants, "You should be as accurate as possible when making your responses. Do not feel rushed, but do try to avoid taking longer than is necessary. If there IS a knife present in the luggage, it will blink on and off after you have made your response to alert you to its hidden location. Otherwise, if you make a false alarm or miss a target, you will receive a text message as feedback." The instructions featured an image of the target knife, but no mention was made of an automated diagnostic aid. Before the practice trials began, the image of the target knife was once more presented to participants.

At the conclusion of the practice trials, participants performed a set of 180 experimental trials. These included 90 target-present trials and 90 target-absent trials, resulting in a signal rate of 50%. The order of trials and the presentation of target-present and target-absent images was randomized for each participant. Participants in the automation-aided experimental groups read a second set of instructions, and were advised that the aid would provide a text diagnosis as to the relative safety or threat of each bag (i.e., "The computer has determined that this bag is safe" or "The computer detects a target"). They were also advised of which type of error the aid would be prone to committing. Participants assisted by a false alarm-prone aid were told, "If the automation errs, it will err by falsely detecting a target when there really is none. It will never err any other

way." Participants assisted by a miss-prone aid were told, "If the automation errs, it will err by missing a target when there really is one present. It will never err any other way." Instructions to participants in the unaided group made no mention of an automated aid. Participants in the unaided group were provided with a neutral message, "Waiting for bag" at the start of each trial. The neutral message served to match the sequence and timing of events within trials across the aided and unaided groups.

A point system assigning different payoffs to correct and incorrect responses encouraged participants to adopt either a conservative, liberal, or neutral response bias. Participants encouraged to adopt a conservative bias were penalized 5 points for false alarms and 1 point for misses. Participants encouraged to adopt a liberal bias were penalized 5 points for misses and 1 point for false alarms. Unaided participants and participants encouraged to adopt a neutral bias were penalized 1 point for any kind of error. In all groups, correct responses were rewarded with 1 point each. The total score that could be obtained was 180 points. A cover story used in the instructions served to motivate the scoring system for participants. Participants were advised, "Management has declared that false alarms (misses) have been affecting traffic flow. Therefore, to try and stop these errors, they have decided to score your performance on each trial. Correct identifications and correct rejections = +1 POINT. Misses = -1(-5) POINTS. False alarms = -1(-5) POINTS."

Each trial was initiated with a "Hit a key to start the next trial" screen against a white background, followed by a 1,000-ms screen displaying the automated aid's text diagnosis. The stimulus image remained on-screen until a key press was made. At the conclusion of each trial, participants would receive a 1,500-ms green feedback message of, "Correct! +1 ; Total score = score" for all correct responses. For any incorrect responses participants would receive a 1,500-ms red feedback message of, "False Alarm! -1/-5; Total score = score" or "Miss! -1/-5; Total score =

score." Figure 2-2 shows the sequence of trial events within an automation-aided trial. An experimental session lasted approximately 60 min.

Experimental Design

Bias of the aid (liberal vs. conservative) and bias of the participant's payoff matrix (liberal, unbiased, conservative) were manipulated in a 2 x 3 between-subjects design. To provide an estimate of unaided performance, an additional group of participants performed the baggage screening task unassisted by the aid. Twenty participants were randomly assigned to each group.

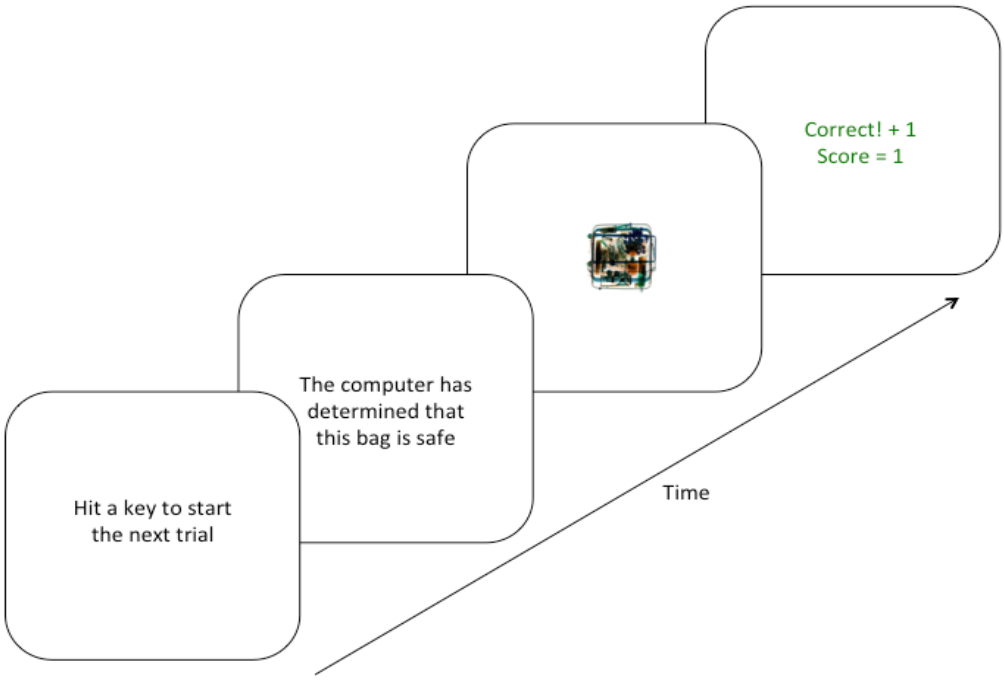

Figure 2-2. The sequence of events within an automation-aided trial.

Measures

Hit rates and false alarm rates were calculated from the participants' responses, and data were converted to signal detection measures of sensitivity and response bias, *d'* and *c* (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). A prior of 0.5 was added

to the raw response frequency value in each cell of the 2 x 2 SDT matrix for each participant to correct for perfect hit and false alarm rates (Hautus, 1995). Performance measures also included reaction times (RTs), calculated separately for target-present and target-absent trials, and measures of compliance and reliance. The first 40 trials for each participant were treated as practice and were excluded from analysis.

Target-present and target-absent RTs were analyzed separately as they are reflective of different cognitive processes involved in visual search: target-present RTs tend to be a direct measure of visual search efficiency, whereas target-absent RTs reflect decisional processes involved with terminating an unsuccessful search (Chun & Wolfe, 1996; Drury & Chi, 1995). RTs for inaccurate responses were excluded from analysis. Analyses including those data, however, produced effects similar to those reported here.

Compliance was measured by calculating the proportion of trials on which the participant agreed with the aid when it made a correct target-present judgment (Rice & McCarley, 2011). Reliance was measured by calculating the proportion of trials on which the participant agreed with the aid when it made a correct target-absent judgment (Rice & McCarley, 2011). Analysis was restricted to trials on which the aid made a correct judgment because aids of different bias made different forms of error—miss-prone aids never produced false-positive responses, FA-prone aids never produced false-negative responses—precluding an unconfounded comparison of agreement rates that included trials on which the aid produced a misjudgment.

Statistical Analyses

In place of conventional null-hypothesis significance tests, statistical analyses employed default Bayesian tests (Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The measures of evidence reported are Bayes factors, likelihood ratios indicating the degree to which the data favor one of two models relative to

the other. We report the Bayes factor calculated with the model including a statistical effect (i.e., a main effect or interaction) of interest in the numerator, and the model excluding the effect in the denominator. These values are labeled B_{10} (Rouder & Morey, 2012). Values of B_{10} greater than 1 give evidence in favor of a statistical effect, and values less than 1 give evidence against it. Main effects and interactions were tested by comparing the fits of a full model, assuming all main effects and interactions, to the fits of models that selectively excluded the effect of interest (Rouder et al., 2012). To ease interpretation, values of B_{10} less than 1 are presented as fractions. For example, a B_{10} of 0.20 is reported as B_{10} = 1/5, indicating a likelihood ratio of 5:1 in favor of the null hypothesis. Terms used to discuss the evidential impact of the reported Bayes factors are borrowed from Wetzels and colleagues (2011), and are as follows: anecdotal or worth no more than a bare mention ($1/3 < B_{10} < 3$), substantial ($1/10 < B_{10} \leq 1/3$ or $3 \leq B_{10} < 10$), strong ($1/30 < B_{10} \leq 1/10$ or $10 < B_{10} < 30$), very strong ($1/100 < B_{10} \leq 1/30$ or $30 \leq B_{10} < 100$), and decisive ($B_{10} < 1/100$ or $B_{10} > 100$). Most notably, values of B_{10} in between 1/3 and 3 are termed anecdotal and by convention are considered to provide very little evidence against or for an effect (Wetzels et al., 2011).

Results

Response Bias. Figure 2-3 presents mean *c* values. Data were first submitted to a series of one-sample tests to compare response bias for the unaided group to zero. Analysis gave decisive evidence that response bias in the unaided group was conservative, $B_{10} > 5e+5$, which is consistent with earlier findings (e.g., Rice & McCarley, 2011). Data were next submitted to a series of one-sample tests to compare response bias for the aided groups to zero. Analysis gave substantial evidence for liberal response bias in the liberal aid/liberal payoff matrix group, $B_{10} = 25.95$, but gave anecdotal to decisive evidence for conservative bias in all the remaining groups (B_{10}'s ranged from 1/3.33 – 1e+5).

To test for effects of aid bias and payoff matrix, data were then submitted to a two-way analysis with aid bias (liberal vs. conservative) and payoff matrix (conservative vs. unbiased vs. liberal) as between-subjects factors. Supporting hypothesis 3, analysis gave decisive evidence in favor of a main effect of aid bias, $F(1,114) = 22.59$, $\eta^2_G = 0.16$, $B_{10} = 3e+3$, and a main effect of payoff matrix, $F(2,114) = 23.73$, $\eta^2_G = 0.29$, $B_{10} > 5e+6$. Failing to support hypothesis 5, however, analysis gave anecdotal evidence against an interaction, $F(2,114) = 1.27$, $\eta^2_G = 0.02$, $B_{10} = 1/2.94$.

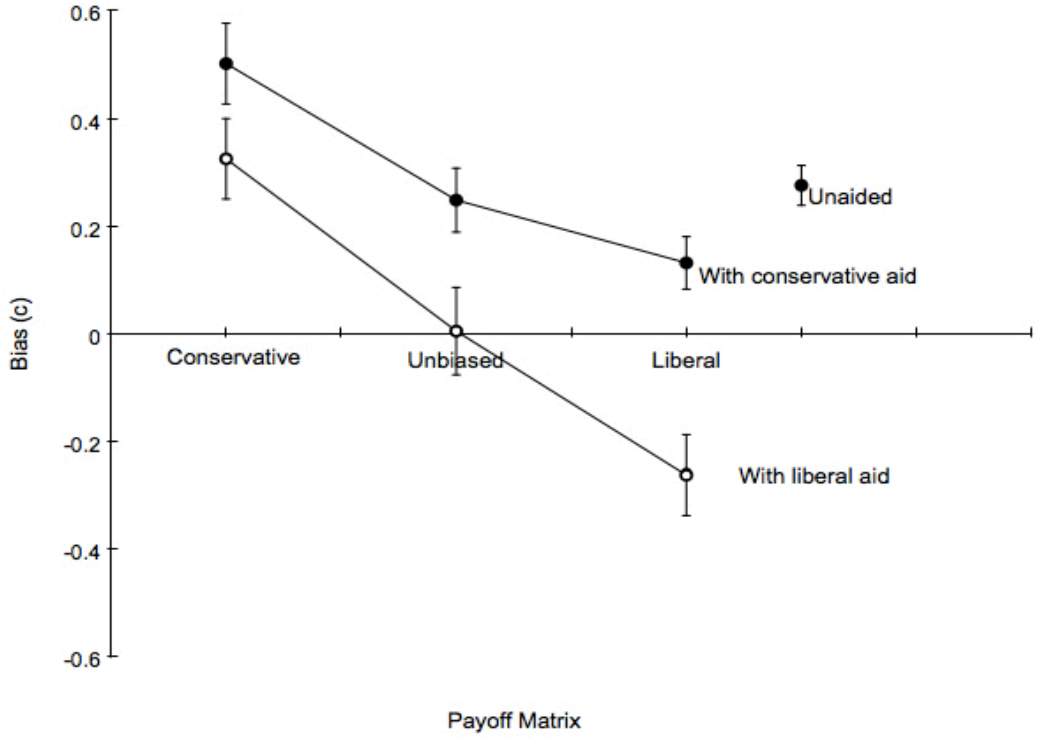

Figure 2-3. Mean *c* values. Error bars indicate standard errors.

Sensitivity. Figure 2-4 presents mean *d'* values. Data were first submitted to a series of pairwise comparisons to compare sensitivity for the unaided group to sensitivity for all other groups. Supporting hypothesis 1, analysis gave very strong to decisive evidence that sensitivity in the unaided group was lower than in any of the other groups (B_{10}'s ranged from 61.01 – 8e+5).

To test whether sensitivity varied with the aid's bias, data were next submitted to a two-way analysis with aid bias and payoff matrix as between-subjects factors. Analysis gave

substantial evidence against a main effect of payoff matrix, $F(2,114) = 1.36$, $\eta^2_G = 0.02$, $B_{10} = 1/4$. Most notably, though, failing to support hypothesis 2 and replicate earlier findings, data gave anecdotal evidence against a main effect of aid bias, $F(1, 114) = 1.84$, $\eta^2_G = 0.02$, $B_{10} = 1/2.27$. Failing to support hypothesis 5, analysis also gave substantial evidence against an interaction, $F(2, 114) = 0.56$, $\eta^2_G = 0.01$, $B_{10} = 1/5$.

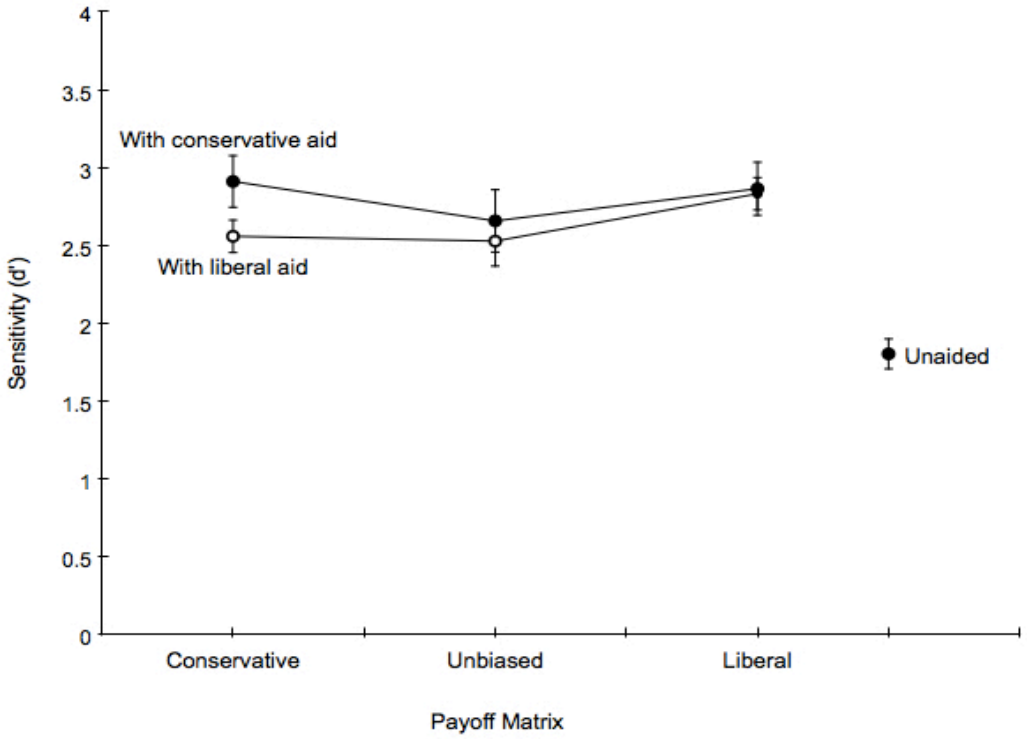

Figure 2-4. Mean *d'* values. Error bars indicate standard errors.

Target-Present RTs. Figure 2-5 presents mean target-present RTs. Data were first submitted to a series of pairwise comparisons to compare target-present RTs for the unaided group to target-present RTs for all other groups. Analysis gave substantial evidence that target-present RTs in the unaided group were longer than target-present RTs in the conservative aid/liberal payoff matrix group ($B_{10} = 3.81$), and anecdotal evidence that target-present RTs in the unaided group were longer than target-present RTs in the conservative aid/unbiased payoff matrix group ($B_{10} = 2.44$).

Analysis gave anecdotal evidence, however, that target-present RTs in the unaided group were no different from target-present RTs in all other groups (B_{10}'s ranged from 1/1.43 – 1/3.23).

To test whether target-present RTs varied as a function of the aid's bias, data were submitted to a two-way analysis with aid bias and payoff matrix as between-subjects factors. Data gave substantial evidence against a main effect of payoff matrix, $F(2,114) = 1.03$, $\eta^2_G = 0.02$, $B_{10} = 1/5.55$, and against an interaction, $F(2,114) = 0.05$, $\eta^2_G = 0.001$ $B_{10} = 1/7.14$. Supporting hypothesis 2, however, analysis did show strong evidence in favor of a main effect of aid bias, $F(1,114) = 10.17$, $\eta^2_G = 0.09$, $B_{10} = 18.8$, indicating that participants produced faster target-present responses when assisted by a miss-prone aid than when assisted by a false alarm-prone aid. This suggests, as expected, that participants were more reluctant to comply with target-present judgments from a false alarm-prone aid than from a miss-prone aid.

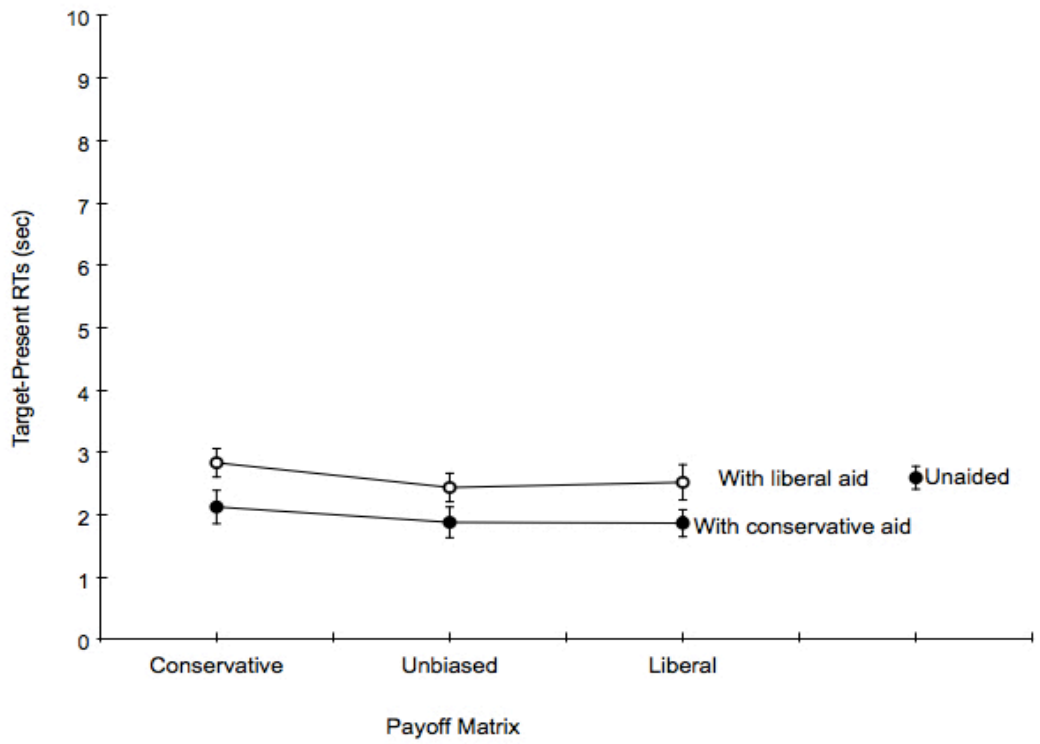

Figure 2-5. Mean target-present RTs. Error bars indicate standard errors.

Target-Absent RTs. Figure 2-6 presents mean target-absent RTs. Data were first submitted to a series of pairwise comparisons to compare target-absent RTs for the unaided group to target-absent RTs for all other groups. Analysis gave very strong to decisive evidence that target-absent RTs in the unaided group were longer than target-absent RTs in any other group (B_{10}'s ranged from 95.8 – 2e+3).

To test whether target-absent RTs varied as a function of the aid's bias, data were submitted to a two-way analysis with aid bias and payoff matrix as between-subjects factors. Analysis gave substantial evidence against a main effect of payoff matrix, $F(2,114) = 0.31$, $\eta^2_G = 0.01$, $B_{10} = 1/4.54$, and an interaction, $F(2,114) = 0.034$, $\eta^2_G = 0.001$, $B_{10} = 1/9.09$, and was indifferent toward a main effect of aid bias, $F(1,114) = 2.41$, $\eta^2_G = 0.02$, $B_{10} = 1/0.84$.

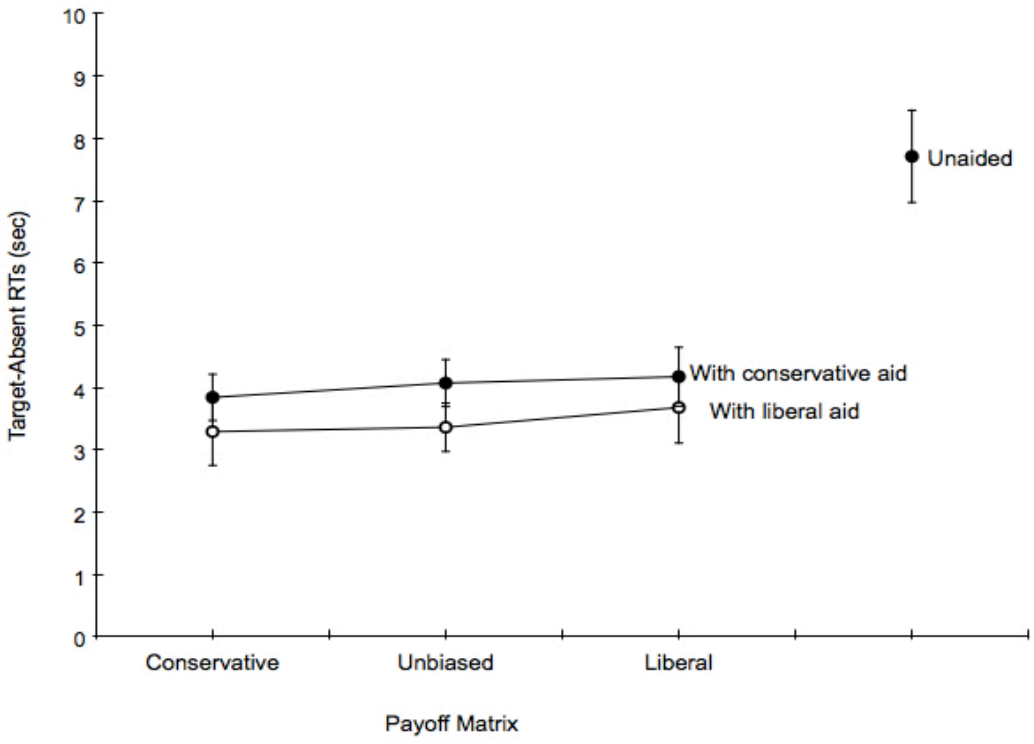

Figure 2-6. Mean target-absent RTs. Error bars indicate standard errors.

Compliance. Figure 2-7 presents mean compliance rates. Data were submitted to a two-way analysis with aid bias and payoff matrix as between-subjects factors. Analysis gave decisive evidence in favor of a main effect of payoff matrix, $F(2,114) = 10.25$, $\eta^2_G = 0.15$, $B_{10} = 339.2$, but substantial evidence against a main effect of aid bias, $F(1,114) = 0.84$, $\eta^2_G = 0.01$, $B_{10} = 1/3.57$, and against an interaction, $F(2,114) = 0.16$, $\eta^2_G = 0.003$, $B_{10} = 1/6.66$. Compliance thus increased as the payoff matrix encouraged more liberal responses, but was invariant across manipulations of the aid's bias.

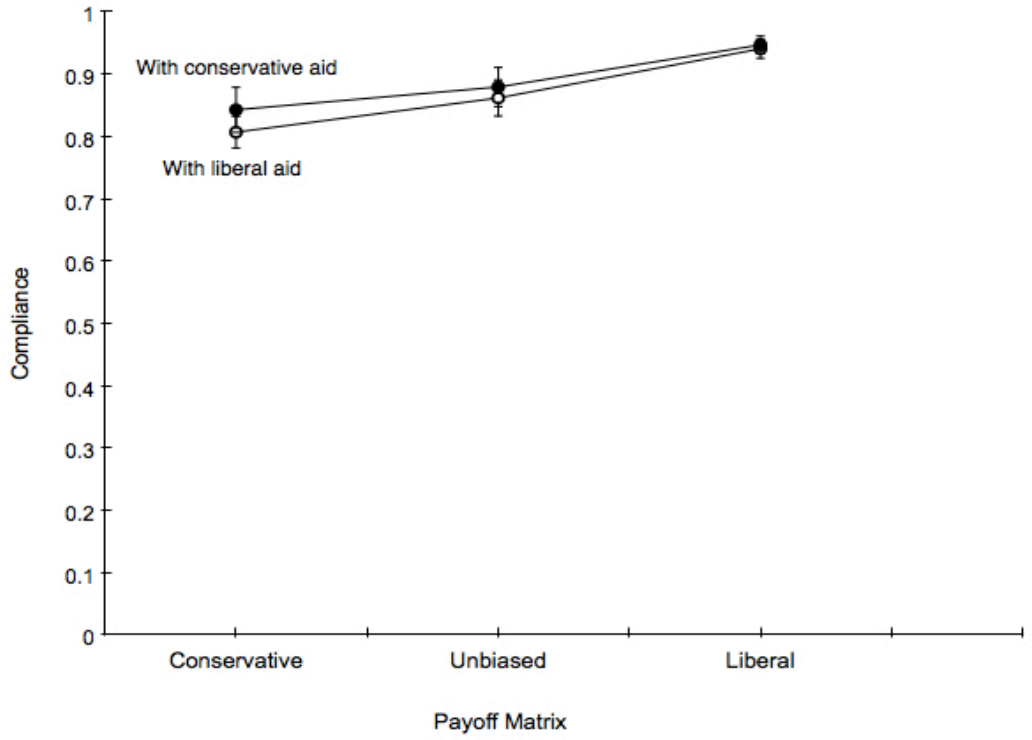

Figure 2-7. Mean compliance rates. Error bars indicate standard errors.

Reliance. Figure 2-8 presents mean reliance rates. Data were submitted to a two-way analysis with aid bias and payoff matrix as between-subjects factors. Analysis gave substantial evidence in favor of a main effect of payoff matrix, $F(2,114) = 0.31$, $\eta^2_G = 0.01$, $B_{10} = 4.54$, but was effectively indifferent toward a main effect of aid bias, $F(1,114) = 2.42$, $\eta^2_G = 0.02$, $B_{10} = 1/0.84$, and substantial evidence against an interaction, $F(2,114) = 0.03$, $\eta^2_G = 0.001$, $B_{10} = 1/9.09$. Reliance

thus increased as the payoff matrix encouraged more conservative responses, but showed no clear effect of the aid's bias. Among the 60 participants assisted by a conservative aid, 13 showed 100% reliance. Among the 60 participants assisted by a liberal aid, only 4 more (13/60) showed 100% reliance.

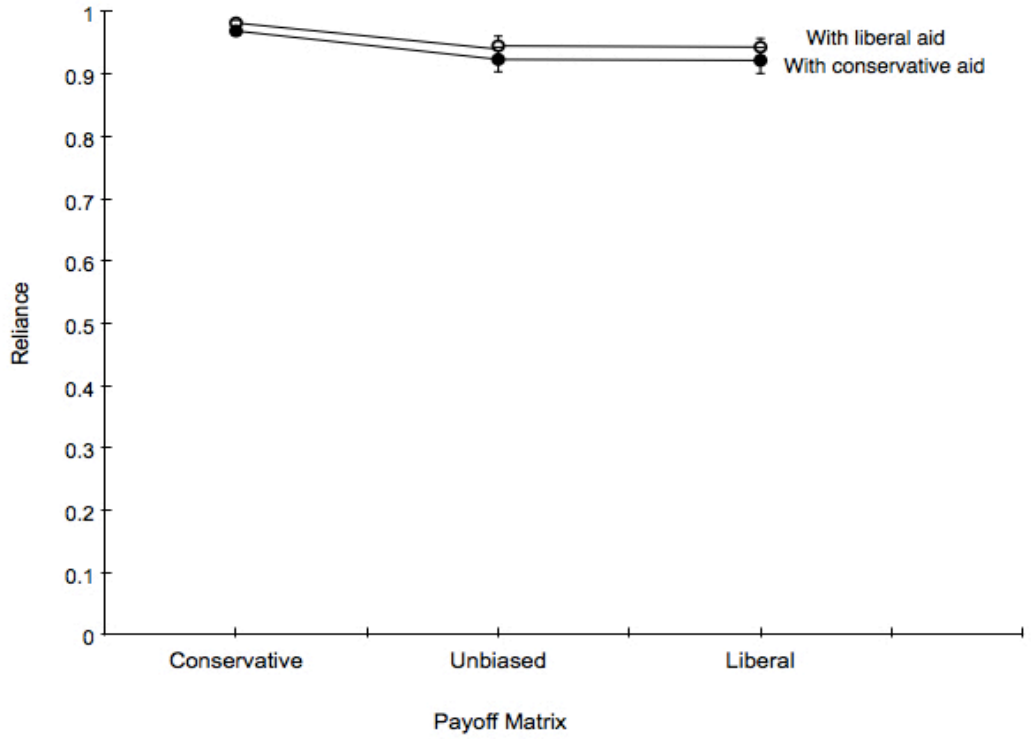

Figure 2-8. Mean reliance rates. Error bars indicate standard errors.

Discussion

Past work has suggested that false alarms from an automated signal detection aid may engender stronger disuse and lower sensitivity from human operators than misses from the aid (Dixon et al., 2007), even when the two forms of error are matched in perceptual characteristics (Rice & McCarley, 2011). The present experiment sought to replicate the asymmetrical effects of automation false alarms and misses and test whether they reflect a tendency for operators to agree with automation whose response bias matches their own. Participants performed a simulated x-ray baggage screening task, either alone or with assistance from a 95%-reliable automated decision aid

prone to either misses or false alarms. The response bias of the aid and the response bias of the participants were manipulated orthogonally. Manipulations were designed to test whether participants' tendency to use the automated aid would vary with the similarity between the aid's bias and the participants'.

Aided sensitivity was uniformly higher than unaided sensitivity, confirming that assistance from the automated aid improved human performance. Contrary to earlier findings (Rice & McCarley, 2011), though, data gave little evidence that false alarms from the aid compromised automation use more than did misses. Neither aided sensitivity nor compliance in the current experiment was statistically higher for participants assisted by a miss-prone aid than for those assisted by a false alarm-prone aid, with both measures trending at least anecdotally in the direction of a null effect. Notably, if participants in the conservative aid groups were interacting with the aid optimally, they would have shown compliance rates of 100%, since the aid never made a false alarm. The finding that compliance rates were below ceiling and did not differ between aided groups therefore indicates clearly suboptimal automation use by the participants assisted by the miss-prone aid. Similarly, if participants in the liberal aid groups were interacting with the aid optimally, they would have shown reliance rates of 100%, since the aid was guaranteed never to miss a target. The finding that reliance rates were below ceiling and did not differ between aided groups therefore indicates suboptimal automation use by participants assisted by the false alarm-prone aid. Evidence for a potential false alarm/miss asymmetry was found only in target-present RTs, which showed faster responses from participants assisted by miss-prone aids. This effect, though, was accompanied by a numeric (though not statistically credible) trend in the opposite direction in target-absent RTs. Moreover, the tendency to make an immediate positive response following a target-present judgment from a miss-

prone aid can be considered rational. It therefore should not be taken as clear evidence for a cognitive bias against false alarm-prone aids.

Thus, the current data largely failed to replicate the finding of asymmetrical costs of automation misses and false alarms reported by Rice and McCarley (2011). This result is surprising, given that the current experimental procedure was modeled closely after that of Rice and McCarley's study, and the current stimuli were the same as they used. The manipulation of participants' payoff matrix was novel to the present experiment, but did not interact with the manipulation of automation bias and therefore did not seem to account for the failure to replicate the expected asymmetry. The two studies differed in participant populations, one drawing participants from a Midwestern U.S. university and the other from an urban Australian university. It's not clear, though, why this difference in participant pools would engender such stark differences in outcome. The present data also leave it unclear whether false alarm- and miss-prone aids might still have asymmetrical effects on performance in cases when the two types of aid differ in perceptual characteristics, as, for example, when positive judgments from the aid are accompanied by an alert or salient visual signal (Dixon et al., 2007). Perhaps one explanation for the failure to replicate is the high reliability of the aid, since on average the aid made only nine incorrect judgments over the course of 180 experimental trials. Several considerations argue against this possibility, though. First, Rice and McCarley (2011) still found an asymmetry with a 95%-reliable decision aid in a procedure very similar to that of the current experiment. Second, the instructions provided at the outset of the current experiment informed participants whether their aids would be miss- or false alarm-prone. Finally, a reanalysis of the current data excluding the first 100 trials of the aided task as practice, allowing participants greater opportunity to witness errors from the aid before beginning experimental trials, produced a pattern of results identical to that reported above.

As little is currently known about the factors underlying operators' inherent tendency to agree with conservative aids, we asked whether this may simply reflect a tendency for operators themselves to be conservative. Understanding whether or not participants tend to agree more with automation whose response bias matches their own carries important implications for human factors practitioners. Firstly, it may allow for practitioners to better tailor the design of automated aids. For example, knowing whether an operator may agree more with a conservative aid than a liberal aid can inform the optimal response bias level of the automated aid (Sorkin & Woods, 1985) that will serve to elicit more efficient human-automation collaboration. Secondly, it may allow practitioners to better train operators. Just as the optimal response bias of the aid depends on the response bias of the operator, so does the benefits or payoffs attached to various decision outcomes (i.e., hits, correct rejections, false alarms, misses) (Rice & McCarley, 2011). Punishing the operator for false alarms, and rewarding the operator for hits, for example, may foster greater automation dependence and ultimately better human-automation collaboration. In any case, however, the current results imply that the asymmetrical effects of automation false alarms and misses on operators' behavior might be less robust than earlier evidence suggested.

Further research will be necessary to generalize this pattern of effects across different forms of signal detection task (e.g., identifying abnormalities in abstract shapes/images/letters; see McCarley et al., 2011), varying levels of aid reliability (e.g., Rice & McCarley, 2011), more realistic signal rates (e.g., 20% signal rate; see Lacson, Gonzalez, & Madhavan, 2008), and/or more extreme payoff schemes (e.g., deducting 40 points for each miss and 10 points for each FA; see Lacson, Wiegmann, & Madhavan, 2005).

This concludes the current published paper.

CHAPTER 3:
STUDY 2

The following manuscript entitled, *Benchmarking aided decision making in a signal detection task,* was published in Human Factors. The version of the manuscript presented here is the final, peer reviewed version, prior to the publisher applying their formatting. This manuscript has been published as:

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors, 59(6),* 881-900. Copyright © [2017] (Human Factors). Reprinted by permission of SAGE Publications.

Link to authoritative document (doi): http://dx.doi.org/10.1177/0018720817700258

All authors were involved in the formulation of the study concept and design, and data analysis. Megan Bartlett collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

In addition, the findings from this experiment have been presented at:

Bartlett, M. L., & McCarley, J. S. (2017, November). *Quantifying Suboptimal Automation Use.* Defence Human Sciences Symposium, Adelaide, Australia.

Introduction

Human operators in everyday and professional contexts work with the assistance of automated decision aids. The assisted tasks often take the form of binary signal detection judgments, which ask a decision maker to classify potentially ambiguous states of the world into either of two discrete categories (Green & Swets, 1966; Macmillan & Creelman, 2005). A credibility assessment aid, for instance, might help organizational decision makers distinguish deceptive from honest responses when questioning interviewees in negotiations or investigations (Jensen, Lowry, & Jenkins, 2011). Analogously, a combat identification system might help soldiers distinguish friends from foes on the battlefield (Wang, Jamieson, & Hollands, 2009). Ideally, assistance from an automated aid will help the human operator to achieve higher levels of *sensitivity*, the ability to distinguish between states of the world. But like the human operator, an automated decision aid performing a signal detection task is typically required to render judgments based on incomplete or uncertain data. The aid's sensitivity will therefore be imperfect, just as the human operator's is, and the aid's judgments will sometimes be wrong.

Imperfect sensitivity does not render an aid inherently useless. Even if the automation errs in occasional judgments, the human operator may be able to achieve a higher sensitivity with the aid's assistance than without it (Wickens & Dixon, 2007). In practice, unfortunately, people often interact with automated aids in a suboptimal way. This may manifest as either *misuse*, a tendency to act on the aid's judgments uncritically, or *disuse*, a tendency to disregard or underweight the aid's judgments (Parasuraman, 2000; Parasuraman & Riley, 1997). These effects compromise the benefits of automated assistance, and in the worst case, operators may even perform a task more poorly when assisted by a decision aid than when unassisted (e.g., Alberdi, Povyakalo, Strigini, & Ayton, 2004).

An important goal of automation design is therefore to encourage more efficient human-automation interaction, allowing the automation-aided operator to achieve higher levels of

sensitivity. Notably, automation-aided performance in a signal detection task can be conceptualized as a form of collaborative decision making in which two agents, the human and the aid, reach separate judgments about the state of the world and then combine their judgments to reach a joint decision (Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001). Understanding the process by which the operator integrates his or her own judgment with that of the aid may thus allow practitioners to better tailor the design of automated aids, to encourage efficient human-automation collaboration. In the worst case, it will allow system designers to better predict automation-aided performance levels.

Using Binary Cues: The Contingent Criterion Model

Many studies of automation-aided decisions have specifically considered the case in which an aid provides the human operator binary judgments (e.g., Botzer, Meyer, Pak, & Parmet, 2010; Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Rice & McCarley, 2011). Robinson and Sorkin's (1985) *contingent criterion* (CC) model has become the modal account of human-automation interaction under these circumstances (e.g., Elvers & Elrif, 1997; Maltz & Meyer, 2001; Meyer, 2001). The model is built on the framework of signal detection theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005). SDT assumes that to render a signal detection judgment, the decision maker first encodes evidence for or against either of two exhaustive and mutually exclusive potential states of the world, one of which is conventionally termed *signal* and the other *noise*. The evidence values are distributed continuously, and unless the task is trivially easy, the evidence distributions corresponding to the two states of the world overlap at least partially.

The decision maker transforms continuous evidence values into discrete judgments by comparing them to a response criterion. Values below the criterion value lead to a judgment of signal absent, and values above it lead to a judgment of signal present. The decision maker's

criterion may be *conservative*, biased toward judgments of noise; *liberal*, biased toward judgments of signal; or *unbiased*. Assuming that the signal and noise evidence distributions are Gaussian with a common standard deviation (Macmillan & Creelman, 2005), sensitivity can be measured by the statistic *d'*,

$$d' = z\,(HR) - z\,(FAR),$$

and bias by the statistic *c*,

$$c = -0.5 \times [z\,(HR) - z\,(FAR)].$$

A value of $d' = 0$ indicates chance performance, and a value of $d' = 5$ indicates near perfect performance. Negative values of *c* indicate liberal bias, positive values indicate conservative bias, and a value of 0 indicates unbiasedness.

The CC model views the aid and the human operator as operating in sequence, with the aid rendering its judgment first and the operator establishing his or her own response criterion contingent on the aid's judgment. The operator is thus presumed to operate with a relatively liberal response criterion following a judgment of signal present from the aid, and with a relatively conservative criterion following a judgment of signal absent from the aid. For ease of exposition, we will refer to a signal present judgment as*Yes* and a signal absent judgment as *No*. Team hit rate under the CC model, HR_{CC}, is,

$$HR_{CC} = HR_{aid}\ (HR_{operator|\text{“Yes”}}) + (1 - HR_{aid})\ HR_{operator|\text{“No”}},$$

where HR_{aid} is the hit rate of the automated aid, $HR_{operator|\text{“Yes”}}$ is the hit rate of the unaided human operator given a *Yes* judgment from the aid, and $HR_{operator|\text{“No”}}$ is the hit rate of the unaided human operator given a *No* judgment from the aid. Team false alarm rate under the CC model, FAR_{CC}, is,

$$FAR_{CC} = FAR_{aid}\ (FAR_{operator|\text{“Yes”}}) + (1 - FAR_{aid})\ FAR_{operator|\text{“No”}},$$

where FAR_{aid} is the false alarm rate of the automated aid, $FAR_{operator|\text{“Yes”}}$ is the false alarm rate of the unaided human operator given a *Yes* judgment from the aid, and $FAR_{operator|\text{“No”}}$ is the false alarm

rate of the unaided human operator given a *No* judgment from the aid.

Team sensitivity under the CC model, d'_{CC}, is thus,

$$d'_{CC} = z\,(HR_{CC}) - z\,(FAR_{CC}).$$

The operator's optimal criterion setting following an aid's judgment is determined by the aid's predictive value (Robinson & Sorkin, 1985). Assuming an unbiased payoff matrix, normative bias following a response *i* from the aid, as measured by the statistic β, is,

$$\beta_{optimal} = [1\text{-}p(\text{signal}|i)] \,/\, p(\text{signal}|i),$$

where *i* is either a *Yes* or a *No*. Normative behavior thus entails larger bias shifts in response to more reliable automated aids. Data have shown that operators' response criteria in fact shift in the expected direction following a *Yes* or *No* judgment from an aid, but that the magnitude of these shifts is smaller than predicted by the normative CC model (Elvers & Elrif, 1997; Meyer, 2001; Wang et al., 2009). These findings have been taken as evidence that operators employ a CC strategy in automation-aided decision tasks, but choose their criteria suboptimally (cf., Botzer et al., 2010).

But while evidence for suboptimal automation use is incontestable, evidence that this is the result of a CC process is more tentative. Bias shifts in the direction of an aid's recommendation are consistent with a CC strategy. Other information integration strategies, however, will also produce differences in response bias conditional on the aid's decision. In fact, any collaborative strategy under which the operator tends to agree with the aid will engender differences in the operator's bias conditional on the aid's decision. Differences in conditional operator bias therefore do not necessarily implicate the decision process postulated by the contingent criterion model. Additionally, the suboptimal CC model by itself offers little help in anticipating the performance benefits that an automated aid will produce. While aided performance will be less than statistically ideal, the model does not specify just how far short of that standard it will fall. Phrased differently, whereas the operator's cued criterion settings are fixed in the optimal CC model, the suboptimal

model makes them free parameters, providing little *a priori* basis for predicting the operator's automation-aided sensitivity. Comparing automation-aided performance to the predictions of alternative, fixed-parameter or parameter-free models may therefore be useful both to identify strategies that provide plausible alternative accounts of human-automation decision making, and to establish benchmarks that help designers predict the performance levels automation-aided operators might attain.

Alternative Models of Binary Cue Use

A very simple strategy for interacting with an automated aid, proposed as a potential strategy for collaborative decision making between pairs of human decision makers, is the *best decides* (BD) model (Bahrami et al., 2010; Denkiewicz, Rączasek-Leonardi, Migdal, & Plewczynski, 2013). This model assumes that the human operator knows whether he or she is more or less sensitive than the aid. If more sensitive, the operator ignores the aid entirely and makes a judgment each trial for him or herself. If less sensitive, the operator defers to the aid's judgments by default. Team sensitivity under the BD model, d'_{BD}, is thus,

$$d'_{BD} = \max (d'_{operator}, d'_{aid}).$$

Although simpler than the CC strategy, the BD strategy makes far less efficient use of the paired decision makers' judgments, producing lower levels of automation-aided sensitivity. Nonetheless, observed automation-aided performance is often poorer still than predicted by the BD model (e.g., Meyer, 2001; Rice & McCarley, 2011).

Another pair of strategies, the *yes/yes* (YY) and *no/no* (NN) decision models proposed by Pollack and Madans (1964), are also inefficient, but again seem to outperform human-automation teams. Under the YY model, both the operator and the aid must report "signal present" for the team to produce a collaborative signal present judgment. Conversely, under the NN model, both the operator and the aid must report "signal absent" to produce a collaborative signal absent judgment.

Since the YY and NN decision models make symmetrical predictions, we will only discuss and report the predictions of the NN model. Team hit rate under the NN model, HR_{NN}, is,

$$HR_{NN} = 1- (1-HR_{operator}) (1-HR_{aid}),$$

where $HR_{operator}$ is the hit rate of the unaided human operator. Team false alarm rate under the NN model, FAR_{NN}, is,

$$FAR_{NN} = 1- (1-FAR_{operator}) (1-FAR_{aid}),$$

where $FAR_{operator}$ is the false alarm rate of the unaided human operator. Team sensitivity under the NN model, d'_{NN} is thus,

$$d'_{NN} = z (HR_{NN}) – z (FAR_{NN}),$$

and team criterion under the NN model, c_{NN}, is,

$$c_{NN} = -1/2 [z (HR_{NN}) + z (FAR_{NN})].$$

Pollack and Madans (1964) found that automation-aided participants achieved sensitivity levels lower than predicted by the NN and YY models.

Adapted to the context of human-automation decision making, Bahrami et al.'s (2010) *coin flip* (CF) model might provide a more plausible and better-fitting process model of human-automation performance. The model assumes that if the human operator and aid agree on a yes-or-no judgment, that's the judgment of the team. If they reach different decisions, the disagreement is effectively resolved by coin flip, that is, by selecting among the two response options randomly and with equal probability. The model thus posits discrete states in which the operator either ignores the model's judgment or defers to it fully. Predictions for the CF model in the current work can be made by estimating team hit rate (*HR*) and false alarm rate (*FAR*) from the individual team member's HR and FAR, then transforming those scores using the standard equation for calculating *d'*. Assuming that the human operates with the same response bias under individual and automation-aided conditions, team hit rate under the CF model, HR_{CF}, is,

$$HR_{CF} = (HR_{operator})\,(HR_{aid}) + 0.5\,(HR_{operator})\,(1\text{-}HR_{aid}) + 0.5\,(1\text{-}HR_{operator})\,(HR_{aid})$$

$$= 0.5\,(HR_{operator} + HR_{aid}),$$

Team false alarm rate under the CF model, FAR_{CF}, is,

$$FAR_{CF} = (FAR_{operator})\,(FAR_{aid}) + 0.5\,(FAR_{operator})\,(1\text{-}FAR_{aid}) + 0.5\,(1\text{-}FAR_{operator})\,(FAR_{aid})$$

$$= 0.5\,(FAR_{operator} + FAR_{aid}).$$

Team sensitivity under the CF model, d'_{CF}, is,

$$d'_{CF} = z\,(HR_{CF}) - z\,(FAR_{CF}),$$

and team criterion under the CF model, c_{CF}, is,

$$c_{CF} = -1/2\,[z\,(HR_{CF}) - z\,(FAR_{CF})].$$

Because the CF model reflects a highly inefficient strategy for combining agents' judgments (Bahrami et al., 2010), it may offer a more plausible account of human-automation collaboration than the models discussed above. Alternatively still, we may consider a model that is similar but potentially more consonant with empirical findings in the study of decision making. Like the CF model, the *probability matching* (PM) model, posits that yes-or-no disagreements between agents are resolved randomly. The PM model, however, assumes that the operator defers to the aid's judgment with a probability equal to the aid's average reliability, mimicking a strategy that participants use in probabilistic choice tasks (see Koehler & James, 2014; Vulkan, 2000, for reviews), including automation-aided decision tasks in which operators have no access to raw data (Bliss, Gilson, & Deaton, 1995; Wiegmann, 2002). Team hit rate under the PM model, HR_{PM}, is,

$$HR_{PM} = R_{aid} \times HR_{aid} + (1\text{-}\,R_{aid}) \times HR_{operator},$$

where R_{aid} is the aid's average reliability rate. Team false alarm rate under the PM model, FAR_{PM}, is,

$$FAR_{PM} = R_{aid} \times FAR_{aid} + (1\text{-}\,R_{aid}) \times FAR_{operator},$$

Team sensitivity under the PM model, d'_{PM} is,

$$d'_{PM} = z\,(HR_{PM}) - z\,(FAR_{PM}),$$

and team criterion under the PM model, c_{PM} is,

$$c_{PM} = -0.5 \times [z\,(HR_{PM}) - z\,(FAR_{PM})].$$

The CF and PM models can be considered variants of the same discrete-state model, differing only in the fixed probability with which the operator defers to the aid. Assuming the automated aid's decisions are more accurate on average than the operator's, the PM model offers an aided decision strategy more efficient than the CF model but nonetheless suboptimal.

Strategies for Using Direct Evidence Values

As noted, the models discussed above presume an aid rendering yes-or-no judgments. Phrased differently, they presume an aid that measures the strength of evidence for a signal and then applies a decision rule to transform that strength estimate into a binary judgment. Some empirical studies have examined variations on this design in which the aid renders confidence-graded judgments on a scale of more than two levels, providing a more fine-grained assessment of the evidence for or against a signal, but even in these cases the aid's judgments have been discretized. Automated aids in one study, for example, provided participants alarms on a 4-level scale, where the lowest level was the absence of a signal and the highest level denoted an urgent alarm (Sorkin, Kantowitz, & Kantowitz, 1988). A visual search aid in another study ranked potential target locations on a 5-level scale (St. John & Manes, 2002). Both of these studies found evidence for better human performance with graded than with binary automated cues, as have some others (Andre & Cutler, 1998; Gupta, Bisantz, & Singh, 2002; McCarley, 2009; Wiczorek & Manzey, 2014). Other research, however, has failed to replicate this benefit (Wickens & Colcombe, 2007; Wiczorek, Manzey, & Zirk, 2014).

An alternative and less-explored design option is to allow the aid to share its evidence estimates directly. By preserving information that is lost when responses are discretized, such direct

evidence sharing offers the potential of better human-automation performance than is achievable with standard, discrete judgments from an aid (Bahrami et al., 2010). The *optimal weighting* (OW) model (Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001), built on the assumption of direct evidence sharing from the aid, in fact offers the strategy for best-possible automation-aided performance. The model assumes that the human and the automated aid both operate as equal-variance Gaussian signal-detectors (Macmillan & Creelman, 2005). Each trial, both agents assess the stimulus independently and estimate the likelihood that it contains a signal. The automated aid reports its likelihood estimate to the human operator, with the automation-aided decision based on a weighted average, *Z*, of these estimates,

$$Z = \Sigma a_i X_i,$$

where *i* indexes the agent, human or automation, a_i is the weight applied to agent *i*'s estimate, and X_i is that estimate. Assuming the human and aid's judgments are stochastically independent, the optimal weight for agent *i* is proportional to the agent's sensitivity, d'_i. In the context of automation-aided decision making, team sensitivity under the OW model, d'_{OW}, is,

$$d'_{OW} = (d'^{2}_{operator} + d'^{2}_{aid})^{1/2}.$$

Another model for using direct evidence judgments from the aid, the *uniform weighting* (UW) model, is identical to the OW model except that it assumes that the operator assigns equal weights to the two estimates of signal likelihood when averaging them, i.e., that $a_{human} = a_{aid}$ (Sorkin et al., 2001). In this case, team sensitivity under the UW model, d'_{UW}, is,

$$d'_{UW} = (d'_{operator} + d'_{aid})/2^{1/2}.$$

If the aid and operator are equally sensitive, the UW model is equivalent to the OW model. Otherwise, d'_{UW} is lower than d'_{OW}.

Comparing the performance of the OW and UW models to the performance of the models discussed above suggests that human operators may benefit more from an aid that shares its

evidence assessments directly, without discretizing responses. As yet, though, this possibility apparently has not been tested empirically.

The Current Experiments

The models above span a range of performance levels, from perfectly efficient to highly inefficient. The present series of experiments tested the performance of automation-aided decision makers in a two-alternative forced choice (2AFC) task against the models to investigate human operators' strategies for interacting with an automated decision aid, and to benchmark empirical automation-aided performance. Participants viewed orange and blue random dot images, and were asked to determine each trial which color was dominant (Voss, Rothermund, & Voss, 2004). They performed the task alone or with assistance from an automated decision aid. The aid rendered its judgment either in the form of a binary diagnosis accompanied by an estimate of signal strength (Experiments 1 and 3), or simply as a binary diagnosis (Experiment 2). The predictions for each collaborative model were calculated from the participant's unaided sensitivity and the sensitivity of the aid. Observed collaborative sensitivity values were then compared to the statistically optimal values predicted by each model.

This research complied with the tenets of the Declaration of Helsinki and was approved by the Social and Behavioural Research Ethics Committee at Flinders University. Informed consent was obtained from all participants.

Experiment 1

Method

Participants. Participants were 40 adults (mean age = 20.97 years, *SD* = 3.76, range = 17-35; 34 females, 6 males) recruited from the Flinders University of South Australia. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 45 min. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. The experimental task was controlled by E-prime (Psychology Software Tools, Inc., Pittsburgh, PA), and stimuli were presented on a 23-inch Samsung monitor with a resolution of 1,920 x 1,080 pixels and a 120 Hz refresh rate. Participants were seated approximately 60 cm from the monitor, with viewing distance unconstrained.

Stimuli were 300 blue and orange random dot images (256 x 256 pixels). Figure 3-1 shows a sample orange-dominant stimulus image. Each stimulus was either blue-dominant or orange-dominant. In the blue-dominant stimuli, each pixel was randomly assigned the color blue with a probability of 0.52 or the color orange with a probability of 0.48. In the orange-dominant stimuli, those probabilities were reversed.

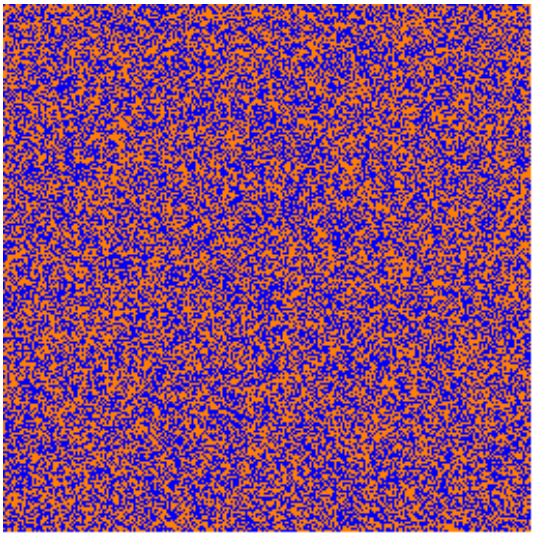

Figure 3-1. A sample orange-dominant stimulus image.

Procedure. Participants performed a 2AFC task requiring them to classify stimulus images as blue- or orange-dominant. A cover story asked the participants to imagine themselves as geologists sorting samples of a fictional mineral Vibranium into blue and orange strains. The instructions informed them, “Unfortunately, the two strains are difficult to tell apart. Both are speckled blue and orange. The only difference visually is that one strain tends to have a little more orange, and the other tends to have a little more blue. For

simplicity, we will call them VBN-ORANGE and VBN-BLUE. However, there is a lot of overlap in their appearance, and it is almost impossible to sort them with 100% accuracy by eye." Participants were asked to press the number 1 on the keyboard if they thought the image was mostly orange, and to press the number 3 on the keyboard if they thought the image was mostly blue.

Participants were also told that on some trials, they would be assisted by an automated decision aid that would provide a binary blue or orange judgment along with an estimate of signal strength. Instructions read, "The aid works by testing the chemical properties of the sample, and then assessing whether the sample is more likely to be VBN-ORANGE or VBN-BLUE. However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is correct 93% of the time and incorrect 7% of the time. To help you predict whether it is right or wrong, the aid will give its assessment along with a numeric rating each trial. A higher rating means that the aid is more likely to be correct. The aid will provide its assessment and rating at the start of each trial. You should use the aid to help you make your decisions, but be aware that you are free to disagree with it any time you wish. Use your own best judgement."

The aid's judgments were calculated using an equal-variance Gaussian signal detection model. Evidence values for blue-dominant images were sampled from a Gaussian distribution with a mean of -1.5 and a standard deviation of 1, and evidence values for orange-dominant images were sampled from a Gaussian distribution with a mean of 1.5 and a standard deviation of 1. Thus, the *d'* of the aid was 3. The aid transformed evidence values into binary judgments using an unbiased response threshold, offering a judgment of blue-dominant if the evidence value sampled for a given trial was less than 0 and a judgment of orange-dominant if the evidence value sampled was greater than 0. Given the aid's *d'* of 3,

the unbiased criterion produced an average accuracy rate of 93%. The aid's estimate of signal strength was simply the absolute value of the sampled evidence value. As noted, participants were informed that a higher value indicated stronger evidence. Because they were generally not expected to have had extensive formal training in statistics, however, they were not provided any additional information about the distribution of evidence values.

Figure 3-2 shows the sequence of events within an automation-aided trial for Experiment 1. Each trial was initiated with a key press from the participant. This was followed by a 1,000-ms fixation screen, a 1,000-ms screen displaying the automated aid's diagnosis, and then the stimulus display. On aided trials participants were provided with the aid's diagnosis, e.g., "Aid judges: Orange 2.14." On unaided trials, participants were instead provided with a neutral message, "Waiting for sample." Presentation of the aid's diagnosis before the stimulus display allowed participants time to attend to the diagnosis carefully, and ensured that the diagnosis and stimulus arrived in the same order in which the CC model presumes they are processed (though see Wiegmann, McCarley, Kramer, & Wickens, 2006, for evidence that automation dependence is similar regardless of the order in which cue and stimulus are presented). Other models make no presumption as to the order of processing. The neutral message served to match the sequence and timing of events across the aided and unaided blocks. The stimulus display remained onscreen until the participant's response. At the end of each trial, participants received a 1,500-ms feedback message of either "Correct!" or "Incorrect!"

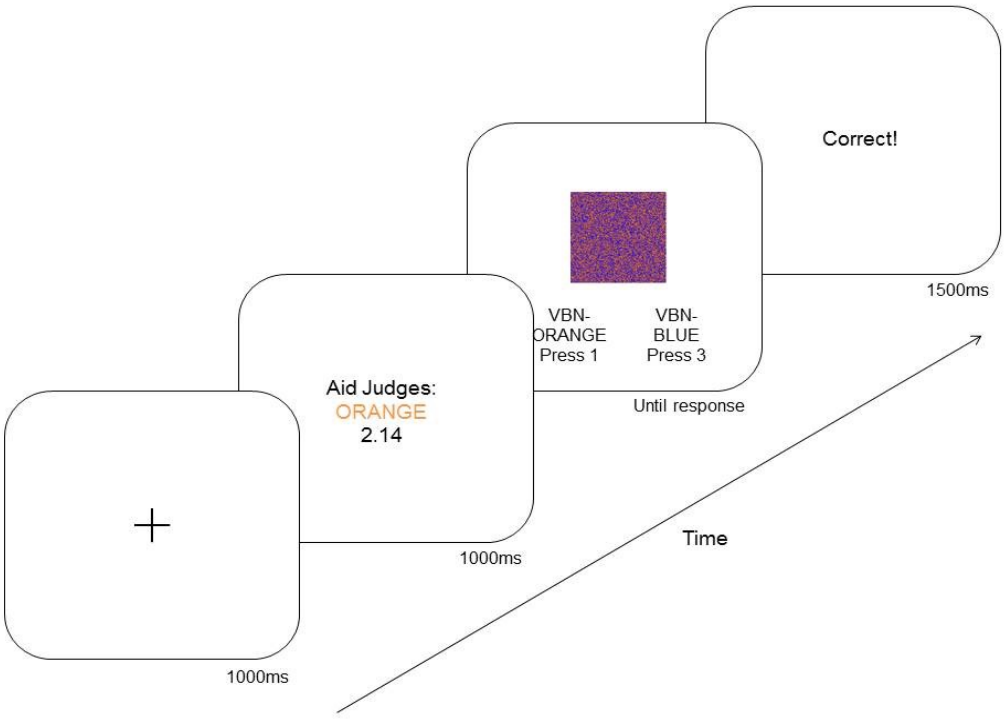

Figure 3-2. The sequence of events within an automation-aided trial for Experiment 1.

Each session comprised a block of 50 unaided practice trials followed by a block of 50 aided practice trials, then a block of 100 unaided experimental trials and a block of 100 aided experimental trials, with the order of the experimental blocks counterbalanced across participants. The order of stimulus images viewed within blocks was randomized across trials. Participants were allowed to rest between blocks. An experimental session lasted approximately 45 min.

Analysis

For analysis, orange-dominant stimuli were treated as signal events and blue-dominant stimuli as noise events. For clarity of exposition below, we refer to orange and blue judgments as *yes* and *no* judgments, respectively. Hit rates and false alarm rates were calculated from the participants' responses, and data were converted to signal detection measures of sensitivity and bias, *d'* and *c* (Green & Swets, 1966; Macmillan & Creelman,

2005; Stanislaw & Todorov, 1999). A prior of 0.5 was added to the raw response frequency value in each cell of the 2 x 2 SDT matrix for each participant to correct for perfect hit and false alarm rates (Hautus, 1995). Data from practice trials were excluded from analysis.

Data analysis employed Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) sampling procedure (Kruschke, 2013, 2015; Lee & Wagenmakers, 2014). This approach begins by assuming a prior distribution on a parameter value of interest, then updates the prior through probabilistic sampling to approximate the posterior distribution on parameter values in light of the observed data.

Analyses were conducted using sampling functions from the package JAGS (Plummer, 2015) in the R programming language (http://www.r-project.org). All parameters were assumed to follow normal distributions, with vague priors on their means and standard deviations (means ~ $N[0, 1 \times 10^6]$; standard deviations ~ $1/\Gamma[.001, .001]$). The use of vague priors ensures that the analysis does not commit *a priori* to strong conclusions, and allows the observed data to dominate the posterior distribution. Each estimate was based on four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step in order to reduce sample autocorrelation, leaving a total of 100,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less, indicating satisfactory convergence of the MCMC chains (Kruschke, 2015).

Descriptive statistics reported include the mean and 95% highest density intervals (HDI) for the estimated posterior distributions (Kruschke, 2013). The 95% HDI is the region that contains 95% of the posterior distribution mass, and within which all values have higher probability than any values outside the region. If the distribution is unimodal and symmetrical, the 95% HDI is equivalent to the central 95% region of the posterior (Gelman

et al., 2013). Where it is useful to compare measures to a value of 0—for example, when examining differences between aided and unaided performance, or between observed data and model predictions—the reported statistics also include the proportion of the estimated posterior distribution that lies above or below 0 (Kruschke, 2013). Values are reported with the nomenclature $x\% < 0 < y\%$. For example, $1\% < 0 < 99\%$ indicates that 1% of the posterior distribution lies below 0, and 99% lies above. We describe an effect as credible if the 95% HDI on the difference between conditions does not overlap 0, and we describe an effect as decisive if more than 99% of the posterior distribution on difference scores falls to one side of 0 (cf. Jeffreys, 1961; Wetzels et al., 2011).

Results

Table 3-1 presents participants' mean hit and false alarm rates for the unaided and aided conditions of Experiments 1–3. The gray bars of Figure 3-3 present the corresponding mean values of *d'*. The gray bars of Figure 3-4 present participants' mean values of the bias measure *c* in the automation-aided conditions of Experiments 1–3, contingent on the aid's binary judgment. Dotted lines in Figures 3-3 and 3-4 present model-predicted values. Results for Experiment 1 appear in the left data column of the table and left panels of the figures.

Data were excluded from four participants in Experiment 1 who failed to achieve an unaided *d'* score of at least 0.5, suggesting a failure to understand or comply with the instructions. Including these participants' data in the analyses below did not change the pattern of results.

Sensitivity. Automation-aided *d'* decisively exceeded unaided *d'*, $M_{diff} = 0.48$, 95% HDI [0.20, 0.75], $0\% < 0 < 100\%$, confirming that assistance from the aid improved participants' sensitivity.

To assess model performance, analyses compared observed *d'* scores from the automation-aided conditions to the model-predicted scores based on the participants' unaided sensitivity. Mean model error scores (predicted scores minus observed scores) are presented in the text, with 95% HDIs. The two models that took into account the aid's graded evidence judgments, the OW model, M_{err} = 1.06, 95% HDI [0.86, 1.28], 0% < 0 < 100%, and the UW model, M_{err} = 0.97, 95% HDI [0.75, 1.19], 0% < 0 < 100%, both decisively overestimated participants' automation-aided sensitivity, as did the three most efficient of the binary-cue models, M_{err} = 0.75, 95% HDI [0.54, 0.96], 0% < 0 < 100% for the optimal CC model, M_{err} = 0.52, 95% HDI [0.31, 0.72], 0% < 0 < 100% for the NN model, and M_{err} = 0.37, 95% HDI [0.18, 0.56], 0% < 0 < 100% for the BD model. In contrast, the CF model decisively underestimated participants' aided sensitivity, M_{err} = -0.23, 95% HDI [-0.43, -0.03], 99% < 0 < 1%. Observed sensitivity did not differ credibly from the predictions of the PM model, M_{err} = 0.16, 95% HDI [-0.04, 0.36], 6% < 0 < 94%.

Table 3-1
Mean Hit and False Alarm Rates and 95% HDIs (in Brackets) for the Unaided and Aided Conditions of Experiments 1, 2, and 3.

	Experiment 1		Experiment 2		Experiment 3	
	Unaided	Aided	Unaided	Aided	Unaided	Aided
Hit rate	0.82 [0.78, 0.87]	0.90 [0.87, 0.92]	0.83 [0.80, 0.87]	0.89 [0.86, 0.91]	0.82 [0.79, 0.86]	0.90 [0.86, 0.93]
False alarm rate	0.14 [0.11, 0.17]	0.10 [0.08, 0.12]	0.13 [0.10, 0.15]	0.08 [0.06, 0.10]	0.14 [0.10, 0.18]	0.10 [0.07, 0.13]

Note. HDI = Highest-density interval.

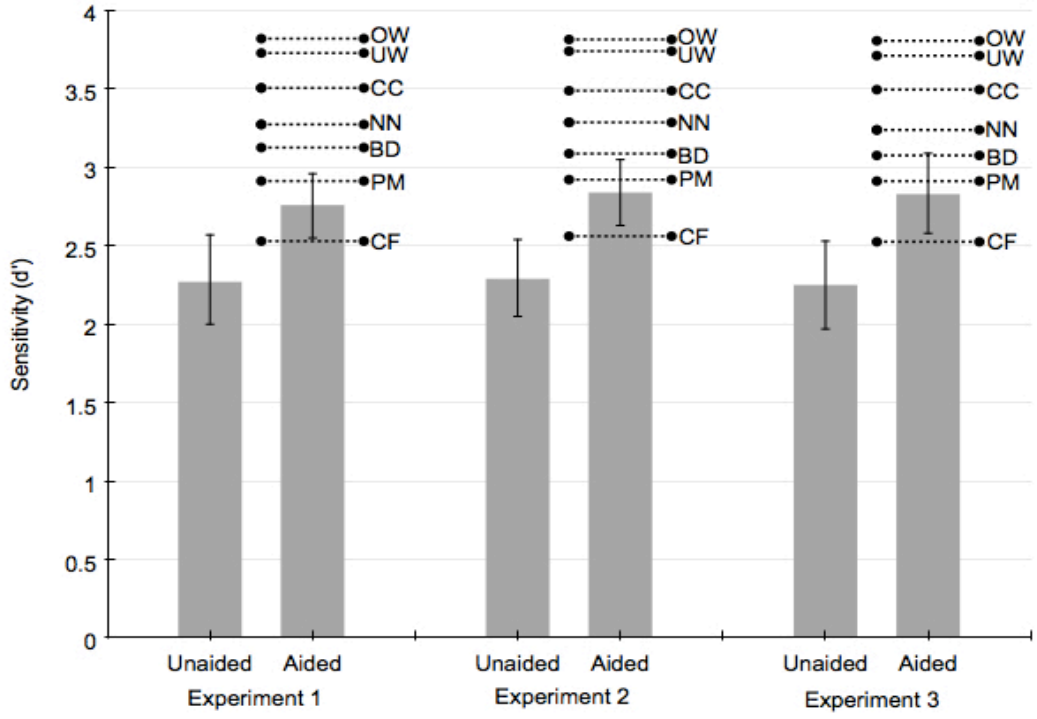

Figure 3-3. Mean *d'* values (gray bars) and model predictions (dotted lines) for Experiments 1, 2, and 3. Error bars indicate 95% highest-density intervals.

Bias. Observed levels of automation-aided sensitivity fell closest to the predictions of the PM model, which holds that the operator defers to the aid with a probability equal to the aid's average reliability. One interpretation of this finding is that participants were in fact using a PM strategy. An alternative possibility, however, is that participants were using a different strategy, but one which happened to mimic the sensitivity of the PM model. As a further test of the models, analyses compared the participants' automation-aided response bias, contingent on the aid's judgment, to the predictions of the NN, CF, optimal CC, and PM models. Note that the predicted bias for trials on which the aid provided a *Yes* judgment is negative infinity under the NN model, and is therefore not shown in Figure 3-4.

As expected, observed bias was decisively more liberal when the aid gave a *Yes* judgment than when it gave a *No* judgment, M_{diff} = 1.26, 95% HDI [1.03, 1.49], 0% < 0 <

100%, confirming that participants' responses were biased in the direction of the aid's judgments. The magnitude of the observed shifts, however, did not closely match the predictions of any of the models under consideration. For trials on which the aid issued a *Yes* judgment, observed bias was decisively more conservative than predicted by the PM model, M_{err} = -1.38, 95% HDI [-1.54, -1.22], 100% < 0 < 0%, the optimal CC model, M_{err} = -0.73, 95% HDI [-0.95, -0.50], 100% < 0 < 0%, or the CF model, M_{err} = -0.24, 95% HDI [-0.41, -0.07], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by the PM model, M_{err} = 1.36, 95% HDI [1.20, 1.51], 0% < 0 < 100%, the optimal CC model, M_{err} = 0.67, 95% HDI [0.47, 0.86], 0% < 0 < 100%, or the CF model, M_{err} = 0.24, 95% HDI [0.08, 0.40], 0% < 0 < 100%, and decisively more conservative than predicted by the NN model, M_{err} = -0.61, 95% HDI [-0.78, -0.44], 100% < 0 < 0%.

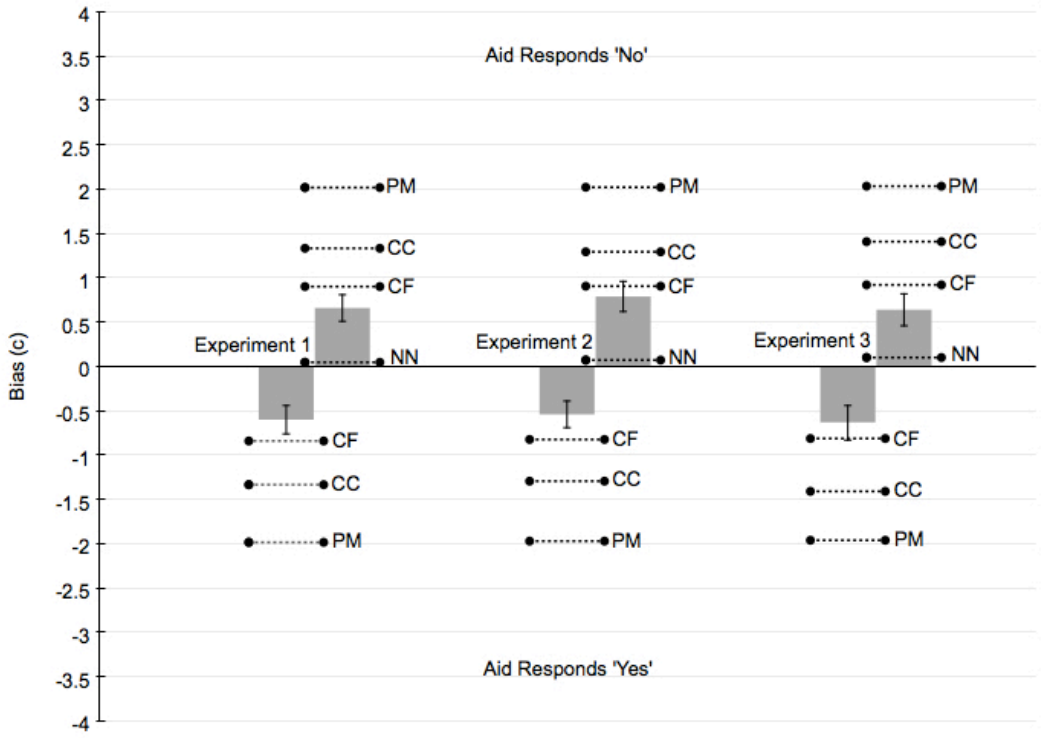

Figure 3-4. Mean of the observed (gray bars) and model-predicted (dotted lines) *c* values for Experiments 1, 2, and 3, contingent on the aid's judgment. The left bar within each panel corresponds to trials on which the aid provided a *Yes* judgment, and the right bar corresponds to trials on which the aid provided a *No* judgment. Error bars indicate 95% highest-density intervals.

Discussion

Automation-aided sensitivity fell closest to the predictions of the PM model. Aided values of *c*, however, were far less extreme than the PM model predicted, and did not match any of the models' predictions closely. But what's perhaps most surprising is that participants appear to have made little or no use of the aid's graded evidence values, substantially underperforming both the OW and UW models. In other words, aided performance was no better than could have been obtained even if the aid had provided only binary judgments. Experiment 2 pursued this result.

Experiment 2

Automation-aided participants in Experiment 1 far underperformed the OW and UW models, suggesting they made little use of the automated aid's graded evidence outputs. Experiment 2 tested this possibility by replicating the procedure of Experiment 1, but only providing participants with a binary judgment from the aid each trial. If participants derived no benefit from the aid's signal strength ratings in the first experiment, performance in Experiment 2 should match that of Experiment 1.

Method

Participants. Participants were 37 adults (mean age = 22.45 years, *SD* = 5.41, range = 17-39; 26 females, 11 males) recruited from the Flinders University of South Australia, none of whom had taken part in Experiment 1. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 45 minutes. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. Apparatus and stimuli were identical to those of Experiment 1, except that participants received only a binary, orange-or-blue judgment from the aid each trial.

Procedure. Experimental procedure and data treatment were similar to those of Experiment 1. Instructions were identical to those of Experiment 1, but modified to omit any mention of continuous values from the aid. Participants were advised simply, "The aid will provide its assessment at the start of each trial."

Results

Results for Experiment 2 appear in the middle data column of Table 3-1 and the middle panels of Figures 3-3 and 3-4. Data were excluded from one participant who failed to

achieve an unaided *d'* score of at least 0.5. Including that participants' data in the analyses below did not change the pattern of results.

Sensitivity. Automation-aided sensitivity exceeded unaided sensitivity by a mean of M_{diff} = 0.55, 95% HDI [0.35, 0.76], 0% < 0 < 100%, with a credible interval that clearly excluded 0, indicating that assistance from the aid again improved participants' sensitivity.

As in the first experiment, however, the participants' aided performance was poor relative to the predictions of the most efficient decision models under consideration. Both the OW model, M_{err} = 0.98, 95% HDI [0.80, 1.15], 0% < 0 < 100%, and the UW model, M_{err} = 0.90, 95% HDI [0.73, 1.08], 0% < 0 < 100%, decisively overestimated aided sensitivity. This is unsurprising, given that OW and UW performance is unattainable based only on binary judgments from the aid. However, aided performance also fell decisively below the levels predicted by the optimal CC model, M_{err} = 0.65, 95% HDI [0.47, 0.82], 0% < 0 < 100%, and the NN model, M_{err} = 0.45, 95% HDI [0.27, 0.62], 0% < 0 < 100%, and credibly below the levels predicted by the BD model, M_{err} = 0.25, 95% HDI [0.05, 0.44], 1% < 0 < 99%. In contrast, aided sensitivity was decisively better than predicted by the CF model, M_{err} = -0.28, 95% HDI [-0.45, -0.11], 100% < 0 < 0%, and again did not differ credibly from the predictions of the PM model, M_{err} = 0.08, 95% HDI [-0.11, 0.28], 20% < 0 < 80%.

Bias. Observed bias was again decisively more liberal when the aid responded *Yes* than when it responded *No*, M_{diff} = 1.33, 95% HDI [1.09, 1.58], 0% < 0 < 100%. For trials on which the aid issued a *Yes* judgment, observed bias was more conservative than predicted by the PM model, M_{err} = -1.43, 95% HDI [-1.58, -1.28], 100% < 0 < 0%, the optimal CC model, M_{err} = -0.75, 95% HDI [-0.99, -0.50], 100% < 0 < 0%, or the CF model, M_{err} = -0.28, 95% HDI [-0.44, -0.13], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by either the PM model, M_{err} =

1.23, 95% HDI [1.05, 1.41], 0% < 0 < 100%, or the optimal CC model, M_{err} = 0.50, 95% HDI [0.28, 0.72], 0% < 0 < 100%, and decisively more conservative than predicted by the NN model, M_{err} = -0.72, 95% HDI [-0.90, -0.53], 100% < 0 < 0%. Observed bias after a *No* from the aid did not differ credibly from that predicted by the CF model, M_{err} = 0.11, 95% HDI [-0.07, 0.30], 11% < 0 < 89%.

Cross-Experiment Comparison. Assistance from the automated aid increased participants' *d'* by 0.48 in Experiment 1 and 0.55 in Experiment 2, M_{diff} = 0.07, 95% HDI [-0.27, 0.41], 35% < 0 < 65%, giving no credible evidence that graded evidence values offered by the aid in Experiment 1 helped participants achieve higher sensitivity. In fact, though the difference was statistically negligible, automation-aided sensitivity trended higher in the second experiment than in the first.

Discussion

Experiment 2 produced a pattern of effects highly similar to that of Experiment 1, suggesting that participants made little use of the aid's graded evidence values in the first experiment. Results affirm more generally that automation-aided performance was highly inefficient, roughly matching the predictions of the PM model, but that participants' cue-contingent response bias did not closely match the predictions of any of the models tested.

Experiment 3

Experiment 1found highly inefficient automation use, even with graded estimates of signal strength from the aid. Experiment 3 sought to confirm this result with a close replication of the first experiment. As a modest extension, a scoring system was incorporated to provide an overt performance incentive and help participants better track their performance over trials.

Method

Participants. Participants were 36 adults (mean age = 22.16 years, *SD* = 4.71, range = 17-35; 30 females, 6 males) recruited from the Flinders University of South Australia, none of whom had taken part in Experiment 1 or 2. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 45 minutes. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. Apparatus and stimuli were identical to those of Experiment 1, except that a point score and running total score was provided with the feedback screen each trial.

Procedure. Experimental procedure and data treatment were similar to those of Experiment 1, except as follows. Instructions were identical to those of Experiment 1, but modified to account for the point system. Participants were advised, "You will be scored on your performance, as Marvel Mining has declared that incorrect sorting of the strains has been detrimental. You will receive 5 POINTS for every correct judgment, and you will be deducted 5 POINTS for every incorrect judgment." The total score that could be obtained in the experimental trials was 1,000 points.

At the conclusion of each trial, participants received a 1500ms feedback message of "Correct! +5, Total score = score" for all correct responses, and "Incorrect! -5, Total score = score" for all errors.

Results

Results for Experiment 3 appear in the right data column of Table 3-1 and the right panels of Figures 3-3 and 3-4.

Sensitivity. As in the first two experiments, assistance from the automated aid decisively improved participants' *d'*, M_{diff} = 0.58, 95% HDI [0.33, 0.83], 0% < 0 < 100%. Again, though, aided performance was highly inefficient. The OW, M_{err} = 0.97, 95% HDI [0.76, 1.19], 0% < 0 < 100%, UW, M_{err} = 0.88, 95% HDI [0.66, 1.10], 0% < 0 < 100%, optimal CC, M_{err} = 0.66, 95% HDI [0.44, 0.87], 0% < 0 < 100%, and NN models, M_{err} = 0.40, 95% HDI [0.19, 0.62], 0% < 0 < 100%, all decisively overestimated aided sensitivity, and the CF model was once more the only model to decisively underestimate it, M_{err} = -0.31, 95% HDI [-0.52, -0.10], 100% < 0 < 0%. Although the BD model again tended to overestimate aided sensitivity, M_{err} = 0.24, 95% HDI [0.00, 0.49], 3% < 0 < 97%, the difference between its predictions and observed performance in this case just failed to reach 95% credibility. As in the earlier experiments, however, observed performance fell closest to the predictions of the PM model, M_{err} = 0.08, 95% HDI [-0.17, 0.31], 26% < 0 < 74%.

Bias. As expected, observed bias was decisively more liberal when the aid responded *Yes*, than when it responded *No*, M_{diff} = 1.27, 95% HDI [0.98, 1.57], 0% < 0 < 100%. For trials on which the aid issued a *Yes* judgment, observed bias was decisively more conservative than predicted by either the PM, M_{err} = -1.33, 95% HDI [-1.52, -1.13], 100% < 0 < 0%, or optimal CC model, M_{err} = -0.78, 95% HDI [-1.13, -0.42], 100% < 0 < 0%. Observed bias trended more liberal than predicted by the CF model, M_{err} = -0.18, 95% HDI [-0.38, 0.01], 96% < 0 < 4%, though the difference was just short of credible. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by either the PM, M_{err} = 1.39, 95% HDI [1.21, 1.57], 0% < 0 < 100%, the optimal CC, M_{err} = 0.77, 95% HDI [0.43, 1.10], 0% < 0 < 100%, or CF model, M_{err} = 0.28, 95% HDI [0.09, 0.47], 0% < 0 < 100%, and decisively more conservative than predicted by the NN model, M_{diff} = -0.54, 95% HDI [-0.75, -0.34], 100% < 0 < 0%.

Cross-Experiment Comparison. Assistance from the automated-aid increased participants' *d'* by 0.48 in Experiment 1 and 0.58 in Experiment 3, M_{diff} = 0.10, 95% HDI [-0.27, 0.47], 29% < 0 < 71%, giving little evidence that the point system of Experiment 3 improved participants' automation use. This does not imply that with more data the modest performance difference between experiments might not become credible, or that stronger incentives or different feedback might not induce more efficient automation use, but it does lend confidence that the effects seen in Experiment 1 are generally robust.

Meta-Analysis

To estimate the discrepancies between observed data and model predictions more precisely, we combined the data of all three experiments and repeated the analyses reported above on the aggregated data.

Consistent with the conclusions above, aggregated sensitivity was decisively higher in the aided condition than in the unaided condition, M_{diff} = 0.54, 95% HDI [0.40, 0.67], 0% < 0 < 100%, but was nonetheless highly inefficient. The five most efficient models under consideration all decisively overestimated automation-aided sensitivity, M_{err} = 1.00, 95% HDI [0.89, 1.12], 0% < 0 < 100% for the OW model, M_{err} = 0.92, 95% HDI [0.80, 1.03], 0% < 0 < 100% for the UW model, M_{err} = 0.69, 95% HDI [0.57, 0.80], 0% < 0 < 100% for the optimal CC model, M_{err} = 0.46, 95% HDI [0.34, 0.57], 0% < 0 < 100% for the NN model, and M_{err} = 0.29, 95% HDI [0.17, 0.41], 0% < 0 < 100% for the BD model, and only the CF model decisively underestimated it, M_{err} = -0.27, 95% HDI [-0.38, -0.16], 100% < 0 < 0%. As above, the PM model came closest to matching observed performance levels. With the additional statistical resolution allowed by the aggregated data set, however, the discrepancy between the model's predictions and observed performance approached 95% credibility, M_{err} = 0.11, 95% HDI [-0.01, 0.22], 4% < 0 < 96%.

For trials on which the aid issued a *Yes* judgment, aggregated bias data were decisively more conservative than the predictions of either the PM, M_{err} = -1.38, 95% HDI [-1.47, -1.28], 100% < 0 < 0%, optimal CC, M_{err} = -0.75, 95% HDI [-0.91, -0.59], 100% < 0 < 0%, or CF model, M_{err} = -0.23, 95% HDI [-0.33, -0.14], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, aggregated bias data were decisively more liberal than the predictions of either the PM, M_{err} = 1.33, 95% HDI [1.23, 1.43], 0% < 0 < 100%, optimal CC, M_{err} = 0.64, 95% HDI [0.50, 0.79], 0% < 0 < 100%, or CF model, M_{err} = 0.21, 95% HDI [0.11, 0.31], 0% < 0 < 100%, and decisively more conservative than the predictions of the NN model, M_{err} = -0.62, 95% HDI [-0.73, -0.52], 100% < 0 < 0%.

In summary, when data were aggregated across experiments, aided sensitivity fell closest to the predictions of the PM model, but differed from them with borderline credibility. Conditionalized bias data remained inconsistent with any of the models under consideration.

Model Comparisons

The results above imply that the PM model may be useful as a heuristic for roughly predicting automation-aided sensitivity, but that participants likely did not employ the PM strategy, or any of the other parameter-free or fixed-parameter strategies under consideration, to make aided decisions. This allows that the data may instead be most compatible with a suboptimal CC model (Robinson & Sorkin, 1985), under which participants make automation-assisted judgments by shifting their response criterion in the direction stipulated by the aid's decision, but to an inadequate degree. However, the analyses above did not test the performance of the suboptimal CC model, and thus provide no direct evidence in support of the model. They also considered sensitivity and bias data separately, rather than jointly. We therefore conducted a model-fitting analysis to compare the performance of the CC model to that of the other models discussed above.

Method

Models were fit using an MCMC Bayesian estimation. Because the empirical data indicated that participants made little use of the aid's graded judgments, only models that relied exclusively on binary cues from the aid were considered. Four models were compared: a CC model, a variant of the CF/PM models that we will call the *discrete-state deferment* model, the BD model, and the NN model. Unaided sensitivity in all four cases was estimated using the hierarchical signal detection model described by Lee and Wagenmakers (2014). At the top level, the model assumes population distributions of sensitivity (*d'*) and criterion (*c*) values. At the level below, it assumes that individual participants render judgments using an equal-variance Gaussian signal detection model with *d'* and *c* values sampled from the population distributions. Finally, individual participants' *d'* and *c* values are reparameterized as hit and false alarm rates, and used to predict raw hit and false alarm counts from a binomial distribution. Population distributions of *d'* and *c* are assumed to be described by normal distributions, with vague priors on their means and standard deviations (means ~ N[0, .00001]; standard deviations ~ 1/Γ[.001, .001]).

Aided sensitivity was estimated differently across the four models. All four models assumed that participants made their own judgments in the aided condition with the same sensitivity as in the unaided condition, and that participants received correct judgments from the aid on 93% of all trials. The models differed in the manner by which they combined the participants' and aid's judgments. The CC model (Robinson & Sorkin, 1985) treated participants' response criteria as free parameters, estimating separate values for trials on which the aid responded *No* and trials on which the aid responded *Yes*. It therefore subsumed the optimal and suboptimal CC models: cue-contingent criterion values that matched the normative values would signal optimal performance, and values that deviated from normative would signal suboptimal performance. The model assumed

that criteria for *Yes* and *No* trials were normally distributed with the same prior distributions as the criteria for unaided trials.

In the discrete-state deferment model, participants resolved disagreements with the aid by deferring to the aid's judgment with a fixed probability. The probability of deferring to the aid was treated as a free parameter described by a beta distribution at the population level. The beta distribution is defined on the interval [0, 1], and is characterized by two parameters (Kruschke, 2015). In the parameterization used here, these parameters were the mode, ω, and concentration, κ, of the distribution. A value of $\kappa = 2$ produces a uniform distribution on the interval [0, 1]. Higher values produce more peaked distributions. The model therefore subsumed the CF and PM models discussed above: a distribution of deferment probabilities peaked tightly around a mode of 0.50 would indicate behavior consistent with the CF model, and a distribution peaked tightly around 0.93 would indicate behavior consistent with the PM model. The parameters ω and κ were assigned vague priors (both $\sim \Gamma[.001, .001]$).

Finally, the NN model assumed that an aided participant issued a *No* response only in the event that both the aid and the participant reached independent judgments of *No*, and the BD model assumed that decisions in aided blocks were made by whichever agent, human or aid, had higher sensitivity.

Each simulation employed four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step, leaving a total of 50,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less. Model performance was compared using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & van Der Linde, 2002), a measure that rewards a model for goodness-of-fit but penalizes it for complexity (Myung & Pitt, 1997). Smaller

values denote a better-fitting model. As a rule of thumb, a difference of DIC in the range of 3 to 7 is regarded as considerable evidence in favor of the better-fitting model (Spiegelhalter et al., 2002).

Results

Of the four models under comparison, the NN produced clearly the worst performance, *DIC* = 4120, followed by the BD, *DIC* = 2652. The discrete-state deferment model performed better, producing a DIC of 2377. The mean posterior value of ω, the modal estimated rate with which participants deferred to the aid, was 0.50, 95% HDI [0.01, 0.96], a value superficially consistent with a CF strategy. However, the mean estimated posterior value of κ, the concentration of deferment rates around the modal value, was 2.00, 95% HDI [2.00, 2.00], indicating that deferment rates across participants were very close to uniformly distributed between 0 and 1. Accordingly, the 95% HDI on ω spanned almost the full range of values between 0 and 1. Model fits thus revealed no tendency for participants to cluster around any particular deferment rate, that is, no consistency in automation use across participants. This implies that fixing the value of κ at 2 should improve the model's DIC, reducing model complexity without sacrificing goodness-of-fit. Consistent with this, running the model with κ as a fixed parameter of value 2 produced a DIC of 2374, better than was achieved by treating κ as a free parameter.

The CC model produced a DIC of 2375, nearly equivalent to that for the discrete-state deferment model with fixed κ. The estimated means of the participants' cued criterion values were less extreme than optimal, both for trials on which the aid issued a *Yes* judgment, M = -0.54, 95% HDI [-0.70, -0.39] for observed c vs. M = -1.19, 95% HDI [-1.28, -1.11], for optimal c, M_{diff} = 0.65, 95% HDI [0.48, 0.81], 0% < 0 < 100%, and for trials on which the aid issued a *No* judgment, M = 0.60, 95% HDI [0.45, 0.76] for observed c vs. M = 1.19, 95% HDI [1.10, 1.28] for optimal c, M_{diff} = -0.59, 95% HDI [-0.76, -0.42], 100% < 0 < 0%. HDIs around the differences between observed and

optimal criterion values clearly excluded 0, indicating that a tendency toward overly conservative criterion shifts was highly consistent across participants.

In total, results suggest that data were roughly indifferent between the discrete-state deferment model with κ fixed at 2, and a suboptimal, overly-conservative CC model. As discussed below, other considerations tilt in favor of the CC model over the discrete state model.

General Discussion

Of the seven fixed-parameter or parameter-free models considered above, the PM model most closely predicted participants' automation-aided sensitivity. Conditionalized on the aid's judgments, however, automation-aided response bias was inconsistent with any of the seven models. Thus, despite the rough match between the observed sensitivity data and the predictions of the PM model, participants do not seem to have used a PM strategy, or in fact to have used any of the fixed-parameter or parameter-free strategies tested.

How, then, did participants reach their automation-aided decisions? Model comparisons were effectively indifferent between a suboptimal CC model and a discrete-state deferment model that subsumes the CF and PM models as special cases. The suboptimal CC model, as explained above, assumes that participants made automation-assisted decisions by shifting their response criterion in the direction stipulated by the aid, but to an inadequate degree. The discrete state model assumes that participants resolved disagreements with the aid by deferring to the automation's judgments with some fixed probability. The models differ functionally in that the CC model implies that a decision maker is more likely to override the aid's recommendation when she is highly confident in her own judgment, for example, on trials when a signal is especially strong. In contrast, the discrete-state model holds that the decision maker is equally likely to override the aid whether or not she is confident in her own judgment. This suggests that future work may be better able to

distinguish the models empirically by examining participants' automation usage across different levels of signal strength.

Until more decisive empirical tests can be conducted, considerations of plausibility (Myung & Pitt, 1997; Spiegelhalter et al., 2002) may best adjudicate between the suboptimal CC and discrete-state deferment models, and seem to favor the suboptimal CC account. The discrete-state model achieved its best fit by assuming that deferment rates across participants were uniformly distributed between 0 and 1. In other words, it posited no consistency across individuals in the tendency to depend on the automation. The suboptimal CC model, in contrast, implied a consistent pattern of behavior across individuals, with HDIs on cue-contingent criteria indicating that participants were unanimously too conservative in their automation dependence. Although decision makers can most certainly differ in their willingness to depend on an automated aid (e.g., Szalma & Taylor, 2011), the possibility that they show no consistent tendencies at all seems unlikely, lending credence to the suboptimal CC model here. The tendency toward inadequate criterion shifts following a cue from the automated decision aid is also consistent with the more general 'sluggish beta' phenomenon (Chi & Drury, 1998; Neyedli, Hollands, & Jamieson, 2011; Wang et al., 2009), a tendency for decision makers in signal detection tasks to adjust their criterion less than they should in response to manipulations of signal rates and event payoffs. These various considerations tentatively suggest that the optimal CC model offers a more plausible account of automation usage than the discrete-state deferment model, even if both produced similar DICs.

As discussed above, other research has also inferred a suboptimal CC strategy from participants' automation-aided sensitivity and criteria (Elvers & Elrif, 1997; Meyer, 2001; Robinson & Sorkin, 1985; Wang et al., 2009). However, the present results go beyond earlier findings by demonstrating that the participants' suboptimal criterion choice produced sensitivity that approached the predictions the PM model. Further research will be necessary to generalize this

pattern across different forms of signal detection task and varying levels of aid reliability, and to identify markers of individual differences (e.g., Merritt & Ilgen, 2008) that allow some users to consistently attain higher benchmarks of automation-aided efficiency than others. But preliminarily, the data imply that, knowing the *d'* of an unaided operator and the *d'* of an automated aid, system designers can use the PM model to roughly predict the operator's aided sensitivity. These predictions can in turn inform analyses of the costs and benefits of building and deploying automated aids.

The tendency for decision makers to disuse decision aids that are not perfectly reliable is of course well-established (Parasuraman & Riley, 1997; Wickens & Dixon, 2007). The finding that participants used automated aids so inefficiently is especially notable here, though, because if used well, the aid's graded strength judgments in Experiments 1 and 3 could have enabled performance well above even the optimal CC level. In fact, aided performance was no better in the first and third experiments than in the second, which offered only binary judgments from the aid.

Data do not make clear why decision makers used the aid's graded cues so inefficiently. Achieving optimal performance would have been challenging in multiple ways. First, participants would have had to know, implicitly or explicitly, the statistical properties of their own sensory representations corresponding to blue-dominant and orange-dominant stimuli. Second, they would have had to know the analogous statistical properties of the aid's evidence distributions. Third, they would have had to know how much better or worse their sensitivity was than the aid's. Armed with all of this knowledge, finally, the participants would have had to calculate an appropriately weighted average of their own judgment and the aid's each trial.

Given these heavy demands, the failure to match the performance of the OW model is unsurprising; researchers have long recognized that limits on information and information-processing abilities place bounds on human cognition that can prevent human decision makers from

reaching putatively normative performance (Simon, 1955; Tversky & Kahneman, 1974). Nonetheless, human decision makers can at least approximate the performance of a linear cue combination rule (Einhorn, Kleinmuntz, & Kleinmuntz, 1979), and though they tend not to weight cues optimally (e.g., Johnson, Cavanagh, Spooner, & Samet, 1973; Montgomery, 1999, 2001; Montgomery & Sorkin, 1996), their deviations from normative weighting are likely to have modest effects on performance (Dawes & Corrigan, 1974; Wainer, 1976). Comparing the predictions of the OW and UW models above, for instance, shows that an equal-weighting rule for combining human and automation judgments would have approached the performance of the optimal-weighting rule. It therefore seems unlikely that participants' inefficient use of graded evidence values was caused by an inability to estimate proper weights for combining judgments. Moreover, even when high cognitive load or imperfect information make a linear decision rule difficult or impracticable, decision makers can often find nonlinear heuristic strategies that allow near-normative performance (Gigerenzer & Gaissmaier, 2011; Hogarth & Karelaia, 2007). In the current tasks, for example, a simple heuristic rule for using the aid's graded judgments to resolve disagreements between the human and aid might have been to defer to the aid when it produced a relatively high evidence value but to override it otherwise.

Despite these possibilities, the data gave little indication that participants made use of the aid's graded evidence judgments. Rather, the null differences between Experiment 2 and Experiments 1 and 3, and the highly inefficient levels of automation-aided performance seen in all three experiments, suggest that participants disregarded the aid's graded outputs entirely. This may indicate a tendency for participants to minimize effort expenditure (Bettman, Johnson, & Payne, 1990), sacrificing decision accuracy in order to forego the short-term cognitive costs of encoding and remembering the aid's graded assessment each trial. Further research will be necessary to determine whether instruction (Sedlmeier & Gigerenzer, 2001), changes to the format in which

information from the aid is presented (Bisantz, 2013; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Todd & Benbasat, 1994), or other task and display manipulations might reduce the effort needed to use the aid's graded assessments and induce more efficient human-automation performance.

This concludes the current published paper.

CHAPTER 4:
STUDY 3

The following series of three experiments are currently under review in a manuscript submitted for publication to Human Factors, entitled *No effect of cue format on automation dependence in an aided signal detection task*. The version of the manuscript presented here is a revised version of the original submission. All authors were involved in the formulation of the study concept and design, and data analysis. Megan Bartlett collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

In addition, the findings from this experiment have been presented at:

Bartlett, M. L., & McCarley, J. S. (2017. November). *Quantifying suboptimal automation use*. Paper presented at the Defence Human Sciences Symposium, Adelaide, Australia.

Bartlett, M. L., & McCarley, J. S. (2017. November). *No effect of information format on performance in an aided signal detection task*. Poster presented at the Australasian Cognitive Neuroscience Society Conference, Adelaide, Australia.

Introduction

Human operators performing tasks such as identifying enemies on the battlefield (Wang, Jamieson, & Hollands, 2009) or detecting fraud in financial reports (Bell & Carcello, 2000) are often assisted by automated decision aids. Due to the probabilistic nature of the judgments they make, however, these aids will rarely operate with perfect accuracy. Unfortunately, human operators tend to interact with an imperfectly reliable aid in a suboptimal way, either *misusing it* by acting on its judgments uncritically, or *disusing* it by ignoring or underweighting its judgments (Parasuraman, 2000; Parasuraman & Riley, 1997).

In contexts like those mentioned above, the decision aid's task is often to sort events into one of two discrete categories (friend or enemy; fraudulent or not) on the basis of uncertain evidence, acting as a classic signal detection system (Green & Swets, 1966; Macmillan & Creelman, 2005). In the simplest case, the aid provides the operator with a binary, yes-or-no judgment each trial. Robinson and Sorkin (1985) proposed a *contingent criterion* (CC) model of human interaction with an aid of this form. The model assumes that the aid and operator work in sequence, with the operator establishing his or her own response criterion contingent on the aid's judgment. To make use of the aid's judgments, the operator establishes a more liberal criterion following a positive judgment from the aid than following a negative judgment. Optimal criterion placement is determined by the aid's predictive validity. As measured by the statistic beta (β^*), the operator's optimal level of bias following a judgment from the aid, is,

$$\beta^* = p(\text{no signal} \mid \text{diagnosis}) / p(\text{signal} \mid \text{diagnosis}),$$

where diagnosis indicates the aid's judgment, positive or negative.

The operator achieves best-possible aided performance by using β^* each trial. Commonly, though, participants appear to adjust their response criterion inadequately (e.g., Robinson & Sorkin, 1985; Wang et al., 2009), achieving performance below ideal levels. Comparing observed performance to the predictions of various models, for instance, Bartlett and McCarley (2017) found

that automation-aided sensitivity not only failed to reach the level predicted by the optimal CC model, it also underperformed a much simpler, *best decides* (BD) model (Bahrami et al., 2010; Denkiewicz, Rączasek-Leonardi, Migdal, & Plewczynski, 2013), which assumes that the operator bases their decision on the judgments of the decision maker (aid or operator) that is on average most reliable. Performance fell closest to the predictions of two highly inefficient models, the *coin flip* (CF) (Bahrami et al., 2010) and *probability matching* (PM) (Bartlett & McCarley, 2017) models. Both of these assume that the aid and the operator reach independent binary judgments and that when those judgments differ, the disagreement is resolved randomly. In the CF model, the operator defers to the aid with a probability of 0.50. In the PM model, the operator defers to the aid with a probability equal to the aid's average reliability level. Both models predict only modest gains from the aid's assistance.

Graded Aids

Decision aids that provide operators a non-binary, graded assessment of evidence strength may offer a method to attenuate to the problem of poor automation dependence, allowing the possibility of better performance than can be achieved with a binary aid and potentially mitigating the risk of suboptimal human-automation interaction (Sorkin, Kantowitz, & Kantowitz, 1988; St. John & Manes, 2002; Woods, 1995). By providing an assessment of any certainty inherent in their judgments, a graded aid allows for a better-calibrated collaborative decision. The *optimal weighting* (OW) and *uniform weighting* (UW) models (Bahrami et al., 2010; Bartlett & McCarley, 2017; Sorkin, Hays, & West, 2001) offer strategies for collaboration in cases where a graded aid provides an estimate of certainty (or conversely, uncertainty) on a continuous scale. Both models assume that the operator averages his or her own estimates of signal strength with the aid's to reach a judgment each trial. Under the OW model, the collaborative decision is based on the average of the operator's and aid's independent judgments, with each judgment weighted proportional to the agent's average

sensitivity. This model produces statistically ideal performance under aided conditions. Under the UW model, the collaborative decision is based on an unweighted average of the two judgments. When the aid and operators have equivalent sensitivity, the UW strategy is equivalent to the OW strategy. Otherwise, the UW model predicts poorer sensitivity. Both the OW and UW strategies, using a continuous measure of evidence from an aid, will tend to outperform the optimal CC strategy, which uses only binary judgments from the aid (Bartlett & McCarley, 2017). See Table 4-1 for equations for the OW, UW, CC, BD, PM, and CF models.

Table 4-1
Equations for the OW, UW, CC, BD, PM, and CF Models.

Model	Equations
OW	$d' = \sqrt{{d'_{\text{operator}}}^2 + {d'_{\text{aid}}}^2}$
UW	$d' = \frac{d'_{\text{operator}} + d'_{\text{aid}}}{\sqrt{2}}$
CC	$HR = HR_{\text{aid}}\left(HR_{\text{operator}\mid\text{"Yes"}}\right) + (1 - HR_{\text{aid}})\, HR_{\text{operator}\mid\text{"No"}}$ $FAR = FAR_{\text{aid}}\left(FAR_{\text{operator}\mid\text{"Yes"}}\right) + (1 - FAR_{\text{aid}})\, FAR_{\text{operator}\mid\text{"No"}}$
BD	$d' = max\left(d'_{\text{operator}}, d'_{\text{aid}}\right)$
PM	$HR = R_{\text{aid}} \times HR_{\text{aid}} + (1 - R_{aid}) \times HR_{\text{operator}}$ $FAR = R_{\text{aid}} \times FAR_{\text{aid}} + (1 - R_{aid}) \times FAR_{\text{operator}}$
CF	$HR = 0.5\left(HR_{\text{operator}} + HR_{\text{aid}}\right)$ $FAR = 0.5\left(FAR_{\text{operator}} + FAR_{\text{aid}}\right)$

Note. d'_{operator} = sensitivity of the unaided human operator; d'_{aid} = sensitivity of the automated aid; HR_{aid} = hit rate of the automated aid; $HR_{\text{operator}\mid\text{“Yes”}}$ = hit rate of the unaided human operator given a Yes judgment from the aid; $HR_{\text{operator}\mid\text{“No”}}$ = hit rate of the unaided human operator given a No judgment from the aid; FAR_{aid} = false alarm rate of the automated aid; $FAR_{\text{operator}\mid\text{“Yes”}}$ = false alarm rate of the unaided human operator given a Yes judgment from the aid; $FAR_{\text{operator}\mid\text{“No”}}$ = false alarm rate of the unaided human operator given a No judgment from the aid; R_{aid} = aid’s average

reliability rate; $HR_{operator}$ = hit rate of the unaided human operator; $FAR_{operator}$ = false alarm rate of the unaided human operator.

But while some studies have found that graded aids can improve decision making, increasing operators' ability to distinguish signal and noise events (e.g., Ragsdale, Lew, Dyre, & Boring, 2012; Sorkin et al., 1988; St. John & Manes, 2002; Wiczorek & Manzey, 2014; Wiczorek, Manzey, & Zirk, 2014), other research has suggested that they do not always do so (Sorkin et al., 1998; Wickens & Colcombe, 2007; Wiczorek & Manzey, 2014). In line with the latter set of results, Bartlett and McCarley (2017) found that participants in an aided signal detection task relied solely on the aid's binary judgments, effectively ignoring the aid's raw graded evidence outputs. Thus, even with analog assessments from the aid, automation-assisted performance fell in the range predicted by the CF and PM strategies of information integration.

Information Format

Why did participants fail to take advantage of the aid's graded assessments? One possibility is that they found the analog judgments difficult to properly interpret. The aid in Bartlett and McCarley's (2017) experiments produced judgments using a conventional Gaussian, equal-variance signal detection model (Macmillan & Creelman, 2005), and the graded judgments provided to participants were simply the raw evidence values sampled by the aid each trial. To use these cues optimally, participants would have had to know the probability distributions of evidence values corresponding to signal and noise stimuli (Bahrami et al., 2010). To properly assess evidence values from the aid, that is, participants would have had to know, at least implicitly, that the evidence values corresponding to the two categories of events were normally distributed with means of ±1.5 and standard deviations of 1.0. Perhaps it is unsurprising, then, that participants failed to use the aid's graded cues with perfect efficiency. Participants could have used the graded cues to improve their performance merely by noting that more extreme values corresponded to a higher probability

that the aid's binary diagnosis was correct. The resulting performance gains might have been too modest to motivate participants to adopt this strategy, however; decision makers often sacrifice accuracy in exchange for a decrease in task effort (Beach & Mitchell, 1978; Russo & Dosher, 1983).

This possibility suggests that automation-aided performance might improve if graded judgments from the aid are rendered in a format participants find easier to interpret. Past work has shown that Bayesian reasoning, for example, can be improved through changes to the format in which information is represented (Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Reasoning is poor when statistical information is presented in the form of normalized frequencies or probabilities (e.g., If *p*[disease] = .01, *p*[positive test | disease] = .90, *p*[positive test | no disease] = .1, what is the probability that someone who tests positive has the disease?), apparently because decision makers tend to underweight base rate information (in this example, the marginal probability, *p*[disease]) (Bar-Hillel, 1990; Kahneman & Tversky, 1973). Performance is better when the same information is presented in the form of natural frequencies (e.g., Out of 1000 people, 10 will have a given disease. Of those 10, 9 will test positive. Of the 990 without the disease, 99 will test positive. If a new patient tests positive, what is the probability that she has the disease?). Gigerenzer and Hoffrage (1995) note that reasoning with natural frequencies demands fewer computational steps than reasoning with probabilities, and does not require the decision maker to explicitly know or account for base rate information. Natural frequencies improve reasoning, that is, by reducing decision makers' informational and computational demands.

Other work has shown similar effects in the context of an automation-aided signal detection task, demonstrating that the format in which statistical information is conveyed can improve automation usage strategy. Participants in a study by Botzer, Meyer, Bak, and Parmet (2010) performed a mock quality control task, judging whether to classify products as intact or flawed

based on the recommendation of an automated aid. To help them optimize their performance, participants were allowed to adjust the aid's threshold in response to changes of signal base rate and the decision payoff matrix. Between groups of participants, however, the influence of threshold settings was described in differing ways. Some participants received predictive values, indicating the probability of a fault conditioned on the aid's diagnosis, while others received diagnostic values, indicating the probability of a positive diagnosis from the aid, conditioned on the presence of a fault. Notably, predictive values, like natural frequencies, inherently vary with signal base rate, sparing the decision maker the need to effortfully incorporate base rates in their judgments. Accordingly, predictive values induced threshold settings closer to optimal than did diagnostic values (Botzer et al., 2010).

The present experiments built on the work of Bartlett and McCarley (2017), by examining the influence of display format on participants' ability to interpret and use cues from a signal detection aid. Participants viewed orange and blue random dot images, and were asked to determine each trial which color was dominant (Voss, Rothermund, & Voss, 2004). They performed the task alone, or with assistance from an automated decision aid. In Experiments 1 & 2, the aid rendered its judgment as a binary diagnosis accompanied by an estimate of signal strength in the form of either a raw evidence value, a likelihood ratio, or a confidence rating. The likelihood ratio in favor of a blue-dominant stimulus given a sampled evidence value x sampled is,

$$p(x \mid \text{blue-dominant stimulus}) / p(x \mid \text{orange-dominant stimulus}).$$

As noted above, in order to render an optimal judgment based on a given value x in the raw cue condition, participants were required to know the forms, means, and standard deviations of the distributions from which the aid samples its evidence values. The likelihood ratio, in contrast, summarizes the evidence provided by x in a manner that requires no knowledge of the underlying evidence distributions (Balakrishnan & Ratcliff, 1996). In other words, the likelihood ratio contains

all of the evidence provided by *x* for adjudicating between the two possible states (Berger et al., 1988; Pawitan, 2013). A likelihood ratio can also be understood in much the same way as a report of natural frequencies. For example, a cue from the aid giving a likelihood ratio of 9:1 in favor of a noise stimulus implies that on average, of ten hypothetical cases with the aid's current level of evidence, nine will correspond to a noise stimuli and one will correspond to a signal. A likelihood ratio therefore can be interpreted with less background information and fewer cognitive operations than can a raw evidence value. The confidence rating format transformed the evidence samples further, mapping log likelihood ratios to integer values on a scale of 0-100. Confidence ratings were thus monotonic with the likelihood ratios from which they were derived (truncated at extreme values), but were rendered in a form that participants might find more familiar or simpler still. Experiments 1 & 2 were identical except that participants received no feedback at the conclusion of each trial in the first experiment, and received feedback on their response accuracy at the conclusion of each trial in the second experiment.

In a third experiment, participants performed the task with assistance from an aid that provided its diagnosis in the form of a binary judgment, a verbal expression of probability, or a verbal expression of probability highlighted within a visuospatial display of a range of verbal expressions. As a converging assessment of performance, automation-aided sensitivity in all three experiments was benchmarked against the predictions of various statistical models of collaborative decision making, as described above, ranging from optimal to highly inefficient (Bartlett & McCarley, 2017).

Finally, in keeping with other studies of decision aid use (e.g., Merritt, Heimbaugh, LaChapell, & Lee, 2013; Merritt, Lee, Unnerstall, & Huber, 2015; Wang et al., 2009; Wiczorek, 2017), regression analyses in all three experiments examined the relationships between a variety of

individual difference measures, including trust in the automated aid, perceived accuracy of the aid, and self-perceived accuracy, and a measure of automation dependence.

Hypotheses

Hypothesis 1: Sensitivity (*d'*) will be higher when participants are assisted by an aid than when they are unassisted. This effect will serve as a manipulation check that participants could and did use the automated aid to improve their performance.

*Hypothesis 2: S*ensitivity will be higher for participants assisted by an aid that provides its diagnosis in the form of a likelihood ratio, than for those assisted by an aid that provides its diagnosis in the form of either a raw evidence value, or a confidence rating.

*Hypothesis 3: S*ensitivity will be higher for participants assisted by an aid that provides its diagnosis in the form of a verbal expression of probability highlighted within a visuospatial display of a range of verbal expressions, than for those assisted by an aid that provides its diagnosis in the form of either a binary judgment, or a verbal expression of probability.

Experiment 1

Method

Participants. Participants were 96 adults (mean age = 21.95 years, *SD* = 4.63, range = 18-38; 75 females, 21 males), recruited from Flinders University. Participants were compensated with $10.00 AUD or 1 hour of course credit for an experimental session that lasted approximately 60 min. All were fluent in English, and were screened for normal color vision and normal or corrected-to-normal visual acuity.

This research complied with the tenets of the Declaration of Helsinki and was approved by the Social and Behavioural Research Ethics Committee at Flinders University. Informed consent was obtained from all participants.

Apparatus and Stimuli. The experimental task was controlled by software written in PsychoPy (Peirce, 2007). Stimuli were presented on a 23-inch Samsung monitor with a resolution

of 1,920 x 1,080 pixels and a 120 Hz refresh rate. Participants sat approximately 60 cm from the monitor, with viewing distance unconstrained.

Stimuli were blue and orange random dot images (256 x 256 pixels), presented against a white background. Each stimulus was either blue- or orange-dominant. In the blue-dominant stimuli, each pixel was randomly assigned the color blue with a probability of 0.52 or the color orange with a probability of 0.48. In the orange-dominant stimuli, those probabilities were reversed. Figure 4-1 shows sample orange-dominant (leftmost) vs blue-dominant (rightmost) stimulus images.

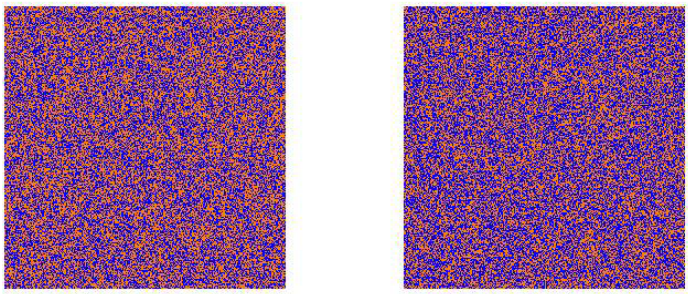

Figure 4-1. Sample orange-dominant (leftmost) vs blue-dominant (rightmost) stimulus images.

Automated Aid. On some blocks of trials, participants were assisted by an automated decision aid that judged whether the stimulus presented each trial had been generated using the parameters of the blue-dominant or orange-dominant distribution. The aid's judgments were generated using a standard equal-variance Gaussian signal detection model (Macmillan & Creelman, 2005). For trials on which the true stimulus categorization was blue-dominant, the aid's evidence value was sampled from a Gaussian distribution with a mean of -1.5 and a standard deviation of 1. For trials on which the true stimulus categorization was orange-dominant, the aid's evidence value was sampled from a Gaussian distribution, with a mean of 1.5 and a standard deviation of 1. Thus, the *d'* of the aid was 3. The aid transformed evidence values into binary judgments using an unbiased response threshold, offering a judgment of blue-dominant if the

evidence value sampled for a given trial was less than 0 and a judgment of orange-dominant if the evidence value sampled was greater than 0. The unbiased criterion combined with a *d'* of 3 produced an average accuracy rate of 93%.

The aid rendered its binary judgment with an estimate of signal strength in one of three forms. In the raw value condition, the aid's estimate of signal strength was simply the absolute value *x* of the sampled evidence value. An example of the aid's judgment for the raw value condition is, "Aid judges: orange; Measure = 1.22." In the likelihood ratio and confidence rating conditions, the aid's raw judgments were transformed and presented as likelihood ratios and confidence ratings, respectively.

For the likelihood ratio condition, the aid's raw sampled evidence value *x* was converted each trial into a likelihood ratio,

$p(x \mid \text{orange-dominant stimulus}) / p(x \mid \text{blue-dominant stimulus})$.

Values were displayed in the format, "Likelihood = *a*:1", where $a \geq 1$. For example, evidence values for orange-dominant trials were normally distributed with a $\mu = 1.5$ and $\sigma = 1$, giving $p(x = 1.22 \mid \text{orange-dominant stimulus}) \approx .383$. Conversely, evidence values for blue-dominant trials were normally distributed with a $\mu = -1.5$ and $\sigma = 1$, giving $p(x = 1.22 \mid \text{blue-dominant stimulus}) \approx .001$. A raw evidence value of 1.22 would therefore have been represented in the likelihood ratio condition as a cue reading, "Aid judges: orange; Likelihood = 39:1."

To generate confidence ratings, a transformation was necessary to convert the aid's sampled evidence values into ratings on a scale of 0–100. For this purpose, the likelihood ratio was calculated as above, and then converted to a log likelihood ratio. The resulting values were arbitrarily truncated at 10, corresponding to a likelihood ratio of approximately 22046 or a raw evidence value of roughly ±4.1. Finally, values were multiplied by 10, putting them on a 0-100 scale. A raw evidence value of 1.22, for example, corresponding to a likelihood ratio of roughly

39:1, would have been transformed to a confidence value of log(39) × 10 ≈ 37. This value would have been presented as a cue reading, “Aid judges: blue; Confidence = 37%.” Note that this transformation preserved the information in the likelihood ratio displays, up to a log likelihood ratio value of 10.

Participants in all cue conditions were informed that higher values indicated stronger evidence. Because they were generally not expected to have had extensive formal training in statistics, however, they were not provided any additional information about the distribution of evidence values.

Individual Difference Measures. Trust in the automated aid, perceived accuracy of the aid, and self-perceived accuracy were measured using items after Merritt (2011). Trust in the aid was assessed using a 6-item self-report measure. Participants responded on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). An example item is, “I believe the aid is a competent performer.” The maximum score that can be obtained is 30.

Perceived accuracy of the aid was assessed with an item asking, “Out of the 100 trials you completed WITH assistance from the automated aid, how many times did you think the aid was correct?” Participants filled in the blank, “I think the aid was correct on ___ out of 100 trials.” The maximum score that can be obtained is 100.

Self-perceived accuracy was assessed with an item asking, “Out of the 100 trials you completed WITHOUT assistance from the automated aid, how many times did you think that you were correct?” Participants filled in the blank, “I think I was correct on ___ out of 100 trials.” The maximum score that can be obtained is 100.

Procedure. Participants performed a two-alternative forced choice (2AFC) task requiring them to classify stimulus images as coming from blue- or orange-dominant distributions. A cover story asked the participants to imagine themselves as geologists sorting samples of a mineral into

blue and orange strains. The instructions informed them, “Unfortunately, the two strains are difficult to tell apart. Both are speckled blue and orange. The only difference visually is that one strain tends to have a little more orange, and the other tends to have a little more blue. However, there is a lot of overlap in their appearance, and it is almost impossible to sort them with 100% accuracy by eye.” Participants were asked to decide each trial if the sample they were presented was orange-dominant or blue-dominant, and to provide an estimate of their decision confidence. They rendered responses by clicking on a six-point rating scale underneath the stimulus image. Responses on the scale were labeled, *Definite blue*, *Probable blue*, *Guess blue*, *Guess orange*, *Probable orange*, and *Definite orange*. Rating scale data were collected to perform anticipated additional analyses.

Participants were also told that on some trials, they would be assisted by an automated decision aid that would provide a blue or orange judgment along with an estimate of certainty. Instructions read, “The aid works by testing the chemical properties of the sample, and then assessing whether the sample is more likely to be ORANGE or BLUE. However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is correct 93% of the time and incorrect 7% of the time.” Figure 4-2 shows sample cue displays from the raw value (leftmost panel), likelihood ratio (middle panel), and confidence rating (rightmost panel) cue conditions of Experiments 1 & 2.

Figure 4-2. Sample cue displays from the raw value (leftmost panel), likelihood ratio (middle panel), and confidence rating (rightmost panel) cue conditions of Experiments 1 & 2.

Participants in the raw value condition were informed, “To help you predict whether it is right or wrong, the aid will give its assessment along with a numeric rating each trial. A higher rating means that the aid is more likely to be correct.” Participants in the confidence rating condition were informed, “To help you predict whether it is right or wrong, the aid will give its assessment along with a confidence rating each trial. A higher confidence rating means that the aid is more likely to be correct.” Participants in the likelihood ratio condition were informed, “To help you predict whether it is right or wrong, the aid will give its assessment along with a likelihood ratio each trial. A higher ratio means that the aid is more likely to be correct.” Regardless of cue condition, participants were advised, “You should use the aid to help you make your decisions, but be aware that you are free to disagree with it any time you wish. Use your own best judgement.”

Figure 4-3 shows the sequence of events within an unaided trial for Experiment 1. Each trial was preceded by a message reading, “Click the circle below to start the next trial.” Each aided trial comprised a 500-ms blank interval, a 1,500-ms screen displaying the

automated aid's diagnosis, another 500-ms blank interval, and then the stimulus display, which remained onscreen until the participant's response. The sequence of events on unaided trials was identical to that on aided trials, except that the aid's diagnosis was replaced by a neutral message, "Waiting for image." Presentation of the aid's diagnosis before the stimulus display allowed participants time to attend to the diagnosis carefully, and ensured that the diagnosis and stimulus arrived in the same order in which the CC model presumes they are processed (though see Wiegmann, McCarley, Kramer, & Wickens, 2006, for evidence that automation dependence is similar regardless of the order in which cue and stimulus are presented). Other models make no presumption as to the order of processing. The neutral message served to match the sequence and timing of events across the aided and unaided blocks.

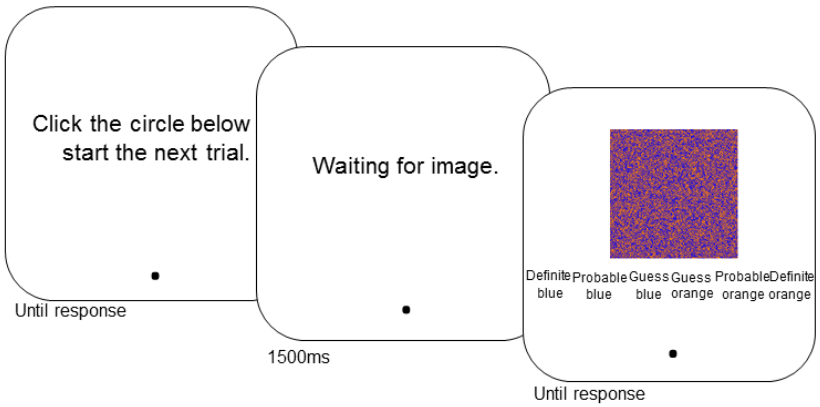

Figure 4-3. The sequence of events within an unaided trial for Experiment 1.

Each session comprised a block of 50 unaided practice trials followed by a block of 50 aided practice trials, then a block of 100 unaided experimental trials and a block of 100 aided experimental trials, with the order of the experimental blocks counterbalanced across participants. Each trial, the stimulus category was selected randomly and with equal probability from among the

two options (i.e., blue- or orange-dominant), and the stimulus image was then generated randomly. At the conclusion of the experimental task, participants were administered with the individual differences questionnaires.

Participants were allowed to rest between blocks. Participants were randomly assigned to the cue format conditions in equal numbers in a between-subjects design.

Analysis

The experiment was originally planned to allow calculation of additional measures, beyond those reported here, based on the analysis of confidence rating data. However, a substantial number of participants failed to use the full range of confidence levels in the response scale, precluding the intended analysis of the rating data. Data were therefore collapsed across confidence ratings to produce binary responses. For analysis, orange-dominant stimuli were treated as signal events and blue-dominant stimuli as noise events. For clarity of exposition, we refer to orange and blue judgments as *yes* and *no* judgments, respectively.

Hit rates and false alarm rates were calculated from the participants' binary responses, and data were converted to signal detection measures of sensitivity and bias, *d'* and *c* (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). A prior of 0.5 was added to the raw response frequency value in each cell of the 2 x 2 signal detection theory (SDT) matrix for each participant to correct for perfect hit and false alarm rates (Hautus, 1995). Data from practice trials were excluded from analysis.

Data analysis employed Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) sampling procedure (Kruschke, 2013, 2015; Lee & Wagenmakers, 2014). This approach begins by assuming a prior distribution on a parameter value of interest, then updates the prior through probabilistic sampling to approximate the posterior distribution on parameter values in light of the observed data. Analyses were conducted using the package

JAGS (Plummer, 2015) in R (http://www.r-project.org). *d'* scores were analyzed in a 2 (Block: unassisted vs. assisted) × 3 (Format: raw value, likelihood ratio, confidence rating) mixed design, with participant treated as an additive effect (Kruschke, 2015). Effects were assumed to follow normal distributions with vague priors on their means and standard deviations. Following Kruschke (2015),

$$Y_{\text{Block, Format, Participant}} \sim N(a0 + a_{\text{Block}} + a_{\text{Format}} + a_{\text{Block} \times \text{Format}} + a_{\text{Participant}}, \sigma_y^2)$$

$$\sigma_y \sim U(SD/1000, SD*1000)$$

$$a0 \sim N(M, [100 \times SD]^2)$$

$$a_{\text{Block}} \sim N(0, \sigma^2_{\text{Block}})$$

$$a_{\text{Format}} \sim N(0, \sigma^2_{\text{Format}})$$

$$a_{\text{Block} \times \text{Format}} \sim N(0, \sigma^2_{\text{Block} \times \text{Format}})$$

$$a_{\text{Participant}} \sim N(0, \sigma^2_{\text{Participant}})$$

$$\sigma_{\text{Block}}, \sigma^2_{\text{Format}}, \sigma^2_{\text{Block} \times \text{Format}}, \sigma_{\text{Participant}} \sim \Gamma(\alpha, \beta)$$

$$\alpha = SD/2$$

$$\beta = 2 * SD$$

where $Y_{\text{Block, Format, Participant}}$ is the *d'* score for a given participant in a given cell of the design, *a0* is the estimated grand mean *d'*, a_{Block} is the effect of Block, a_{Format} is the effect of Format, $a_{\text{Block} \times \text{Format}}$ is the effect of the Block x Format interaction, $a_{\text{Participant}}$ is the participant effect, σ_y is the estimated standard deviation of the normally distributed *d'* scores, *M* is the grand mean of the observed *d'* scores, and *SD* is the standard deviation of the observed *d'* scores. Use of the data sample mean and standard deviation to set parameters of the priors ensured that the prior distributions were scaled appropriately to the data (Kruschke, 2015). The use of vague priors ensured that the analysis did not commit *a priori* to strong conclusions, allowing the observed data to dominate the posterior distribution. Predictions for the various

models of automation use were based on the hierarchically-estimated grand mean *d'* score for unaided performance on each iteration of the sampling procedure.

Parameter estimation was based on four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step in order to reduce sample autocorrelation, leaving a total of 100,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less, indicating satisfactory convergence of the MCMC chains (Kruschke, 2015).

Descriptive statistics reported include the mean and 95% highest density intervals (HDI) for the estimated posterior distributions (Kruschke, 2013). The 95% HDI is the region that contains 95% of the posterior distribution mass, and within which all values have higher probability than any values outside the region. If the distribution is unimodal and symmetrical, the 95% HDI is equivalent to the central 95% region of the posterior (Gelman et al., 2013). Where it is useful to compare measures to a value of 0—for example, when examining differences between aided and unaided performance, or between observed data and model predictions—the reported statistics also include the proportion of the estimated posterior distribution that lies above or below 0 (Kruschke, 2013). Values are reported with the nomenclature $x\% < 0 < y\%$. For example, $1\% < 0 < 99\%$ indicates that 1% of the posterior distribution lies below 0, and 99% lies above. We describe an effect as credible if the 95% HDI on the difference between conditions does not overlap 0, and we describe an effect as decisive if more than 99% of the posterior distribution on difference scores falls to one side of 0 (cf. Jeffreys, 1961; Wetzels et al., 2011).

To test more directly for effects of cue format on performance, an additional analysis compared the fit of the full model described above to the fit of a reduced model excluding the main effect of Format and the interaction of Block × Format. Model fits were assessed

using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & van Der Linde, 2002), a measure that rewards a model for goodness-of-fit but penalizes it for complexity (Myung & Pitt, 1997). Smaller values denote a better-fitting model. A lower DIC value for the reduced model would therefore give evidence that, accounting for parsimony, the data favored a null effect of Format and Block × Format. As a rule of thumb, a difference of DIC in the range of 3 to 7 is regarded as considerable evidence in favor of the better-fitting model (Spiegelhalter et al., 2002).

Results

Data were excluded and replaced from four participants in Experiment 1 who failed to achieve an unaided *d'* score of at least 0.5, suggesting a failure to understand or comply with the instructions. Including these participants' data in the analyses below did not change the pattern of results. An additional two participants were excluded due to technical difficulties. This left 30 participants in each cue format condition.

Sensitivity. Table 4-2 presents participants' mean hit and false alarm rates, *d'* and *c* scores for the raw value, likelihood ratio, and confidence rating conditions of Experiment 1. The gray bars of Figure 4-4 present the hierarchically-estimated group mean values of *d'*. Dotted lines in Figure 4-4 present mean model-predicted values.

Automation-aided *d'* bordered on being credibly higher than unaided *d'* in the raw value condition, M_{diff} = 0.28, 95% HDI [0.00, 0.54], 3% < 0 < 97%, and was decisively higher than unaided *d'* in the likelihood ratio, M_{diff} = 0.37, 95% HDI [0.12, 0.64], 0% < 0 < 100%, and confidence rating conditions, M_{diff} = 0.39, 95% HDI [0.14, 0.66], 0% < 0 < 100%. Further analyses compared performance across groups to examine the effects of cue format on automation usage. Unaided sensitivity did not differ credibly between the likelihood ratio and raw value groups, M_{diff} = -0.13, 95% HDI [-0.48, 0.18], 79% < 0 < 21%,

confidence rating and raw value groups, M_{diff} = -0.10, 95% HDI [-0.44, 0.22], 72% < 0 < 28%, or likelihood ratio and confidence rating groups, M_{diff} = 0.04, 95% HDI [-0.28, 0.36], 41% < 0 < 59%, suggesting that groups were similar in their baseline performance levels. Crucially, aided groups likewise failed to differ credibly between the likelihood ratio and raw value groups, M_{diff} = -0.04, 95% HDI [-0.37, 0.28], 59% < 0 < 41%, confidence rating and raw value groups, M_{diff} = 0.02, 95% HDI [-0.30, 0.35], 46% < 0 < 54%, or likelihood ratio and confidence rating groups, M_{diff} = 0.06, 95% HDI [-0.25, 0.39], 36% < 0 < 64%.

Table 4-2
Mean Hit and False Alarm Rates, d' and c Scores with 95% HDIs [in brackets] for the Raw Value, Likelihood Ratio, and Confidence Rating Conditions of Experiment 1.

	Raw		Likelihood		Confidence	
	Unaided	Aided	Unaided	Aided	Unaided	Aided
Hit rate	0.85 [0.79, 0.89]	0.88 [0.83, 0.92]	0.86 [0.80, 0.91]	0.89 [0.84, 0.93]	0.82 [0.76, 0.87]	0.88 [0.83, 0.92]
False alarm rate	0.12 [0.08, 0.17]	0.10 [0.06, 0.14]	0.17 [0.11, 0.23]	0.12 [0.08, 0.17]	0.12 [0.08, 0.17]	0.10 [0.06, 0.14]
d'	2.19 [1.95, 2.45]	2.47 [2.23, 2.71]	2.06 [1.81, 2.29]	2.43 [2.18, 2.67]	2.09 [1.85, 2.33]	2.49 [2.25, 2.73]
c	0.05 [-0.13, 0.24]	0.03 [-0.15, 0.22]	-0.07 [-0.27, 0.12]	-0.04 [-0.24, 0.14]	0.11 [-0.07, 0.31]	0.04 [-0.14, 0.24]

Note. HDI = Highest-density interval.

Consistent with these effects, data showed no credible Block × Format interaction, as the difference between unaided and aided conditions did not differ credibly between the likelihood ratio and raw value conditions, M_{diff} = 0.10, 95% HDI [-0.22, 0.46], 29% < 0 < 71%, confidence rating and raw value conditions, M_{diff} = 0.12, 95% HDI [-0.18, 0.50], 25% < 0 < 75%, or likelihood ratio and confidence rating conditions, M_{diff} = 0.02, 95% HDI [-0.30, 0.36], 45% < 0 < 55%. Overall, performance was similar whether the aid provided its diagnosis in the form of a raw value, likelihood ratio, or confidence rating.

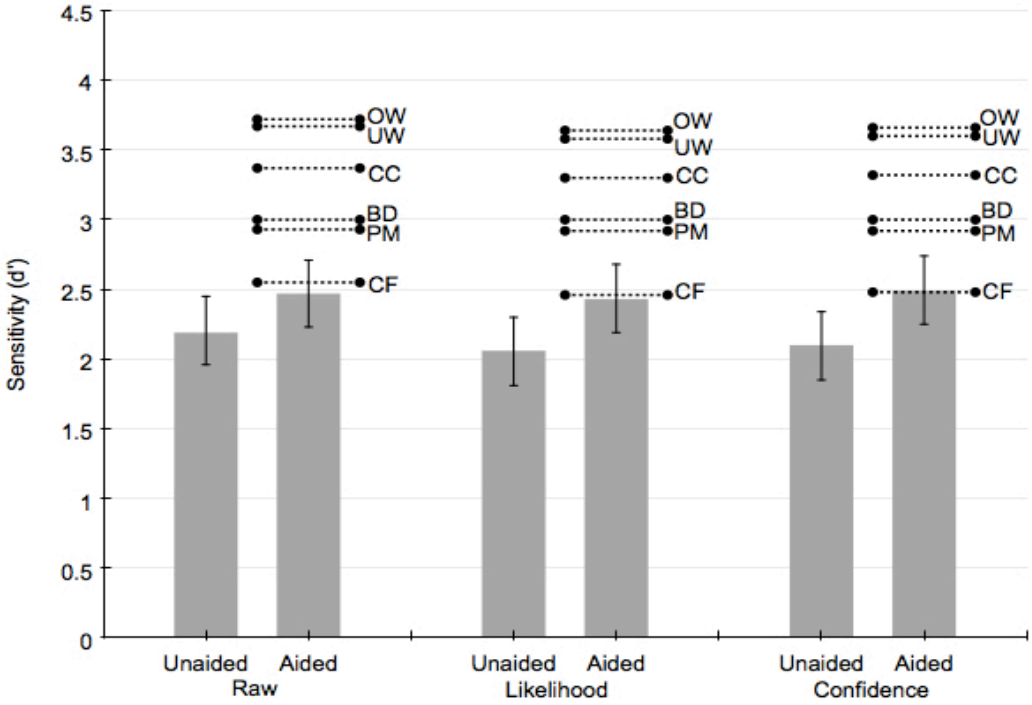

Figure 4-4. Hierarchically-estimated group mean values (gray bars) and model-predicted values (dotted lines) of *d'* for the raw value, likelihood ratio, and confidence rating cue conditions of Experiment 1. Error bars indicate 95% highest-density intervals.

Giving more direct evidence against an influence of cue format, model fits favored a model excluding a main effect of Format and an interaction of Block × Format over the full model, which included those effects, DIC = 825.69 for the reduced model and DIC = 830.71 for the full model.

Model Predictions. To assess model performance, analyses compared observed *d'* scores from the automation-aided conditions to the model-predicted scores. Mean model error scores (predicted scores minus observed scores) are presented in the text, with 95% HDIs. Since no differences were found between aided formats, we collapsed across conditions for analysis of model error scores.

The OW model decisively overestimated participants' automation-aided sensitivity, M_{err} = 1.21, 95% HDI [1.06, 1.36], 0% < 0 < 100%, as did the UW model, M_{err} = 1.15, 95%

HDI [1.00, 1.31], 0% < 0 < 100%, the optimal CC model, M_{err} = 0.86, 95% HDI [0.71, 1.01], 0% < 0 < 100, the BD model, M_{err} = 0.53, 95% HDI [0.38, 0.69], 0% < 0 < 100%, and the PM model, M_{err} = 0.46, 95% HDI [0.30, 0.61], 0% < 0 < 100%. Observed sensitivity did not differ credibly from the predictions of the CF model, M_{err} = 0.03, 95% HDI [-0.12, 0.19], 34% < 0 < 66%.

Individual Difference Measures. Since no differences were found between aided formats, we collapsed across conditions for analysis of the self-report measures.

Bayesian linear regression analyses were carried out on the relationship between self-perceived accuracy and efficiency, trust and efficiency, perceived aid accuracy and trust, and finally, perceived aid accuracy and efficiency. Efficiency provides an index of how far observed group performance falls from statistically ideal performance (Sorkin et al., 2001; Tanner & Birdsall, 1958), normalizing automation-aided sensitivity for each observer by unaided sensitivity. In the context of automation-aided decision making, efficiency, η, is,

$$\eta = \left(\frac{d'_{aided}}{d'_{OW}}\right)^2$$

where d'_{aided} is the sensitivity of the human-automation team, and d'_{OW} is the sensitivity of the ideal group.

Participants who perceived the aid to be highly accurate reported greater trust in the aid, r = 0.65, 95% HDI [0.48, 0.81], 0% < 0 < 100%, and showed greater efficiency, r = 0.48, 95% HDI [0.30, 0.67], 0% < 0 < 100%. Participants who trusted the aid also displayed greater efficiency, r = 0.40, 95% HDI [0.21, 0.60], 0% < 0 < 100%. No credible relationship was evident between self-perceived accuracy and efficiency, r = 0.09, 95% HDI [-0.13, 0.29], 21.30% < 0 < 78.70%.

Discussion

Aided performance was similar whether the aid provided its diagnosis in the form of a raw value, a likelihood ratio, or a confidence rating. Performance under all three formats was highly inefficient, falling well short of the levels that could have been attained with just binary cues from the aid. These results suggest that participants' inefficient use of the raw signals from the graded aid, here and in earlier work (Bartlett & McCarley, 2017), was not caused by a failure to properly infer the probabilistic distribution of raw evidence values. Automation-aided performance was poor even when participants were provided a direct estimate of the relative likelihood of the cue's diagnosis being correct, obviating the need for any knowledge of the raw evidence distributions.

On average, in fact, automation-aided sensitivity fell closest to the predictions of the least efficient model of collaborative decision making that was considered, the CF model. This result differs from that of Bartlett and McCarley (2017), who found performance closer to the predictions of the PM model in a task very similar to that used here. The experiments differed, though, in one potentially important aspect of procedure. Specifically, whereas participants in Bartlett and McCarley's experiments received a message after each trial to tell them whether their judgment had been right or wrong, participants in the current experiment did not. Past work has found operators show closer-to-optimal automation interactions when performance feedback is provided than when it is not (Beck, Dzindolet, & Pierce, 2007; Dzindolet, Pierce, Peterson, Purcell, & Beck, 2002). Experiment 2 therefore investigated whether post-trial feedback would help participants collaborate with the aid more efficiently, or might engender performance differences across aid formats.

Experiment 2

Experiment 2 replicated the procedure of Experiment 1, but provided participants trial-by-trial feedback.

Method

Participants. Participants were 90 adults (mean age = 21.65 years, *SD* = 4.99, range =18-39; 72 females, 18 males) recruited from Flinders University, none of whom had taken part in Experiment 1. Participants were compensated with $10.00 AUD or 1 hour of course credit for an experimental session that lasted approximately 60 min. All were fluent in English, and were screened for normal color vision and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. Apparatus and stimuli were identical to those of Experiment 1.

Procedure and Analysis. Experimental procedure, data treatment, and analysis were identical to those of Experiment 1, except that at the conclusion of each trial, participants received a 1,500-ms feedback message of either “Correct!” or “Incorrect!

Results

Sensitivity. Table 4-3 presents participants’ mean hit and false alarm rates, *d’* and *c* scores for the raw value, likelihood ratio, and confidence rating conditions of Experiment 2. The gray bars of Figure 4-5 present the hierarchically-estimated group mean values of *d’*. Dotted lines in Figure 4-5 present model-predicted values.

Automation-aided *d’* was decisively higher than unaided *d’* in the raw value, M_{diff} = 0.45, 95% HDI [0.20, 0.70], 0% < 0 < 100%, likelihood ratio, M_{diff} = 0.45, 95% HDI [0.19, 0.70], 0% < 0 < 100%, and confidence rating conditions, M_{diff} = 0.42, 95% HDI [0.16, 0.66], 0% < 0 < 100%.

Again, unaided sensitivity did not differ credibly between the likelihood ratio and raw value groups, M_{diff} = -0.03, 95% HDI [-0.32, 0.25], 58% < 0 < 42%, confidence rating

and raw value groups, M_{diff} = 0.04, 95% HDI [-0.24, 0.33], 40% < 0 < 60%, or likelihood ratio and confidence rating groups, M_{diff} = 0.07, 95% HDI [-0.21, 0.37], 32% < 0 < 68%. Aided sensitivity likewise failed to differ credibly between the likelihood ratio and raw value groups, *M* = -0.03, 95% HDI [-0.32, 0.25], 59% < 0 < 41%, confidence rating and raw value groups, *M* = 0.00, 95% HDI [-0.28, 0.29], 49% < 0 < 51%, or likelihood ratio and confidence rating groups, *M* = 0.04, 95% HDI [-0.25, 0.32], 40% < 0 < 60%.

Table 4-3
Mean Hit and False Alarm Rates, d' and c Scores with 95% HDIs [in brackets] for the Raw Value, Likelihood Ratio, and Confidence Rating Conditions of Experiment 2.

	Raw		Likelihood		Confidence	
	Unaided	Aided	Unaided	Aided	Unaided	Aided
Hit rate	0.88 [0.84, 0.91]	0.92 [0.89, 0.94]	0.90 [0.87, 0.92]	0.93 [0.90, 0.95]	0.89 [0.86, 0.92]	0.93 [0.91, 0.95]
False alarm rate	0.08 [0.06, 0.11]	0.06 [0.04, 0.08]	0.11 [0.08, 0.14]	0.07 [0.05, 0.09]	0.09 [0.07, 0.12]	0.07 [0.05, 0.09]
d'	2.54 [2.32, 2.76]	2.99 [2.77, 3.21]	2.51 [2.28, 2.72]	2.95 [2.73, 3.17]	2.57 [2.36, 2.80]	2.99 [2.77, 3.21]
c	0.08 [-0.03, 0.21]	0.05 [-0.06, 0.18]	-0.03 [-0.16, 0.08]	-0.01 [-0.14, 0.10]	0.01 [-0.10, 0.13]	-0.02 [-0.15, 0.09]

Note. HDI = Highest-density interval.

As the pairwise comparisons above suggest, data showed no credible evidence of a Block x Format interaction, as the difference between unaided and aided conditions did not differ credibly between the likelihood ratio and raw value conditions, M_{diff} = -0.00, 95% HDI [-0.32, 0.31], 51% < 0 < 49%, confidence rating and raw value conditions, M_{diff} = -0.03, 95% HDI [-0.37, 0.27], 58% < 0 < 42%, or likelihood ratio and confidence rating conditions, M_{diff} = -0.03, 95% HDI [-0.37, 0.27], 58% < 0 < 42%.

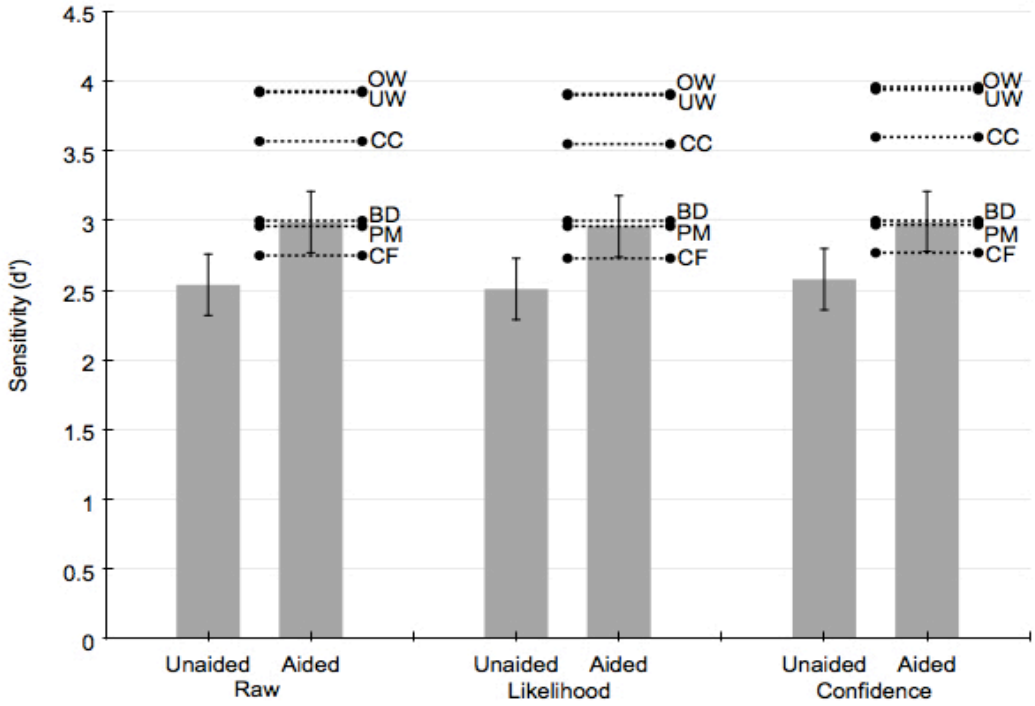

Figure 4-5. Hierarchically-estimated group mean values (gray bars) and model-predicted values (dotted lines) of *d'* for the raw value, likelihood ratio, and confidence rating cue conditions of Experiment 2. Error bars indicate 95% highest-density intervals.

Analysis of model fits again favored a model excluding the main effect of Format and the interaction of Block × Format, DIC = 783.23, over the full model incorporating those effects, DIC = 798.10.

Model Predictions. Again, since no differences were found between aided formats, we collapsed across conditions for analysis of model performance.

As in the first experiment, the OW model decisively overestimated participants' automation-aided sensitivity, M_{err} = 0.95, 95% HDI [0.80, 1.10], 0% < 0 < 100%, as did the UW model, M_{err} = 0.94, 95% HDI [0.78, 1.10], 0% < 0 < 100%, and the optimal CC model, M_{err} = 0.59, 95% HDI [0.44, 0.75], 0% < 0 < 100%. In contrast to the results of Experiment 1, however, the CF model decisively underestimated participants' aided sensitivity, M_{err} = -0.23, 95% HDI [-0.38, -0.08], 100% < 0 < 0%, and neither the predictions of the BD model,

M_{err} = 0.02, 95% HDI [-0.13, 0.16], 40% < 0 < 60%, or the PM model, M_{err} = -0.02, 95% HDI [-0.16, 0.13], 59% < 0 < 41%, differed credibly from observed performance.

Individual Difference Measures. Again, because no differences were found between aided formats, we collapsed across conditions for analysis of the self-report measures.

Participants who perceived the aid to be highly accurate reported greater trust in the aid, r = 0.44, 95% HDI [0.25, 0.63], 0% < 0 < 100%. No credible relationships were evident, however, between perceived aid accuracy and efficiency, r = -0.02, 95% HDI [-0.23, 0.19], 58.52% < 0 < 41.48%, self-perceived accuracy and efficiency, r = 0.02, 95% HDI [-0.20, 0.23], 43.73% < 0 < 56.27%, or trust and efficiency, r = 0.15, 95% HDI [-0.06, 0.36], 8.32% < 0 < 91.68%.

Discussion

Experiment 2 found automation use slightly more efficient than observed in Experiment 1, falling in the range of PM and BD model predictions, consistent with earlier evidence that feedback improves automation use (Beck et al., 2007; Dzindolet et al., 2002). Replicating the findings of Experiment 1, however, performance was similar whether the aid provided its diagnosis in the form of a raw value, likelihood ratio, or confidence rating. Experiment 3 explored two further forms of representation.

Experiment 3

Experiment 3 examined whether verbal representations of the aid's probabilistic cues (e.g., "Absolutely impossible", "Very unlikely," "Rather likely"; Bisantz, Marsiglio, & Munch, 2005) would bolster human-automation interaction. Verbal descriptors of probability are vaguer than numeric ones and more subject to individual differences in interpretation (Beyth-Marom, 1982; Budescu, Weinberg, & Wallsten, 1988; Wallsten,

Budescu, Rapoport, Zwick, & Forsyth, 1986). They do not seem to significantly compromise decision making performance, however, (e.g., Budescu et al., 1988; Budescu & Wallsten, 1990; Erev & Cohen, 1990), and because they are considered easier or more intuitive to interpret, are preferred over numeric descriptors by some decision makers (Wallsten, Budescu, Zwick, & Kemp, 1993). This suggests the possibility that decision makers in the current context might find verbal expressions of the automated aid's confidence easier to use than numeric expressions.

Experiment 3 tested this possibility by replicating the general procedure of Experiment 2 but using an aid that provided its diagnosis in the form of a binary judgment, or a binary judgment coupled with a likelihood ratio mapped onto a verbal expression of confidence. As an additional manipulation, verbal expressions were presented in either of two forms. In one case, the cue appeared as a single expression in the centre of the screen. In the alternative case, the cue appeared as a highlighted item within a scale showing all six of the possible verbal expressions, ordered by level of confidence. The latter displays were intended to provide an additional, visuospatial cue to aid participants' memory for the cue values (Darling, Allen, & Havelka, 2017; Paivio, 1990) and emphasize the confidence level of each judgment within the range of potential values. Figure 4-6 shows sample cue displays from the binary (leftmost panel), verbal (middle panel) and verbal-spatial (rightmost panel) cue conditions of Experiment 3.

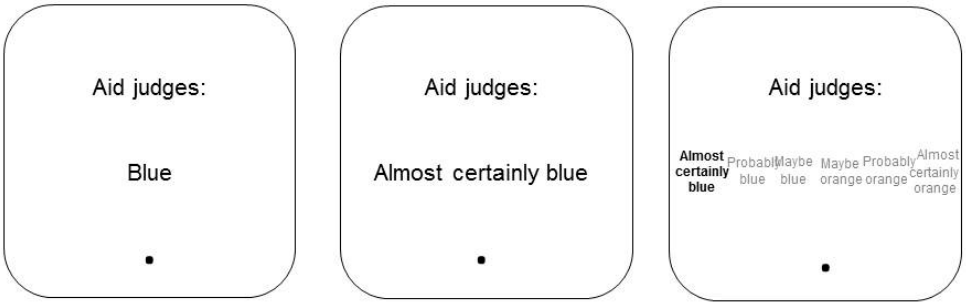

Figure 4-6. Sample cue displays from the binary (leftmost panel), verbal (middle panel) and verbal-spatial (rightmost panel) cue conditions of Experiment 3.

Method

Participants. Participants were 91 adults (mean age = 21.94 years, *SD* = 5.40, range = 18-39; 67 females, 24 males) recruited from Flinders University, none of whom had taken part in Experiments 1 or 2. Participants were compensated with 1 hour of course credit for an experimental session that lasted approximately 60 min. All were fluent in English, and were screened for normal color vision and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. Apparatus and stimuli were identical to those of Experiments 1 & 2.

Automated Aid. The aid rendered its judgment in one of three formats. In the binary cue condition, the aid's judgment was simply presented as a binary blue-or-orange judgment. In the verbal and verbal-spatial cue conditions, the aid's raw judgments were transformed and presented as verbal descriptions. First, the aid's raw sampled evidence value x was converted each trial into a likelihood ratio,

$$p(x \mid \text{blue-dominant stimulus}) / p(x \mid \text{orange-dominant stimulus}),$$

which was in turn transformed to a probability,

$$p(\text{blue-dominant stimulus} \mid x) = 1 - p(\text{orange-dominant stimulus} \mid x).$$

Finally, the *p*(blue-dominant stimulus) was mapped onto a set of verbal expressions adapted from Hamm (1991) and Bisantz et al. (2005). The verbal expressions that were used were "Almost certainly", "Probably" and "Maybe." The aid offered a judgment of "Almost certainly blue" if $p(\text{blue-dominant stimulus} \mid x) > 0.90$, a judgment of "Probably blue" if $0.90 \geq p(\text{blue-dominant stimulus} \mid x) > 0.55$, and a judgment of "Maybe blue" if $0.55 \geq p(\text{blue-dominant stimulus} \mid x) > 0.50$. Alternatively, the aid offered a judgment of "Maybe orange" if $0.50 \geq p(\text{blue-dominant stimulus} \mid x) > 0.45$, a judgment of "Probably orange" if $0.45 \geq p(\text{blue-dominant stimulus} \mid x) > 0.10$, and a judgment of "Almost certainly orange" if $p(\text{blue-dominant stimulus} \mid x) < 0.10$.

In the binary and verbal cue conditions, the aid's judgment each trial appeared by itself, in the center of the screen. In the verbal-spatial cue condition, the whole range of verbal expressions was displayed in order, from "Almost certainly blue" to "Almost certainly orange," with the expression denoting the aid's judgment being presented in black, bold-faced font, and the other remaining expressions being presented in gray font. Figure 4-6 shows sample cue displays from the binary (leftmost panel), verbal (middle panel) and verbal-spatial (rightmost panel) cue conditions of Experiment 3.

Procedure. Experimental procedure and data treatment were identical to those of Experiment 2 except that participants were asked to provide their judgment by clicking on one of two text boxes underneath the stimulus image. As participants in Experiments 1 & 2 failed to use the full rating scale, we instead used a binary response scale. Participants in the binary cue condition were advised, "To help you predict whether it is right or wrong, the aid will give its judgment each trial," while participants in the verbal and verbal-spatial cue conditions were advised, "To help you predict whether it is right or wrong, the aid will give an assessment of its confidence along with its judgment each trial." Regardless of cue condition, participants were

advised, “You should use the aid to help you make your decisions, but be aware that you are free to disagree with it any time you wish. Use your own best judgement.”

Analysis

Data treatment and analysis were the same as those of Experiments 1 & 2. *d’* scores were analyzed in a 2 (Block: unassisted vs. assisted) × 3 (Format: binary, verbal, verbal-spatial) mixed design, with participant treated as an additive effect (Kruschke, 2015).

Results

Data were excluded and replaced from one participant in the binary cue condition who failed to achieve an unaided *d’* score of at least 0.5, suggesting a failure to understand or comply with the instructions. Including this participant’s data in the analyses below did not change the pattern of results. This left 30 participants in each cue format condition.

Sensitivity. Table 4-4 presents participants’ mean hit and false alarm rates, *d’* and *c* scores for the binary, verbal, and verbal-spatial conditions of Experiment 3.The gray bars of Figure 4-7 present the hierarchically-estimated group mean values of *d’*. Dotted lines in Figure 4-7 present mean model-predicted values.

Automation-aided *d’* was decisively higher than unaided *d’* in the binary, M_{diff} = 0.35, 95% HDI [0.12, 0.58], 0% < 0 < 100%, verbal, M_{diff} = 0.46, 95% HDI [0.23, 0.69], 0% < 0 < 100%, and verbal-spatial cue conditions, M_{diff} = 0.44, 95% HDI [0.22, 0.67], 0% < 0 < 100%

Again, unaided sensitivity did not differ credibly between the verbal and binary cue groups, *M* = -0.23, 95% HDI [-0.59, 0.10], 91% < 0 < 9%, verbal-spatial and binary cue groups, *M* = -0.33, 95% HDI [-0.68, 0.03], 97% < 0 < 3%, or verbal-spatial and verbal cue groups, *M* = -0.09, 95% HDI [-0.43, 0.24], 72% < 0 < 28%. Aided sensitivity likewise failed to differ credibly between the verbal and binary cue groups, *M* = -0.13, 95% HDI [-0.47, 0.21], 77% < 0 < 23%, verbal-spatial and

binary cue groups, M = -0.24, 95% HDI [-0.59, 0.10], 91% < 0 < 9%, or verbal-spatial and verbal cue groups, M = -0.11, 95% HDI [-0.44, 0.22], 75% < 0 < 25%.

Table 4-4
Mean Hit and False Alarm Rates, d' and c Scores with 95% HDIs [in brackets] for the Binary, Verbal, and Verbal-Spatial Conditions of Experiment 3.

	Binary		Verbal		Verbal-Spatial	
	Unaided	Aided	Unaided	Aided	Unaided	Aided
Hit rate	0.92 [0.89, 0.94]	0.94 [0.92, 0.96]	0.89 [0.85, 0.92]	0.93 [0.90, 0.95]	0.88 [0.85, 0.92]	0.92 [0.89, 0.94]
False alarm rate	0.10 [0.07, 0.14]	0.07 [0.04, 0.09]	0.10 [0.07, 0.14]	0.07 [0.05, 0.10]	0.12 [0.09, 0.16]	0.08 [0.05, 0.11]
d'	2.71 [2.46, 2.97]	3.06 [2.81, 3.31]	2.48 [2.23, 2.72]	2.93 [2.69, 3.18]	2.38 [2.13, 2.63]	2.82 [2.57, 3.07]
c	-0.08 [-0.22 0.04]	-0.05 [-0.19, 0.07]	0.00 [-0.12, 0.14]	-0.02 [-0.15, 0.11]	-0.03 [-0.16, 0.09]	-0.03 [-0.16, 0.09]

Note. HDI = Highest-density interval

Consistent with these pairwise comparisons, data again gave no credible evidence of a Block × Format interaction, as the difference between unaided and aided conditions did not differ credibly between the verbal and binary conditions, M_{diff} = 0.10, 95% HDI [-0.18, 0.44], 25% < 0 < 75%, verbal-spatial and binary conditions, M_{diff} = 0.09, 95% HDI [-0.19, 0.41], 28% < 0 < 72%, or verbal-spatial and verbal conditions, M_{diff} = -0.02, 95% HDI [-0.32, 0.28], 54% < 0 < 46%.

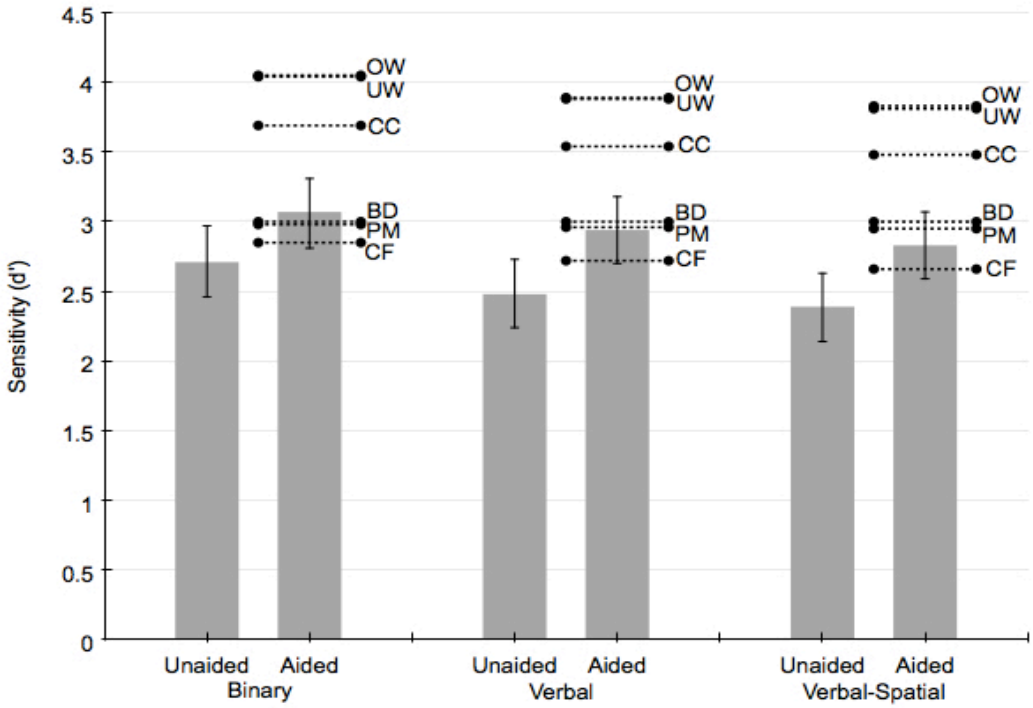

Figure 4-7. Hierarchically-estimated group mean values (gray bars) and model-predicted values (dotted lines) of *d'* for the binary, verbal, and verbal-spatial cue conditions of Experiment 3. Error bars indicate 95% highest-density intervals.

DIC scores again favoured a model excluding the main effect of Format and the Block × Format interaction, DIC = 737.62, over the full model including those effects, DIC = 755.86.

Model Predictions. Since no differences were found between aided formats, we again collapsed across conditions for analysis of model performance.

As in Experiments 1 & 2, the OW model decisively overestimated participants' automation-aided sensitivity, M_{err} = 0.98, 95% HDI [0.85, 1.11], 0% < 0 < 100%, as did the UW model, M_{err} = 0.96, 95% HDI [0.83, 1.10], 0% < 0 < 100%, and the optimal CC model, M_{err} = 0.62, 95% HDI [0.49, 0.76], 0% < 0 < 100%.

As in Experiment 2, the CF model decisively underestimated participants' aided sensitivity, M_{err} = -0.20, 95% HDI [-0.33, -0.07], 100% < 0 < 0%, and neither the predictions of the BD model, M_{err} = 0.06, 95% HDI [-0.09, 0.21], 23% < 0 < 77%, or the PM model, M_{err} = 0.02, 95% HDI [-0.13, 0.16], 39% < 0 < 61%, differed credibly from observed performance.

Individual Difference Measures. Since no differences were found between aided formats, we collapsed across conditions for analysis of the self-report measures. It is important to note, however, that OW performance would be statistically unachievable for participants in the current experiment, as the evidence assessments they received from the aid were not continuous but discretized (to either six or two levels). We are scaling performance, therefore, relative to hypothetical OW performance.

No credible relationships were evident, however, between perceived aid accuracy and trust, r = 0.12, 95% HDI [-0.09, 0.34], 12.58% < 0 < 87.42%, trust and efficiency, r = 0.12, 95% HDI [-0.10, 0.33], 13.73% < 0 < 86.27%, self-perceived accuracy and efficiency, r = -0.00, 95% HDI [-0.22, 0.20], 50.56% < 0 < 49.44%, or perceived aid accuracy and efficiency, r = -0.02, 95% HDI [-0.23, 0.20], 57.50% < 0 < 42.50%.

Discussion

Experiment 3 once again found inefficient automation use, with aided sensitivity falling in the range of PM and BD model predictions, consistent with the findings of Experiment 2. Most importantly, however, performance was similar whether the aid provided its diagnosis as a binary judgment, a single verbal expression of probability, or a verbal expression of probability mapped onto a visuospatial cue.

General Discussion

The present experiments investigated whether manipulating the format of an automated aid's cues would encourage users to collaborate with the aid more efficiently in a signal detection task. In Experiments 1 & 2, the aid rendered a binary judgment each trial along with an estimate of signal strength in the form of either a raw value, a likelihood ratio, or a confidence rating. In Experiment 3, it provided its judgments as binary cues, verbal expressions of probability, or verbal expressions of probability mapped onto a visuospatial cue. Assistance from the aid consistently improved participants' sensitivity. Aided performance was suboptimal, however, and contrary to expectations, was similar across cue formats; rendering probabilistic cues in formats that should have made them easier to interpret did not induce participants to use the aid more efficiently.

The null effects of cue format, combined with the finding that aided sensitivity fell near the levels predicted by the least efficient information integration strategies examined, buttress Bartlett and McCarley's (2017) conclusion that participants relied exclusively on the aid's binary diagnoses, ignoring its graded displays of evidence strength. The current data suggest this was true even when cues were designed to minimize the cognitive overhead involved with interpreting, combining and remembering the aid's graded judgments (Bettman, Johnson, & Payne, 1990).

The conclusion that participants ignored the aid's non-binary cues contrasts with the results of other studies that have found benefits of graded decision aids to human performance. Wiczorek and Manzey (2014) and Wiczorek et al. (2014), for instance, found that 3- and 4-level alarm systems enabled better decisions in a simulated quality control task than did a binary alarm system. Sorkin et al. (1988) likewise found higher sensitivity on an aided signal detection task when cues from the aid were presented on a 4-level scale rather than a binary scale. St. John and Manes (2002) demonstrated that target detection in a serial visual search task was faster when spatial cues were color-coded on a 6-level scale to indicate the likelihood of a target at each cued location than when

they were binary. Neyedli, Hollands, and Jamieson (2011) found that participants performing an aided combat identification task could adjust their response bias in response to cues that specified the aid's reliability on a trial-by-trial basis.

In all these cases, though, task demands differed from those of the present study. In the experiments described by Ragsdale et al. (2012), Sorkin et al. (1988), Wiczorek and Manzey (2014), and Wiczorek et al. (2014), the participants' signal detection task was embedded within a multi-task scenario. An alert from an aid therefore was not just a source of information on which to base the signal detection judgment, but was a cue for the participant to shift attention away from an ongoing task. One benefit of the graded aid may therefore have been to help operators better balance attention between concurrent tasks, knowing when it was most important to attend to a stimulus on the signal detection channel (Wiczorek & Manzey, 2014). Consistent with this speculation, Sorkin et al. (1988) found that a graded signal detection aid outperformed a binary aid only when the detection task was paired with a demanding concurrent task. When the concurrent task was easier, binary and graded aids were equally useful.

The current study also differed from some others in that it required participants to integrate judgments from the aid with their own assessments of the visual stimulus. In earlier studies demonstrating benefits of graded cues, the cues have often been presented without raw stimulus data for the participant to examine (Ragsdale et al., 2012; St. John & Manes, 2002; Wiczorek et al., 2014). Wiczorek and Manzey (2014) in fact found that the advantage of 3-level cues over binary cues to sensitivity was eliminated when participants were allowed to cross-check the aid's judgments against raw system data. This might imply that the availability of raw data encourages operators to discount or ignore graded judgments from a decision aid. A notable exception to this pattern, though, comes from Neyedli et al. (2011), whose participants made visual friend-or-foe judgments, under single-task conditions, with assistance from a combat identification aid. As

described above, participants adjusted their response bias in response to visual icons that specified the aid's reliability, taking advantage of graded cues from the aid in a way that participants in the current studies didn't. The current stimuli and procedure differed from Neyedli et al.'s in a variety of potentially important ways, including the nature of the task—color discrimination versus combat identification—and the form of the reliability cues—textual/numeric versus graphical. Further study, manipulating task and procedural characteristics like these, will be necessary to determine which of these task characteristics modulates the usefulness of graded cues.

The current results do reiterate earlier findings that performance feedback can improve the efficiency of automation use. In the absence of feedback (Experiment 1), aided sensitivity fell near the predictions of the CF model (Bahrami et al., 2010), a highly inefficient strategy of information integration. In the presence of feedback (Experiments 2 & 3), aided performance became slightly more efficient, rising to the level of the PM and BD strategies. Even with performance feedback, however, aided sensitivity remained well short of optimal. The current data do not tell us whether aided decisions might have become more efficient with extended training on the task or experience with the aid, but they do imply that performance feedback, over a short time on task, will improve automation use only modestly. Moreover, the finding that performance falls consistently in the range of the CF/PM strategies' predictions provides a heuristic for predicting the level of performance that operators might achieve in a signal detection task with the support of a decision aid.

Individual Differences

The three experiments produced inconsistent results concerning the relationship between individual difference measures and automation-aided efficiency. Regardless of feedback, participants in Experiments 1 & 2 who perceived the aid to be highly reliable also reported greater trust in the aid, replicating earlier findings (e.g., Lee & See, 2004; Merritt, 2011; Merritt & Ilgen,

2008; Wang et al., 2009). The efficiency of automation use co-varied with the perceived reliability of, and reported trust in the aid, but only in Experiment 2. No relationship was evident between self-perceived accuracy and efficiency in any experiments, and no credible correlations were found amongst individual difference measures in Experiment 3, or between the individual difference measures and efficiency of automation use. Further research will be necessary to determine whether inconsistencies in the patterns of individual differences were driven by differences in task characteristics between experiments, or were simply the result of random variation.

Constraints on Generality

Our results replicate and extend earlier findings that automation-aided decision making is highly inefficient, and suggest that confidence-graded cues from an aid may be of limited value to performance, even when they are rendered in a format designed to minimize the cognitive overhead of interpreting probabilistic information. A variety of considerations constrain the generality of these results though. Our effects were observed with a task and stimuli that were tightly controlled and abstract, using an aid of a single, fairly high sensitivity level ($d' = 3$) and cues represented as text. Replications will be necessary to determine whether similar results obtain with different and more naturalistic forms of signal detection task, with aids of lower or higher sensitivity, or with alternative cue formats. Indeed, as discussed above, earlier data have shown that participants can take advantage of graphical reliability cues in a combat identification task, and can sometimes make use of multi-level alerts in multitask environments. Work is therefore necessary to identify the task characteristics that make graded cues valuable. Additional study will be needed as well to determine whether a different participant population, more expert or perhaps better motivated, might make better of use confidence-graded cues, or interact more efficiently with decision aids in general.

This concludes the current paper under review.

CHAPTER 5:
STUDY 4

The following experiment, entitled *Ironic Efficiency* is currently in preparation. The version of the manuscript presented here is the author's original.

All authors were involved in the formulation of the study concept and design, and data analysis. Megan Bartlett collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

Introduction

While automated decision aids can assist the human operator to perform challenging decision tasks, such as identifying enemies on the battlefield (e.g., Wang, Jamieson, & Hollands, 2009), such aids will rarely be 100%-reliable (Wickens, Thomas, & Young, 2000). Unfortunately, human operators tend to interact with such imperfectly reliable aids in a suboptimal way, either over-relying or under-relying on the aid's judgments (Parasuraman, 2000; Parasuraman & Riley, 1997). As a result, automation-aided performance falls short of achievable levels (e.g., Meyer, 2001; Rice & McCarley, 2011; Robinson & Sorkin, 1985).

Operators' tendency toward poor automation also complicates the task of predicting just how much an aid will improve human users' performance. An ideal-user model of human-automation interaction places a statistical upper limit on the decision accuracy achievable by a human-automation team (Robinson & Sorkin, 1985; Sorkin, Hays, & West, 2001). Without an accurate model of the operators' actual strategy for using an aid, though, suboptimal performance levels may be difficult to estimate *a priori*. As a step toward enabling better predictions of automation benefits, Bartlett and McCarley (2017) benchmarked automation-aided performance to the predictions of various models of collaborative decision making (Green & Swets, 1966; Macmillan & Creelman, 2005). Participants performed a visual discrimination task with assistance from a 93%-reliable decision aid, and automation-aided performance was compared to the predictions of models that ranged from statistically optimal to highly inefficient. These included,

- the *optimal weighting* (OW) model (Bahrami et al., 2010; Sorkin et al. 2001), which assumes that to reach an automation-assisted decision, the operator averages his or her own estimate of signal likelihood with that of the aid, weighting each estimate by the agent's average sensitivity;

- the *uniform weighting* (UW) model (Bahrami et al., 2010; Sorkin et al. 2001), which assumes that the operator's and aid's judgments are averaged in an unweighted manner;
- the c*ontingent criterion* (CC) model (Robinson & Sorkin, 1985), which assumes that the aid and operator work in sequence, with the operator establishing his or her own response criterion contingent on the aid's judgment;
- the *best decides* (BD) model (Bahrami et al., 2010; Denkiewicz, Rączasek-Leonardi, Migdal, & Plewczynski, 2013), which assumes that the operator bases their decision on the judgments of the more reliable decision maker (aid or operator);
- the *coin flip* (CF) model (Bahrami et al., 2010), which assumes that the operator defers to the aid with a probability of 0.50, and;
- the *probability matching* (PM) model (Bartlett & McCarley, 2017), which assumes that the operator defers to the aid with a probability equal to the aid's average reliability level.

Observed sensitivity in aided conditions fell far short of the predictions of the OW, UW, CC, and BD models, approximating the predictions of the PM and CF models. Importantly, participants did not appear to use either of these strategies; taking into account response bias as well as sensitivity, data appeared most consistent with a suboptimal form of the CC model, in which decision makers adjusted their response criterion insufficiently in response to cues from the aid (Robinson & Sorkin, 1985). Results nonetheless suggested that the CF and PM models can be used as heuristic models for predicting automation-aided sensitivity.

The generalizability of this conclusion is limited, though, as Bartlett and McCarley (2017) tested aids of only one, fairly high, reliability level. The CF and PM models might thus be of less heuristic value for aids higher or lower in reliability. Existing data in fact imply that the CF and PM

strategies are unlikely to be useful heuristic models for low reliability aids. The benefits of an automated decision aid decline as the reliability of the aid decreases, with automation-aided performance generally leveling off around unaided levels (e.g., Rice & McCarley, 2011; Rovira, McGarry, & Parasuraman, 2007; Skitka, Mosier, & Burdick, 1999; Wickens & Dixon, 2007). In contrast to these results, the CF and PM models, along with the UW model, imply that an aid substantially less reliable than the human operator will compromise the human operator's judgments, reducing sensitivity below unaided levels. The CF and PM strategies therefore seem likely to underestimate automation-aided sensitivity when the automated aid is highly unreliable. Conversely, they may tend even to overestimate aided sensitivity when the aid is more reliable than that tested by Bartlett and McCarley (2017). See Figure 5-1 for model prediction simulations across varying levels of aid sensitivity (i.e., *d'* of 0, 1, 2, 3, 4, or 5), assuming a fixed level of unaided *d'* (i.e., 3).

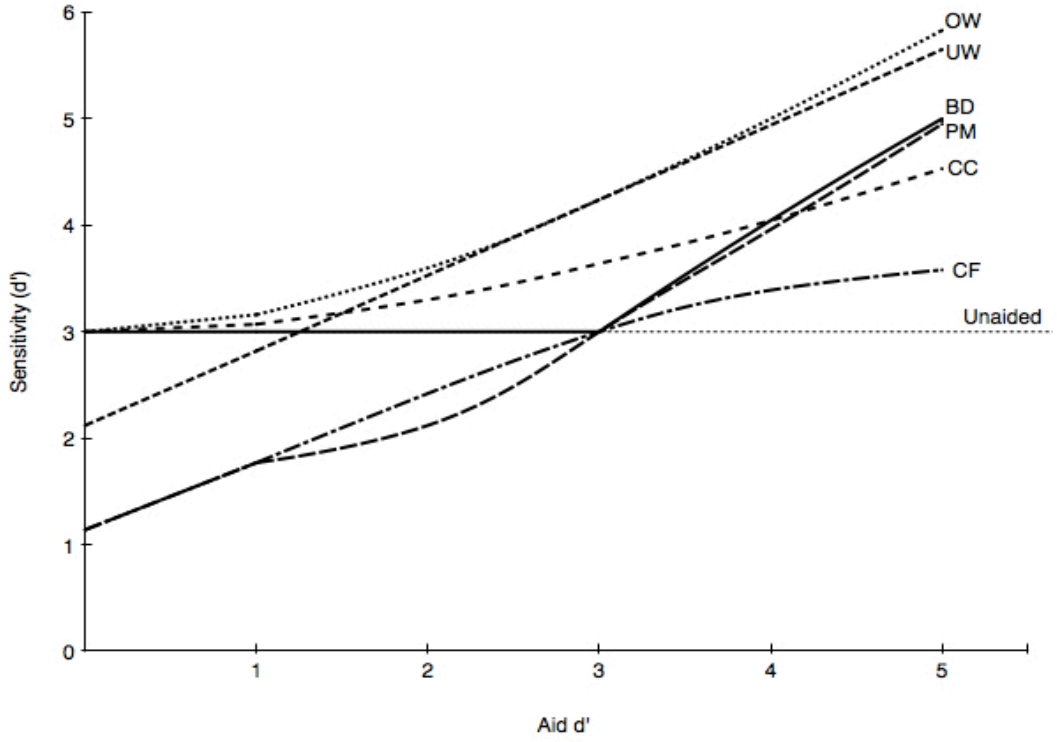

Figure 5-1. Model prediction (dotted lines) simulations across varying levels of aid sensitivity. The horizontal dotted line indicates unaided sensitivity.

To test the generality of the CF/PM models as a heuristic for predicting automation-aided performance, the present experiment benchmarked automation-aided sensitivity against the predictions of the models described above across different levels of aid reliability. Participants performed a two-alternative forced choice (2AFC) task requiring them to view and classify the dominant color of a series of orange and blue random dot images each trial (Bartlett & McCarley, 2017; Voss, Rothermund, & Ross, 2004). They performed the task alone and with assistance from a 60%, 85%, or 96% reliable automated decision aid. The aid rendered its judgment as a binary diagnosis accompanied by an estimate of signal strength in the form of a likelihood ratio indicating how strongly data favored the proffered binary diagnosis.

Finally, in keeping with other studies of decision aid usage (e.g., Merritt, Heimbaugh, LaChapell, & Lee, 2013; Merritt, Lee, Unnerstall, & Huber, 2015; Wang et al., 2009; Wiczorek, 2017), regression analyses examined the relationship between a variety of individual difference measures and a measure of automation usage.

This research complied with the tenets of the Declaration of Helsinki and was approved by the Social and Behavioural Research Ethics Committee at Flinders University. Informed consent was obtained from all participants.

Method

Participants. Participants were 48 adults (mean age = 21.71 years, *SD* = 4.59, range = 18-34; 38 females, 10 males) recruited from Flinders University. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 60 min. Participants were fluent in English, and were screened for normal color vision and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. The experimental task was controlled by software written in PsychoPy (Peirce, 2007), and stimuli were presented on a 23-inch Samsung monitor with a

resolution of 1,920 x 1,080 pixels and a 120 Hz refresh rate. Participants were seated approximately 60 cm from the monitor, with viewing distance unconstrained.

Stimuli were blue and orange random dot images (256 x 256 pixels), presented against a white background. Each stimulus was either blue-dominant or orange-dominant. In the blue-dominant stimuli, each pixel was randomly assigned the color blue with a probability of 0.52 or the color orange with a probability of 0.48. In the orange-dominant stimuli, those probabilities were reversed. Figure 5-2 shows a sample orange-dominant stimulus image.

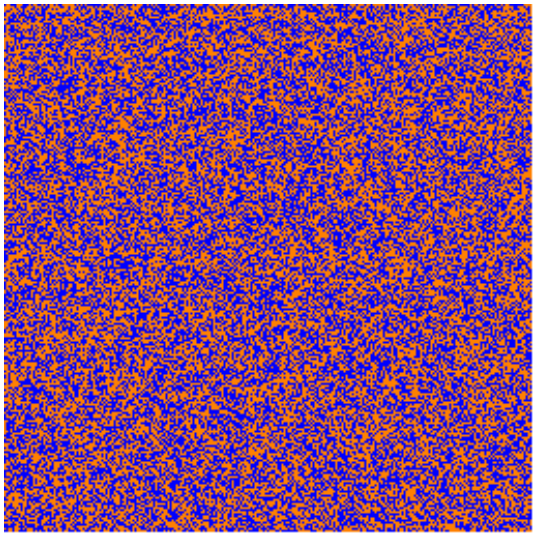

Figure 5-2. A sample orange-dominant stimulus image.

Automated Aid. On some blocks of trials, participants were assisted by an automated decision aid that judged whether the stimulus presented each trial had been generated using the parameters of the blue-dominant or orange-dominant distribution. The aid's judgments were generated using a standard equal-variance Gaussian signal detection model (Macmillan & Creelman, 2005).

Participants were assisted by aids of varying reliability. For trials on which the true stimulus categorization was blue-dominant, the aid's evidence value was sampled from a Gaussian distribution with either a mean of -1.75, -1, or -0.25 and a standard deviation of 1. For trials on

which the true stimulus categorization was orange-dominant, the aid's evidence value was sampled from a Gaussian distribution, with either a mean of 1.75, 1, or 0.25 and a standard deviation of 1. Thus, the *d'* of the aid could either be 3.5, 2, or 0.5. The aid transformed evidence values into binary judgments using an unbiased response threshold, offering a judgment of blue-dominant if the evidence value sampled for a given trial was less than 0 and a judgment of orange-dominant if the evidence value sampled was greater than 0. The unbiased criterion combined with a *d'* of 3.5 produced an average accuracy rate of 96%. The unbiased criterion combined with a *d'* of 2 produced an average accuracy rate of 85%. Finally, the unbiased criterion combined with a *d'* of 0.5 produced an average accuracy rate of 60%.

The aid rendered its binary judgment with an estimate of signal strength in the form of a likelihood ratio. The aid's raw sampled evidence value *x* was converted each trial into a likelihood ratio,

$p(x \mid \text{blue-dominant stimulus}) / p(x \mid \text{orange-dominant stimulus})$.

Values were displayed in the format, "Likelihood = *a*:1", where $a \geq 1$. An example of the aid's judgment is, "Aid judges: blue; Likelihood = 39:1."

Participants were informed that higher values indicated stronger evidence. Because they were generally not expected to have had extensive formal training in statistics, however, they were not provided any additional information about the distribution of evidence values.

Individual Difference Measures. Trust in the automated aid, perceived accuracy of the aid, and self-perceived accuracy were measured using items after Merritt (2011). Trust in the aid was assessed using a 6-item self-report measure. Participants responded on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). An example item is, "I believe the aid is a competent performer."

Perceived accuracy of the aid was assessed with an item asking, “Out of the 100 trials you completed WITH assistance from the automated aid, how many times did you think the aid was correct?” Participants filled in the blank, “I think the aid was correct on ___ out of 100 trials.”

Self-perceived accuracy was assessed with an item asking, “Out of the 100 trials you completed WITHOUT assistance from the automated aid, how many times did you think that you were correct?” Participants filled in the blank, “I think I was correct on ___ out of 100 trials.”

Procedure. Participants performed a 2AFC task requiring them to classify stimulus images as coming from blue- or orange-dominant distributions. A cover story asked the participants to imagine themselves as geologists sorting samples of a mineral into blue and orange strains. The instructions informed them, “Unfortunately, the two strains are difficult to tell apart. Both are speckled blue and orange. The only difference visually is that one strain tends to have a little more orange, and the other tends to have a little more blue. However, there is a lot of overlap in their appearance, and it is almost impossible to sort them with 100% accuracy by eye.” Participants were asked to decide each trial if the sample they were presented was orange-dominant or blue-dominant, and to provide an estimate of their decision confidence. They rendered responses by clicking on a six-point rating scale underneath the stimulus image. Responses on the scale were labeled, *Definite blue*, *Probable blue*, *Guess blue*, *Guess orange*, *Probable orange*, and *Definite orange*. Rating scale data were collected to perform anticipated additional analyses.

Participants were also told that on some trials, they would be assisted by an automated decision aid that would provide a blue or orange judgment along with an estimate of certainty. Instructions read, “The aid works by testing the chemical properties of the sample, and then assessing whether the sample is more likely to be ORANGE or BLUE.” Participants assisted by an aid with a *d’* of 3.5 were advised, “However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is

correct 96% of the time and incorrect 4% of the time." Participants assisted by an aid with a *d'* of 2 were advised, "However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is correct 85% of the time and incorrect 15% of the time." Finally, participants assisted by an aid with a *d'* of 0.5 were advised, "However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is correct 60% of the time and incorrect 40% of the time."

Participants in all reliability conditions were informed, "To help you predict whether it is right or wrong, the aid will give its assessment along with a likelihood ratio each trial. A higher ratio means that the aid is more likely to be correct…You should use the aid to help you make your decisions, but be aware that you are free to disagree with it any time you wish. Use your own best judgement."

Figure 5-3 shows the sequence of events within an unaided trial. Each trial was preceded by a message reading, "Click the circle below to start the next trial." Each aided trial comprised a 500-ms blank interval, a 1,500-ms screen displaying the automated aid's diagnosis, another 500-ms blank interval, and then the stimulus display, which remained onscreen until the participant's response. At the conclusion of each trial, participants received a 1,500-ms feedback message of either "Correct!" or "Incorrect!" The sequence of events on unaided trials was identical to that on aided trials, except that the aid's diagnosis was replaced by a neutral message, "Waiting for image." Presentation of the aid's diagnosis before the stimulus display allowed participants time to attend to the diagnosis carefully, and ensured that the diagnosis and stimulus arrived in the same order in which the CC model presumes they are processed (though see Wiegmann, McCarley, Kramer, & Wickens, 2006, for evidence that automation dependence is similar regardless of the order in which cue and stimulus are presented). Other models make no presumption as to the order of processing.

The neutral message served to match the sequence and timing of events across the aided and unaided blocks.

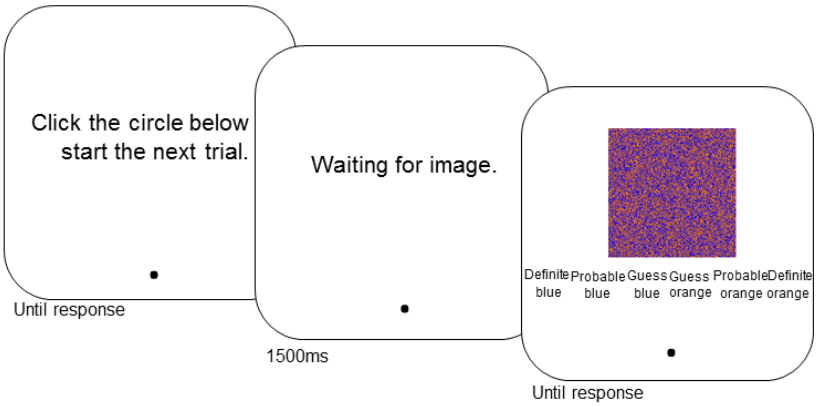

Figure 5-3. The sequence of events within an unaided trial.

Each session comprised a block of 50 unaided practice trials followed by a block of 50 aided practice trials, then a block of 100 unaided experimental trials and a block of 100 aided experimental trials, with the order of the experimental blocks counterbalanced across participants. Each trial, the stimulus category was selected randomly and with equal probability from among the two options (i.e., blue- or orange-dominant), and the stimulus image was then generated randomly. Participants were allowed to rest between blocks. Participants were randomly assigned to the aid reliability conditions in equal numbers in a between-subjects design.

Analysis

For analysis, orange-dominant stimuli were treated as signal events and blue-dominant stimuli as noise events. For clarity of exposition, we refer to orange and blue judgments as *yes* and *no* judgments, respectively. Hit rates and false alarm rates were calculated from the participants' responses, and data were converted to signal detection measures of sensitivity and bias, *d'* and *c* (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). A prior of 0.5 was added to the raw response frequency

value in each cell of the 2 x 2 signal detection theory (SDT) matrix for each participant to correct for perfect hit and false alarm rates (Hautus, 1995). Data from practice trials were excluded from analysis.

Data analysis employed Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) sampling procedure (Kruschke, 2013, 2015; Lee & Wagenmakers, 2014). This approach begins by assuming a prior distribution on a parameter value of interest, then updates the prior through probabilistic sampling to approximate the posterior distribution on parameter values in light of the observed data. Analyses were conducted using sampling functions from the package JAGS (Plummer, 2015) in R (http://www.r-project.org). *d'* scores were analyzed in a 2 (Block: unassisted vs. assisted) × 3 (Aid reliability: 60%, 85%, 96%) mixed design, with participant treated as an additive effect (Kruschke, 2015). Effects were assumed to follow normal distributions with vague priors on their means and standard deviations. Following Kruschke (2015),

$$Y_{\text{Block, AidReliability, Participant}} \sim N(a0 + a_{\text{Block}} + a_{\text{AidReliability}} + a_{\text{Block} \times \text{AidReliability}} + a_{\text{Participant}}, \sigma_y^2)$$

$$\sigma_y \sim U(SD/1000, SD*1000)$$

$$a0 \sim N(M, [100 \times SD]^2)$$

$$a_{\text{Block}} \sim N(0, \sigma^2_{\text{Block}})$$

$$a_{\text{AidReliability}} \sim N(0, \sigma^2_{\text{AidReliability}})$$

$$a_{\text{Block} \times \text{AidReliability}} \sim N(0, \sigma^2_{\text{Block} \times \text{AidReliability}})$$

$$a_{\text{Participant}} \sim N(0, \sigma^2_{\text{Participant}})$$

$$\sigma_{\text{Block}}, \sigma^2_{\text{AidReliability}}, \sigma^2_{\text{Block} \times \text{AidReliability}}, \sigma_{\text{Participant}} \sim \Gamma(\alpha, \beta)$$

$$\alpha = SD/2$$

$$\beta = 2 * SD$$

where $Y_{\text{Block, AidReliability, Participant}}$ is the *d'* score for a given participant in a given cell of the design, *a0* is the estimated grand mean *d'*, a_{Block} is the effect of Block, $a_{\text{AidReliability}}$ is the effect of Aid Reliability, $a_{\text{Block} \times \text{Format}}$ is the effect of the Block x Aid Reliability interaction, $a_{\text{Participant}}$ is the participant effect, σ_y is the estimated standard deviation of the normally distributed *d'* scores, *M* is the grand mean of the observed *d'* scores, and *SD* is the standard deviation of the observed *d'* scores. Use of the data sample mean and standard deviation to set parameters of the priors ensured that the prior distributions were scaled appropriately to the data (Kruschke, 2015). The use of vague priors ensured that the analysis did not commit *a priori* to strong conclusions, and allowed the observed data to dominate the posterior distribution. Predictions for the various models of automation use were based on the Bayesian-estimated grand mean *d'* score for unaided performance on each iteration of the sampling procedure.

Parameter estimation was based on four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step in order to reduce sample autocorrelation, leaving a total of 100,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less, indicating satisfactory convergence of the MCMC chains (Kruschke, 2015).

Descriptive statistics reported include the mean and 95% highest density intervals (HDI) for the estimated posterior distributions (Kruschke, 2013). The 95% HDI is the region that contains 95% of the posterior distribution mass, and within which all values have higher probability than any values outside the region. If the distribution is unimodal and symmetrical, the 95% HDI is equivalent to the central 95% region of the posterior (Gelman et al., 2013). Where it is useful to compare measures to a value of 0—for example, when examining differences between aided and unaided performance, or between observed data

and model predictions—the reported statistics also include the proportion of the estimated posterior distribution that lies above or below 0 (Kruschke, 2013). Values are reported with the nomenclature $x\% < 0 < y\%$. For example, $1\% < 0 < 99\%$ indicates that 1% of the posterior distribution lies below 0, and 99% lies above. We describe an effect as credible if the 95% HDI on the difference between conditions does not overlap 0, and we describe an effect as decisive if more than 99% of the posterior distribution on difference scores falls to one side of 0 (cf. Jeffreys, 1961; Wetzels et al., 2011).

Results

As responses clustered toward the ends of the rating scales, we collapsed the ratings into binary responses for analysis.

Sensitivity. The gray bars of Figure 5-4 present the hierarchically-estimated group mean values of *d'*. Dotted lines in Figure 5-4 present mean model-predicted values.

Automation-aided *d'* bordered on being credibly higher than unaided *d'* for participants assisted by the 85%-reliable aid, $M_{diff} = 0.31$, 95% HDI [-0.04, 0.65], $4\% < 0 < 96\%$, and was decisively higher for participants assisted by the 96%-reliable aid, $M_{diff} = 0.60$, 95% HDI [0.22, 0.98], $0\% < 0 < 100\%$. Aided *d'*, however, failed to differ credibly from unaided *d'* for participants assisted by the 60%-reliable aid, $M_{diff} = -0.04$, 95% HDI [-0.42, 0.34], $58\% < 0 < 42\%$, consistent with earlier findings that automation less than 70%-reliable in its diagnoses produces little benefit to the human operator (Rice & McCarley, 2011; Rovira et al., 2007; Skitka et al., 1999; Wickens & Dixon, 2007).

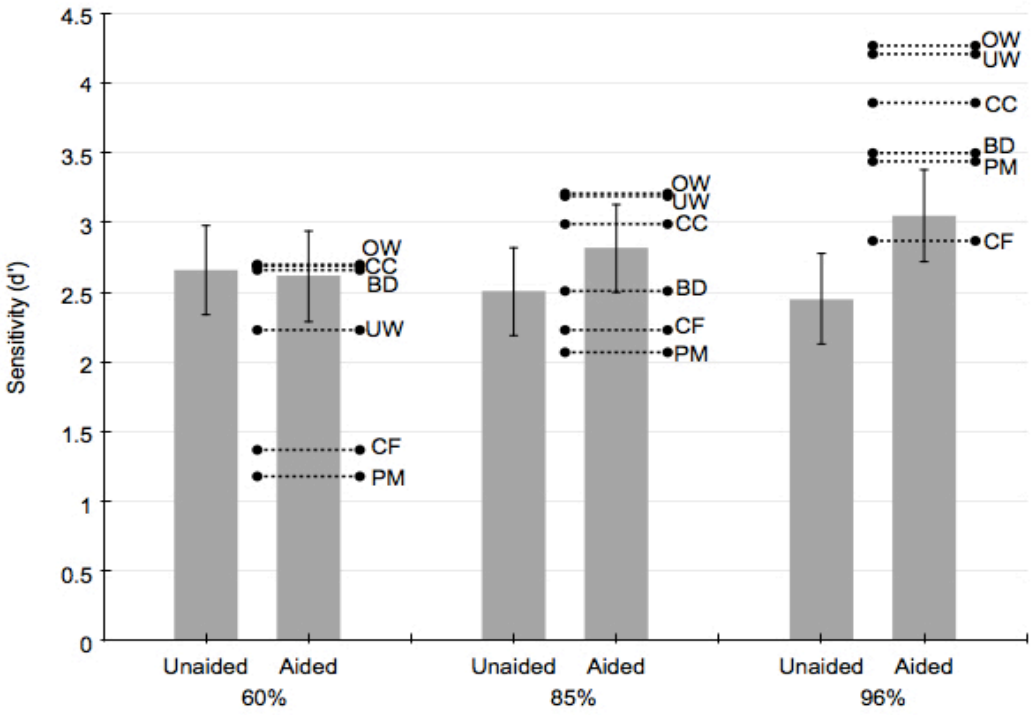

Figure 5-4. Hierarchically-estimated group mean values (gray bars) and model-predicted values (dotted lines) of *d'* for the 60%, 85%, and 96% aid reliability conditions. Error bars indicate 95% highest-density intervals.

Further analyses compared performance across conditions to examine the effects of aid reliability on automation usage. Unaided sensitivity did not differ credibly between the 85% and 60% aid reliability conditions, M_{diff} = -0.15, 95% HDI [-0.60, 0.28], 75% < 0 < 25%, 96% and 60% aid reliability conditions, M_{diff} = -0.21, 95% HDI [-0.66, 0.24], 82% < 0 < 18%, or 96% and 85% aid reliability conditions, M_{diff} = -0.06, 95% HDI [-0.49, 0.37], 61% < 0 < 39%, suggesting that conditions were similar in their baseline performance levels. Aided conditions likewise failed to differ credibly between conditions for the 85% and 60% aid reliability conditions, M_{diff} = 0.20, 95% HDI [-0.23, 0.64], 18% < 0 < 82%, or 96% and 85% aid reliability conditions, M_{diff} = 0.23, 95% HDI [-0.20, 0.68], 15% < 0 < 85%. Aided sensitivity did however differ credibly between the 96% and 60% aid reliability conditions, M_{diff} = 0.43, 95% HDI [-0.04, 0.89], 3% < 0 < 97%, indicating that assistance

from a highly reliable, relative to an imperfectly reliable aid improved participants' sensitivity.

Consistent with these effects, data showed a credible block × aid reliability interaction, as the difference between unaided and aided conditions differed credibly between the 96% and 60% aid reliability conditions, M_{diff} = 0.64, 95% HDI [0.04, 1.19], 1% < 0 < 99%. The interaction of block × aid reliability fell short of credibility when comparing the 85% and 60% aid reliability conditions, M_{diff} = 0.35, 95% HDI [-0.13, 0.84], 8% < 0 < 92%, and the 96% and 85% aid reliability conditions, M_{diff} = 0.29, 95% HDI [-0.17, 0.79], 11% < 0 < 89%, but trended in the expected direction in both cases.

Model Predictions. To assess model performance, analyses compared observed *d'* scores from the automation-aided conditions to the model-predicted scores. Mean model error scores (predicted scores minus observed scores) are presented in the text, with 95% HDIs.

For participants assisted by the 60%-reliable aid, the CF model decisively underestimated participants' automation-aided sensitivity, M_{err} = -1.25, 95% HDI [-1.56, -0.94], 100% < 0 < 0%, as did the PM model, M_{err} = -1.44, 95% HDI [-1.75, -1.13], 100% < 0 < 0%, 85%, and UW model, M_{err} = -0.39, 95% HDI [-0.72, -0.05], 99% < 0 < 1%. Observed sensitivity did not differ credibly from the predictions of the OW model, M_{err} = 0.09, 95% HDI [-0.29, 0.46], 33% < 0 < 67%, CC model, M_{err} = 0.07, 95% HDI [-0.31, 0.45], 36% < 0 < 64%, or BD model, M_{err} = 0.04, 95% HDI [-0.34, 0.42], 42% < 0 < 58%.

For participants assisted by the 85%-reliable aid, the OW model decisively overestimated participants' automation-aided sensitivity, M_{err} = 0.39, 95% HDI [0.08, 0.71], 1% < 0 < 99%, as did the UW model, M_{err} = 0.37, 95% HDI [0.06, 0.67], 1% < 0 < 99%, whereas the CF model decisively underestimated participants' automation-aided sensitivity,

M_{err} = -0.58, 95% HDI [-0.87, -0.29], 100% < 0 < 0%, as did the PM model, M_{err} = -0.75, 95% HDI [-1.04, -0.44], 100% < 0 < 0%, and BD model, M_{err} = -0.31, 95% HDI [-0.65, 0.04], 96% < 0 < 4%. Observed sensitivity did not differ credibly from the predictions of the CC model, M_{err} = 0.18, 95% HDI [-0.14, 0.50], 13% < 0 < 87%.

For participants assisted by the 96%-reliable aid, the OW model decisively overestimated participants' automation-aided sensitivity, M_{err} = 1.22, 95% HDI [0.90, 1.55], 0% < 0 < 100%, as did the UW model, M_{err} = 1.16, 95% HDI [0.82, 1.50], 0% < 0 < 100%, and CC model, M_{err} = 0.81, 95% HDI [0.50, 1.14], 0% < 0 < 100%, whereas the CF model decisively underestimated participants' automation-aided sensitivity, M_{err} = -0.18, 95% HDI [-0.50, 0.16], 85% < 0 < 15%. Observed sensitivity did not differ credibly from the predictions of the BD model, M_{err} = 0.45, 95% HDI [0.12, 0.78], 0% < 0 < 100%, or PM model, M_{err} = 0.39, 95% HDI [0.07, 0.71], 1% < 0 < 99%.

Efficiency. As an alternative way of assessing performance changes across levels of aid reliability, we can examine the efficiency of automation-aided performance. Efficiency provides an index of how far observed group performance in a collaborative signal detection task falls from the statistical ideal (Sorkin et al., 2001; Tanner & Birdsall, 1958), giving a measure of automation usage that effectively normalizes automation-aided sensitivity for each observer by unaided sensitivity. In the context of automation-aided decision making, efficiency, η, is,

$$\eta = \left(\frac{d'_{aided}}{d'_{OW}}\right)^2$$

where d'_{aided} is the sensitivity of the human-automation team, and d'_{OW} is the sensitivity of the ideal group. Efficiency of 1.0 indicates statistically optimal human-automation sensitivity. Values below 1.0 indicate performance short of achievable levels. Efficiency

was calculated from the estimated values of unaided and aided *d'* at every step of the MCMC chain used in the analyses of sensitivity reported above.

Figure 5-5 presents the hierarchically-estimated group mean values of automation-aided efficiency. Unlike absolute values of automation-aided sensitivity, which trended downward as the aid's reliability declined, aided efficiency trended upward as the aid became less reliable. Whereas participants assisted by a 96%- reliable aid achieved efficiency in the range of 0.5, decisively below 1.0, participants assisted by a 60%-reliable aid achieved near-perfect efficiency. Thus, as the aid became less reliable, participants' automation usage more closely approached optimal levels.

Efficiency failed to differ credibly between conditions for the 85% and 60% aid reliability conditions, M_{diff} = -0.17, 95% HDI [-0.49, 0.12], 87% < 0 < 13%, but was credibly different between the 96% and 85% aid reliability conditions, M_{diff} = -0.26, 95% HDI [-0.45, -0.07], 100% < 0 < 0%, and between the 96% and 60% aid reliability conditions, M_{diff} = -0.44, 95% HDI [-0.75, -0.15], 100% < 0 < 0%.

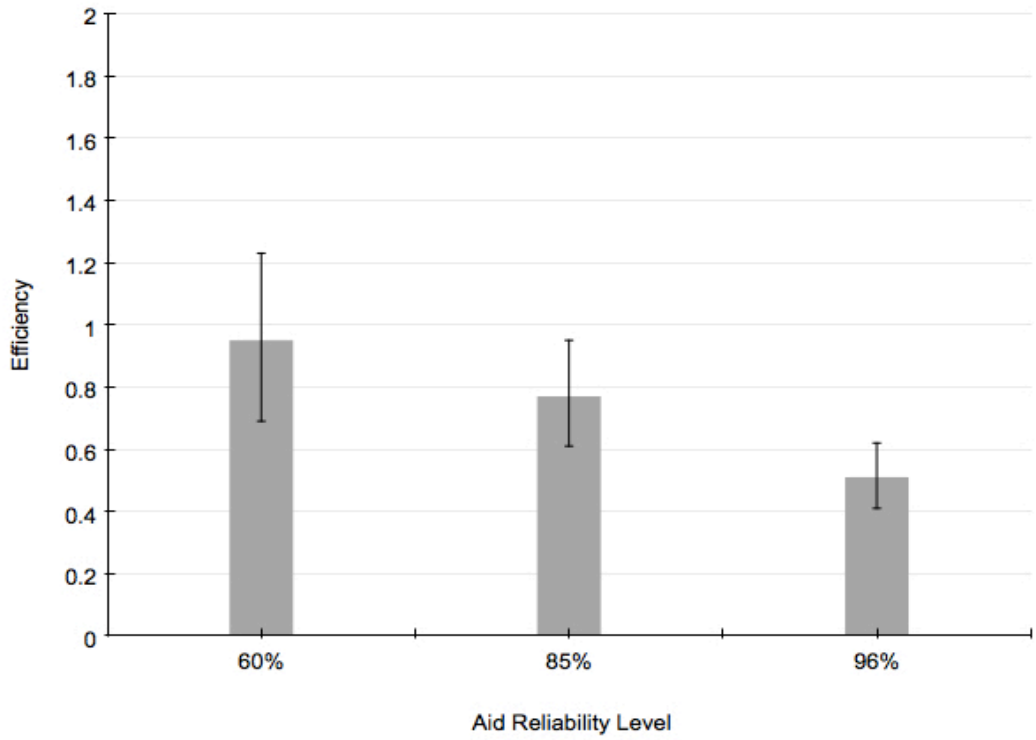

Figure 5-5. Hierarchically-estimated group mean values (gray bars) of automation-aided efficiency for the 60%, 85%, and 96% aid reliability conditions. Error bars indicate 95% highest-density intervals.

Individual Difference Measures. Bayesian linear regressions were carried out on the relationship between self-perceived accuracy and efficiency, trust and efficiency, perceived aid accuracy and trust, and finally, perceived aid accuracy and efficiency.

Participants assisted by the 60%-reliable aid who perceived themselves to be highly accurate bordered on showing credibly greater efficiency, r = 0.49, 95% HDI [-0.03, 1.02], 3.21% < 0 < 96.79%. No credible relationship was evident, however, between perceived aid accuracy and trust, r = 0.35, 95% HDI [-0.20, 0.91], 10.17% < 0 < 89.83%, perceived aid accuracy and efficiency, r = 0.09, 95% HDI [-0.52, 0.68], 37.65% < 0 < 62.35%, or trust and efficiency, r = -0.13, 95% HDI [-0.72, 0.46], 67.82% < 0 < 32.18%.

Participants assisted by the 85%-reliable aid who trusted the aid showed greater efficiency, r = 0.58, 95% HDI [0.08, 1.05], 1.20% < 0 < 98.80%. No credible relationship

was evident, however, between perceived aid accuracy and trust, $r = 0.35$, 95% HDI [-0.21, 0.91], 10.34% < 0 < 89.66%, perceived aid accuracy and efficiency, $r = 0.28$, 95% HDI [-0.28, 0.87], 15.62% < 0 < 84.38%, or self-perceived accuracy and efficiency, $r = 0.17$, 95% HDI [-0.42, 0.76], 27.29% < 0 < 72.71%.

Participants assisted by the 96%-reliable aid who perceived themselves to be highly accurate bordered on showing credibly greater efficiency, $r = 0.49$, 95% HDI [-0.03, 1.00], 3.17% < 0 < 96.83%. Participants assisted by the 96%-reliable aid who perceived the aid to be highly accurate also bordered on reporting credibly greater trust in the aid, $r = 0.49$, 95% HDI [-0.04, 1.01], 3.14% < 0 < 96.86%. No credible relationship was evident, however, between trust and efficiency, $r = 0.34$, 95% HDI [-0.23, 0.90], 10.91% < 0 < 89.09%, or perceived aid accuracy and efficiency, $r = 0.16$, 95% HDI [-0.42, 0.76], 28.06% < 0 < 71.94%.

Discussion

The present experiment investigated the effect of varying aid reliability on participants' information integration in an automation-assisted signal detection task, comparing observed performance levels to the predictions of various statistical models of collaborative decision making. Assistance from a highly reliable decision aid (85% or 96%) improved discrimination performance, as expected, while assistance from a moderately reliable aid (60%) did not, consistent with earlier results that automation that is less than 70%-reliable loses any benefit to the human operator (Rice & McCarley, 2011; Rovira et al., 2007; Skitka et al., 1999; Wickens & Dixon, 2007). Absolute levels of automation-aided sensitivity thus trended downward as the aid became less reliable, just as expected.

Relative to the best-achievable sensitivity levels, however, aided performance improved as aid reliability decreased. Replicating the findings of Bartlett and McCarley (2017), automation-

aided sensitivity fell closest to the predictions of the highly inefficient CF model when the aid itself was highly reliable (96%). In contrast, when the aid was moderately reliable (85%), aided sensitivity instead fell near the level predicted by the optimal CC model, and when the aid was less reliable still, performance approached the predictions of the OW model, the best-achievable level of performance. Taken together these findings suggest that the predictions of the CF and PM models impose an upper limit on performance efficiency when the aid is highly reliable, but that when the aid decreases in reliability, more efficient models instead become better descriptors of human performance.

Why did automation-aided efficiency improve even as aid reliability, and absolute aided performance level, declined? This is just the effect expected, interestingly, if participants largely ignore low-reliability aids. As described above, the OW model assumes that team judgments are based on a weighted average of the operator's and aid's individual judgments, where the judgment from each agent is weighted proportional to the agent's baseline sensitivity. As the aid's sensitivity approaches 0, the ideal usage strategy therefore becomes to give its judgments a weight of 0, that is, to ignore them. Thus, ironically, the tendency to disuse an imperfect aid produces efficient performance only when the aid performs near chance.

Near-optimal performance at low levels of aid reliability does not imply that participants were actually using an optimal weighing strategy to reach their aided decisions. As the reliability of the aid approaches chance, rather, the OW, BD, and optimal CC models all converge on the prediction that participants will simply ignore the aid. Which strategy were participants actually using? It could be the case that participants were using a range of different models across different levels of aid reliability, for example, that participants assisted by a highly reliable aid employed a CF strategy while participants assisted by a moderately reliable aid employed a nearly optimal CC strategy. This hypothesis, though, is obviously unparsimonious, and rests poorly with the data of

Bartlett and McCarley (2017), who found that performance with a highly reliable aid was best accounted for by a suboptimal CC model. It therefore appears more plausible to suggest that participants employed a CC strategy, but one in which the magnitude of cued-criterion shifts was poorly calibrated to the aid's reliability. One possibility the data suggest is a model in which the magnitude of cued criterion shifts was a negatively accelerating function of the optimal size of the shift. Or more simply, participants' cued-criterion shift following a cue might be proportional to the optimal shift, such that absolute deviation of the criterion from optimal increases as cue reliability increases. A criterion shift that was ½ of the optimal magnitude would have little absolute effect on performance when participants were assisted by a 60%-reliable aid, for instance, but would compromise performance substantially when participants were assisted by a 96%-reliable aid. Either of these models, notably, is consistent with the more general finding that decision makers tend to employ a 'sluggish beta' (Wickens, Hollands, Banbury, & Parasuraman, 2015) in signal detection tasks, adjusting their response bias inadequately in response to variation of signal rate and payoff (Green & Swets, 1966).

Inconsistent with previous research (i.e., Lee & See, 2004; Merritt, 2011; Merritt & Ilgen, 2008; Wang et al., 2009), no credible relationship emerged between perceived aid reliability and trust in the aid, across all aid reliability levels. In fact, no support was found for most of the relationships measured via self-report. Further research will be necessary to qualify or generalize these findings, identifying individual differences that might allow selection of efficient automation users.

This concludes the current paper in preparation.

CHAPTER 6:
GENERAL DISCUSSION

The current thesis investigated decision makers' automation usage strategies and examined elements of automation interface design that shape human-automation interaction. The current thesis had four main aims: 1) to explore and explain earlier findings that different forms of automation errors produce asymmetrical effects on the human operator's willingness to trust the automation, 2) to better understand operators' inefficient use of aids by comparing participants' aided performance levels to the predictions of statistical models of collaborative decision making, 3) to determine whether aided performance can be bolstered by manipulating the format of the aid's cues, and 4) to determine whether these results can be generalized across aids of varying reliability. The present chapter will summarize all findings, discuss implications for human factors practitioners, and finally suggest directions for future research.

Study 1 aimed to replicate the asymmetry between automation false alarms and misses on operator behavior, that is, that automation false alarms may hinder performance and subsequent automation use from human operators more so than automation misses, as previously reported (e.g., Dixon, Wickens, & McCarley, 2007; Rice & McCarley, 2011), and to explore its potential cause. Specifically, we speculated whether such an asymmetry would reflect either a tendency for operators to agree with automation whose response bias matches their own (Rice & McCarley, 2011), or an inherent tendency to disuse false alarm-prone aids more than miss-prone aids (McCarley, Rubinstein, Steelman, & Swanson, 2011; McCarley, Steelman, & Rubinstein, 2013).

To investigate this asymmetry, participants performed a simulated baggage screening task, either alone, or with assistance from a 95%-reliable decision aid, whereby we manipulated both the human operator's and automated aid's response bias. The response bias of the human operator was manipulated via a point system assigning different payoffs to correct and incorrect responses,

whereby participants were encouraged to respond conservatively, neutrally, or liberally. The response bias of the aid was manipulated such that the aid was prone to committing either false alarms or misses. We hypothesized that being assisted by an aid, as against performing the task alone, would bolster performance, as reflected by higher levels of sensitivity and shorter RTs. We also expected, critically, that a conservative (miss-prone) aid would produce larger benefits than a liberal (false alarm-prone) aid. Furthermore, we hypothesized that a false alarm-prone aid would produce lower levels of compliance than a miss-prone aid, while a miss-prone aid would produce lower levels of reliance than a false alarm-prone aid. Most importantly, we speculated that an interaction would emerge between payoff matrix and aid bias on compliance and reliance, if operators held an inherent tendency to agree with automation whose response bias matched their own.

While results confirmed that assistance from an aid improved human performance, data failed to provide conclusive evidence either for or against an asymmetry in the effects of automation false alarms and misses. The failure to replicate the asymmetry does not provide strong evidence against the effect, but it does suggest that the effect may be modest at best. What could be the reason for the failure to replicate? The results are slightly puzzling as the general procedure was similar and the stimuli used were identical to that of Rice and McCarley's study. Differences in participant populations may have been the reason, as participants in Rice and McCarley's study came from a Midwestern U.S. university, while participants in the current study came from an urban Australian university. It is unclear why different participant populations would induce such contrasting results. Some research, however, does point toward cultural variation in individual's trust and perception of automation (i.e., Huerta, Glandon, & Petrides, 2012; Li, Rau, & Li, 2010; Rau, Li, & Li, 2009), which may help to explain our results. Regardless, the asymmetry between automation false alarms and misses may not be as strong as previously reported.

Study 2 aimed to investigated participants' decision making strategies for interacting with automated aids, benchmarking aided sensitivity to the predictions of seven statistical models of collaborative decision making, spanning from highly efficient to highly inefficient. Participants performed a two-alternative forced choice task requiring them to determine the dominant color in a series of random dot images (Voss, Rothermund, & Voss, 2004). They either performed this task alone or with assistance from a 93%-reliable automated decision aid. In two experiments (Experiments 1 & 3), the aid provided participants with a binary judgment along with an estimate of signal strength, while in another (Experiment 2) the aid instead provided only a binary diagnosis. Additionally, participants in Experiment 3 received a running score to help them track their performance.

Though assistance from an aid improved human performance in all three experiments, participants interacted with the aid suboptimally, with aided sensitivity hewing closest to the predictions of the probability matching (PM) model, which assumes that the operator defers to the aid with a probability equal to the aid's average reliability. Participants' response bias conditionalized on the aid's judgments, however, was both inconsistent with the PM model, or any of the other models tested. Model-fitting analyses concluded that a suboptimal form of the contingent criterion (CC) model best explained the data. Under this model, participants shift their response criterion inadequately in response to an aid's judgment. This tendency for participants to use a suboptimal CC strategy in aided tasks is both consistent with the 'sluggish beta' phenomenon (Chi & Drury, 1998; Neyedli, Hollands, & Jamieson, 2011; Wang, Jamieson, & Hollands, 2009), and prior research (e.g., Elvers & Elrif, 1997; Meyer, 2001; Robinson & Sorkin, 1985; Wang et al., 2009).

Interestingly, performance was similar whether the aid provided graded or binary judgments, and scoring participants on their performance did little to bolster human-automation

interaction. The fact that participants derived no benefit from graded cues from the aid suggests that participants may have simply failed to use these judgments. If used optimally, the graded cues could have induced performance that exceeded the levels predicted by the CC model. Why did participants fail to use the aid's graded evidence assessments? It could be the case that participants sacrificed decision accuracy to minimize the effort involved with remembering the aid's cues (Bettman, Johnson, & Payne, 1990), as the steps involved with using the cues optimally may have been cognitively demanding for participants. Moreover, past research has shown that decision makers often fall short of optimal performance due to limitations in cognitive and information-processing abilities (Simon, 1955; Tversky & Kahneman, 1974).

As participants in Study 2 effectively ignored the aid's graded evidence assessments, and instead relied solely on the aid's binary judgments, Study 3 investigated whether participants' information integration strategies could be improved if we manipulated the format of an automated decision aid's cues. Participants performed the same two-alternative forced choice task as in Study 2, again unassisted or with assistance from a 93%-reliable automated decision aid. In two experiments (Experiments 1 & 2), the aid provided participants with a binary judgment along with an estimate of signal strength in the form of either a raw value, a likelihood ratio, or a confidence rating, while in another (Experiment 3), the aid instead provided a binary judgment along with either a verbal or verbal/visuospatial expression of confidence (Experiment 3). Participants in Experiment 1 were not provided with performance feedback, while participants in Experiment 2 were. Aided sensitivity was again benchmarked to the predictions of various statistical models. Trust in the aid, perceived accuracy of the aid, and self-perceived accuracy of the participant were investigated using self-report measures.

Results again confirmed that aided performance was suboptimal, most closely matching the predictions of the coin flip (CF) model, which assumes that the human operator defers to the aid

with a 50% probability, in the absence of performance feedback (Experiment 1), and the PM model and best decides (BD) model, which assumes that the operator defers to the most reliable team member, in the presence of performance feedback (Experiments 2 & 3). Trial-by-trial performance feedback thus appeared to improve participants' usage strategies, a finding consistent with previous research (Beck, Dzindolet, & Pierce, 2007; Dzindolet, Pierce, Peterson, Purcell, & Beck, 2002).

More importantly, performance was similar across aided formats—raw values, likelihood ratios, or confidence rating formats in two experiments, verbal and verbal-spatial representations of probability in another. The finding that aided performance was similar across these formats suggests that participants' failure to use the graded evidence judgments in Study 2 was not due to difficulties estimating the probabilistic distribution of raw evidence values; the distributional information necessary to interpret raw evidence values optimally was encoded inherently within the alternative formats tested in Study 3, to no evident effect on participants' decision making.

Studies 2 and 3 provided consistent evidence that automation-aided sensitivity fell in the range predicted by the CF/PM models, implying that these models may be used as heuristics for predicting automation-aided sensitivity. However, these studies all used an aid that was 93%-reliable. Study 4 tested the generalizability of this conclusion using aids of 60%, 85%, and 96% reliability. Participants performed the same task as used in Studies 2 and 3, but this time alone or with assistance from a 60%, 85%, or 96%-reliable automated decision aid. The aid provided a binary judgment, along with an estimate of signal strength in the form of a likelihood ratio. Trust in the aid, perceived accuracy of the aid, and self-perceived accuracy were also investigated using self-report. Automation-aided efficiency was also assessed.

Aided performance was superior to unaided performance when participants were assisted by an 85% or 96%-reliable aid, but not when they were assisted by a 60%-reliable aid, a finding

consonant with earlier research (e.g., Rice & McCarley, 2011; Rovira, McGarry, & Parasuraman, 2007; Skitka, Mosier, & Burdick, 1999; Wickens & Dixon, 2007). Consistent with the findings of Studies 2 and 3, furthermore, aided sensitivity was highly inefficient when the aid itself was highly reliable (96%). When participants were assisted by a less reliable aid, however, efficiency improved. An 85%-reliable aid produced sensitivity roughly matching the predictions of the optimal CC model, and the 60%-reliable aid produced sensitivity matching the predictions of the optimal weighting (OW) model, which assumes that judgments from both the aid and operator are averaged, with higher weighting given to judgments from the more sensitive member. Performing optimally when assisted by a 60%-reliable aid, however, does not mean that participants were using the OW model to make their decisions, as the predictions of the OW, BD, and optimal CC models all converge in these circumstances to recommend that the operator ignore the aid altogether. So how did participants make their decisions, and which strategies did they use? Participants could have used different strategies at different levels of aid reliability, such that participants assisted by a 96%-reliable aid used a CF strategy, while participants assisted by an 85%-reliable aid used an optimal CC strategy. This possibility, however, is far from parsimonious, and contradicts the conclusions of Study 2, which found that a suboptimal CC model best accounted for participants' interaction with a 93%-reliable decision aid. Thus, a simpler conclusion to unify the findings across all three levels of aid reliability is that participants used a suboptimal CC strategy, where the size of the criterion shifts were incorrectly calibrated with the reliability of the aid. More particularly, the data suggest that cue-contingent criterion shifts are a negatively accelerated function of the optimal shift size, as determined by the aid's reliability.

Implications & Future Directions

Results lend insight into participants' automation-aided decision strategies and provide benchmarks for assessing automation-aided performance levels. Knowing the sensitivity of an

operator and automated aid, system designers can use the models to help predict the levels of aided sensitivity the operator will attain. Such predictions can inform cost-benefit analyses of designing, building, and deploying automated aids.

Results also add to the growing literature, confirming that, while automated decision aids can bolster human performance (e.g., Maltz & Meyer, 2001; Meyer, 2001; Wickens & Dixon, 2007), this benefit is far from optimal, with participants largely under-utilizing the aid's advice (e.g., Lee, 2008). Individuals' propensity to disuse a highly reliable aid may stem from lack of trust (e.g., Lee, 2006; Lee & Moray, 1992; Liu, Fuld, & Wickens, 1993; Muir, 1987; Rice & McCarley, 2011), overconfidence (e.g., Dzindolet, Pierce, Beck, & Dawe, 2002), or algorithm aversion (Dietvorst, Simmons, & Massey, 2015).

The results of Study 1 demonstrate that, although previous work has indicated that automation false alarms compromise performance more than automation misses (e.g., Dixon et al., 2007; Rice & McCarley, 2011), the effect may be less robust than we assumed (cf., Wickens et al., 2009). This suggests that automation designers need not feel compelled to minimize false alarms against the recommendation of normative factors (i.e., signal rate and payoff matrix). This may be especially true when such alerts are perceived as reasonable or plausible to the human operator, rather than as strictly random or unrelated to the state of the world (Lees & Lee, 2007).

Taken together, the results of Studies 2 and 4 demonstrate that the efficiency of automation-aided performance is near to optimal when the aid is close to chance reliability, but becomes highly inefficient as the aid's reliability increases. This suggests that when working with assistance from highly reliable aids, performance is likely to fall in the range of the CF and PM models, but when assisted by a less reliable aid, however, more optimal models instead become better descriptors of human performance. Of course, as the aid's reliability approaches chance levels, the absolute

benefit to performance that the aid offers necessarily decreases even if performance efficiency improves.

The results of Study 3 indicate that participants combined aided cues with their own signal detection judgments highly inefficiently, regardless of the format in which judgments from the aid were displayed. Future research is necessary to determine how more efficient human-automation collaboration can be induced. This could involve investigating how to make the cues more interpretable to the user either through instruction or training (Sedlmeier & Gigerenzer, 2001), or an alternative form of representation, such as the graphical or pictorial display of the aid's numeric probabilistic cues (e.g., Brase, 2009; Neyedli, Hollands, & Jamieson, 2011). Pie charts, for example, are effective representations for conveying probabilistic information (Hollands & Spence, 1992, 1998; Spence & Lewandowsky, 1991) that can be interpreted accurately by decision makers (Hollands & Spence, 1998).

Further research will be needed to test these various suggestions, of course, and to generalize the pattern of effects reported in these studies across different and more realistic forms of signal detection task, different participant populations, more realistic signal rates, and/or more extreme payoff schemes.

REFERENCES

Adams, J. A. (2009). Multiple robot/single human interaction: Effects on perceived workload. *Behaviour & Information Technology, 28*, 183–198.

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology, 11*, 909-918.

Andre, A. D., & Cutler, H. A. (1998). Displaying uncertainty in advanced navigation systems. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 31-35). Santa Monica, CA: Human Factors and Ergonomics Society.

Bahner, J. E., Hüper A., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*, 688-699.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*, 1081- 1085.

Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., & Scott, L. A. (2006). Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors, 48*, 693-709.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*, 775-779.

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 615-633.

Bar-Hillel, M. (1990). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 39-61). Amsterdam: Elsevier.

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors, 59,* 881-900.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. The *Academy of Management Review, 3,* 439-449.

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decisions: Controlling intent and appraisal errors in a target detection task. *Human Factors, 49,* 429-437.

Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory, 19,* 169-184.

Berger, J. O., Wolpert, R. L., Bayarri, M. J., DeGroot, M. H., Hill, B. M., Lane, D. A., & LeCam, L. (1988), *The Likelihood Principle* [Lecture Notes-Monograph Series], Vol. 6, Hayward, CA: Institute of Mathematical Statistics.

Bettman, J. R., Johnson, E. J., & Payne, J. W. (1990). A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes, 45,* 111-139.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. Journal of Forecasting, 1, 257-269.

Bhanu, B. (1986). Automatic target recognition: State of the art survey. *IEEE Transactions on Aerospace and Electronic Systems, AES-22,* 364-379.

Billings, C. E. (1996). *Toward a human-centered approach to automation*. Englewood Cliffs, NJ: Erlbaum.

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach*. Mahwah, NJ: Erlbaum.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation, 13,* 173-189.

Bisantz, A. M. (2013). Uncertainty visualization and related techniques. In J. D. Lee & A. Kirlik (Eds.), *The oxford handbook of cognitive engineering* (pp. 579-594). New York, NY: Oxford University Press.

Bisantz, A. M., Marsiglio, S. S., & Munch, J. (2005). Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors, 47,* 777-796.

Bisantz, A. M., & Pritchett, A. R. (2003). Measuring judgment in complex, dynamic environments: A lens model analysis of collision detection behavior. *Human Factors, 45,* 266–280.

Bliss, J. P. (1997). Alarm reaction patterns by pilots as a function of reaction modality. *International Journal of Aviation Psychology, 7,* 1-14.

Bliss, J. P., Dunn, M., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills, 80,* 1231-1242.

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics, 38,* 2300-2312.

Botzer, A., Meyer, J., Bak, P., & Parmet, Y. (2010). User settings of cue thresholds for binary categorization decisions. *Journal of Experimental Psychology: Applied, 16,* 1-15.

Bowers, C.A., Oser, R.L., Salas, E., & Cannon-Bowers, J.A. (1996). Team performance in automated systems. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 243-263). Hillsdale, NJ: Erlbaum.

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology, 23*, 369–381.

Breazeal, C. (2002). *Designing sociable robots*. Cambridge, MA: MIT press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd Edition). Berkeley, CA: University of California Press.

Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes, 46*, 240–263.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 281–294.

Chen, S. I., Visser, T. A. W., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied, 23*, 240-262.

Chi, C., & Drury, C. G. (1998). Do people choose an optimal response criterion in an inspection task? *IIE Transactions, 30*, 257-266.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology, 30*, 39-78.

Cotté, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger driver's reliance on collision warning systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.

Dadashi, N., Stedmon, A. W., & Pridmore, T. P. (2012). Semi-automated CCTV surveillance: The effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. *Applied Ergonomics, 44*, 730–738.

Darling, S., Allen, R. J., & Havelka, J. (2017). Visuospatial bootstrapping: When visuospatial and verbal memory work together. *Current Directions in Psychological Science, 26*, 3-9.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Degani, A. (2001). Taming HAL: Designing interfaces beyond 2001. New York, NY: Macmillan.

Denkiewicz, M., Rączaszek-Leonardi, J., Migdal, P., & Plewczynski, D. (2013). Information-sharing in three interacting minds solving a simple perceptual task. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2172-2176). Austin, TX: Cognitive Science Society.

de Vries, P., Midden, C., & Bowhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies, 58,* 719-735.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144,* 114-126.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. *Human Factors, 48,* 474-486.

Dixon, S. R., Wickens, C.D., & McCarley, J. S. (2007). On the independence of compliance and reliance: are automation false alarms worse than misses? *Human Factors, 49,* 564-572.

Drury, C. G., & Chi, C. F. (1995). A test of economic models of stopping policy in visual search. *IIE Transactions, 27,* 382-393.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies, 58,* 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13,* 147-164.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44,* 79-94.

Dzindolet, M., Pierce, L., Peterson, S., Purcell, L., & Beck, H. (2002). The influence of feedback on automation use, misuse, and disuse. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 551-555). Santa Monica, CA: Human Factors and Ergonomics Society.

Edwards, E. (1977). Automation in civil transport aircraft. *Applied Ergonomics, 4*, 194-198.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*, 465-485.

Elvers, G. C., & Elrif, P. (1997). The effects of correlation and response bias in alerted monitor displays. *Human Factors, 39*, 570-580.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*, 381-394.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*, 1–18.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2007). Reliance on automation as a function of expectation of reliability, cost of verification, and age. In *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting* (pp. 6-10). Santa Monica, CA: Human Factors and Ergonomics Society.

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system*. Columbus, OH: Ohio State University Foundation.

Fuld, R. B. (2000). The fiction of function allocation, revisited. *International Journal of Human-Computer Studies, 52*, 217-233.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-511.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1,* 19-33.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62,* 451-482.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gilbert, F. J., Astley, S. M., Gillan, M. G. C., Agbaje, O. F., Wallis, M. G., James, J., Boggis, C. R. M., & Duffy, S. W. (2008). Single reading with computer-aided detection for screening mammography. *The New England Journal of Medicine, 359,* 1675-1684.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Gupta, N., Bisantz, A. M., & Singh, T. (2002). The effects of adverse condition warning system characteristics on driver performance: An investigation of alarm signal type and threshold level. *Behaviour & Information Technology, 21,* 235-248.

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes, 48*, 193–223.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers, 27*, 46-51.

Helmreich, R. L. (1984). Cockpit management attitudes. *Human Factors, 26*, 583-589.

Ho, G., Wheatley, D., Scialfa, C. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers, 17*, 690–710.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*, 2261-2262.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733-758.

Hollands, J. G., & Spence, I. (1992). Judgments of change and proportion in graphical perception. *Human Factors, 35*, 313-334.

Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: The summation model. *Applied Cognitive Psychology, 12*, 173-190.

Hudson, L., Bateman, F., Bergstrom, P., Cerra, F., Glover, J., Minniti, R., Seltzer, S., & Tosh, R. (2012). Measurements and standards for bulk-explosives detection. *Applied Radiation and Isotopes, 70*, 1037-1041.

Huerta, E., Glandon, T., & Petrides, Y. (2012). Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems, 13*, 316-333.

Hurst, K., & Hurst, L. (1982). *Pilot error: The human factors*. New York, NY: Aronson.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Jen, C.-C., & Yu, S.-S. (2015). Automatic detection of abnormal mammograms in mammographic images. *Expert Systems with Applications, 42,* 3048-3055.

Jensen, M., Lowry, P. B., & Jenkins, J. L. (2011). Effects of automated participative decision support in computer-aided credibility assessment. *Journal of Management Information Systems, 28,* 201-234.

Johnson, E. M., Cavanagh, R. C., Spooner, R. L., & Samet, M. G. (1973). Utilization of reliability measurements in Bayesian inference: Models and human performance. *IEEE Transactions on Reliability, 22,* 176-183.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors, 35,* 221-242.

Koehler, D. J., & James, G. (2014). Probability matching, fast and slow. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 103-131). San Diego, CA: Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142,* 573-603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Waltham, MA: Academic Press.

Lacson, F. C., Gonzalez, C., & Madhavan, P. (2008). Framing and context effects in visual search training. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 348-352). Santa Monica, CA: Human Factors and Ergonomics Society.

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 482-486). Santa Monica, CA: Human Factors and

Ergonomics Society.

Latane, B., Williams, K. D., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*, 822–832.

Lee, J. D. (2006). Human factors and ergonomics in automation design. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp.1570-1596). Hoboken, NJ: Wiley.

Lee, J. D. (2008). Review of a pivotal human factors article: "Humans and automation: Use, misuse, disuse, abuse" *Human Factors, 50*, 404-410.

Lee, J. D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human–Computer Studies, 1*, 153–184.

Lee, J. D., & Sanquist, T. F. (2000). Augmenting the operator function model with cognitive operations: Assessing the cognitive demands of technological innovation in ship navigation. *IEEE Transactions on Systems, Man, & Cybernetics – Part A: Systems & Humans, 30*, 273-285.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors, 46*, 50-80.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge, UK: Cambridge University Press.

Lees, M.N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics, 50*, 1264-1286.

Lehto, M. R., Papastavrou, J. D., Ranney, T. A., & Simmons, L. A. (2000). An experimental comparison of conservative versus optimal collision avoidance warning system thresholds.

Safety Science, 36, 185-209.

Leveson, N. G. (1995). *Safeware: System safety and computers*. New York, NY: Addison-Wesley.

Li, D., Rau, P. P., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics, 2*, 175-186.

Lin, L., Vicente, K. J., & Doyle, D. J. (2001). Patient safety, potential adverse drug events, and medical device design: A human factors engineering approach. *Journal of Biomedical Informatics, 34*, 274-284.

Liu, Y., Fuld, R., & Wickens, C. D. (1993). Monitoring behaviour in manual and automated scheduling systems. *International Journal of Man-Machine Studies, 39*, 1015-1029.

Liu, X., & Zeng, Z. (2015). A new automatic mass detection method for breast cancer with false positive reduction. *Neurocomputing, 152*, 388-402.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin, 107*, 401-413.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Madhavan, P., & Wiegmann, D. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors, 49*, 773–785.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors, 43*, 217-226.

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors, 45*, 281-295.

McCarley, J. S. (2009). Response criterion placement modulates the benefits of graded alerting systems in a simulated baggage screening task. In *Proceedings of the Human Factors and*

Ergonomics Society 53rd Annual Meeting (pp. 1106-1110). Santa Monica, CA: Human Factors and Ergonomics Society.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15,* 302-306.

McCarley, J. S., Rubinstein, J., Steelman, K. S., & Swanson, L. (2011). Estimating user's preferred response bias in an automated diagnostic aid: A psychophysical approach. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting* (pp. 326-329). Santa Monica, CA: Human Factors and Ergonomics Society.

McCarley, J. S., Steelman, K. S., & Rubinstein, J. (2013, October). *Target detection aids in a visual search task: A comparison of fixed-threshold and adjustable-threshold aids*. Paper presented at the Annual Meeting of the Human Factors and Ergonomics Society, San Diego, CA.

McClumpha, A., & James, M. (1994). Understanding automated aircraft. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Recent research and trends* (pp. 314-319). Hillsdale, NJ: Erlbaum.

McDaniel, J. W. (1988). Rules for fighter cockpit automation. In *Proceedings of the IEEE National Aerospace and Electronics Conference* (pp. 831-838). New York, NY: IEEE.

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors, 53,* 356-370.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55,* 520-534.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50,* 194-210.

Merritt, S., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors, 57,* 34-47.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43,* 563-572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46,* 196-204.

Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors, 44,* 343-353.

Meyer, J., & Sheridan, T. B. (2017). The intricacies of user adjustments of alerting thresholds. *Human Factors, 59,* 901-910.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38,* 311-322.

Montgomery, D. A. (1999). Human sensitivity to variability information in detection decisions. *Human Factors, 41,* 90-105.

Montgomery, D. A. (2001). Sampling methods for identifying differences in source reliability. *Journal of General Psychology, 128,* 5-20.

Montgomery, D. A., & Sorkin, R. D. (1996). Observer sensitivity to element reliability in a mutlielement visual display. *Human Factors, 38,* 484-494.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science, 1,* 354-365.

Mosier, K., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201-220). Hillsdale, NJ: Erlbaum.

Muir, B. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*, 527-539.

Muir, B. M. (1988). Trust between humans and machines, and the design of decision aids. In E. Hollnagel, G. Mancini, & D. D. Woods (Eds.), *Cognitive engineering in complex dynamic worlds* (pp. 71-83). London, UK: Academic.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*, 1905–1922.

Muir, B. M., & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*, 429–460.

Murrell, G. A. (1977). Combination of evidence in a probabilistic visual search and detection task. *Organizational Behavior and Human Performance*, *18*, 3-18.

Myung, I.J., & Pitt, M. A. (1997). Applying Occam's razor in modelling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Nercessian, S., Panetta, & Agaian, S. (2008). Automatic detection of potential threat objects in x-ray luggage scan images. In *IEEE Conference on Technologies for Homeland Security* (pp. 504-509). Waltham, MA: IEEE.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors, 53*, 338-355.

Nishikawa, R. M. (2007). Current status and future directions of computer-aided diagnosis in mammography. *Computerized Medical Imaging and Graphics, 31*, 224-235.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*, 413–417.

Paivio, A. (1990). *Mental representations: A dual-coding approach*. New York, NY: Oxford.

Papastavrou, J. D., & Lehto, M. R. (1996). Improving the effectiveness of warnings by increasing the appropriateness of their information content: Some hypotheses about human compliance. *Safety Science, 21*, 175-189.

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics, 43*, 931-951.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics, 40*, 390-399.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*, 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3*, 1-23.

Parasuraman, R., Mouloua, M., Molloy, R., & Hilburn, B. (1996). Monitoring of automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 91-115). Mahwah, NJ: Lawrence Erlbaum.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics, 30*, 286-297.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556-569.

Pawitan, Y. (2013). *In all likelihood: Statistical modelling and inference using likelihood.* New York, NY: Oxford.

Pereira, D. C., Ramos, R. P., & do Nascimento, M. Z. (2014). Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer Methods and Programs in Biomedicine, 114*, 88-101.

Plummer, M. (2015). JAGS Version 4.0.0 user manual. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

Pollack, I., & Madans, A. B. (1964). On the performance of a combination of detectors. *Human Factors, 6*, 523-531.

Prinzel, L. J., Freeman, F. C., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2000). A closed-loop system for examining psychophysical measures for adaptive task allocation. *International Journal of Aviation Psychology, 10*, 393-410.

Qiu, L., & Benbasat, I. (2010). A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human-Computer Studies, 68*, 669–688.

Ragsdale, A., Lew, R., Dyre, B. P., & Boring, R. L. (2012). Fault diagnosis with multi-state alarms in a nuclear power control simulator. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (pp. 2167-2171). Santa Monica, CA: Human Factors and Ergonomics Society.

Rau, P. L., Li, Y., & Li, D. (2009). Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior, 25*, 587-595.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied, 17*, 320-331.

Riley, V. (1989). A general model of mixed-initiative human-machine systems. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 124-128). Santa Monica, CA: Human

Factors and Ergonomics Society.

Riley, V. (1994). A theory of operator reliance on automation. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Recent research and trends* (pp. 8-14). Hillsdale, NJ: Erlbaum.

Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. In R. E. Eberts, & C. G. Eberts (Eds.), *Trends in ergonomics/human factors II* (pp. 75-82). Amsterdam: North-Holland.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47,* 877-903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356-374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49,* 76-87.

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 676-696.

Sanquist, T. F., Minsk, B., & Parasuraman, R. (2008). Cognitive engineering in radiation screening for homeland security. *Journal of Cognitive Engineering and Decision Making, 2,* 204-219.

Scerbo, M.W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 37-63). Hillsdale, NJ: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130,* 380-400.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors, 39*, 14-29.

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics, 38*, 608–625.

Sheridan, T. (1980). Computer control and human alienation. *Technology Review, 10*, 61-73.

Sheridan, T. B. (1988). Trustworthiness of command and control systems. In *Proceedings of IFAC Man–Machine Systems* (pp. 427-431). Oulu, Finland: International Federation of Automatic Control.

Sheridan, T. B. (1992). *Telerobotics, automation and supervisory control*. Cambridge, MA: MIT Press.

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. New York, NY: Wiley.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99-118.

Simonson, I. & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51*, 416-446.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision making? *International Journal of Human-Computer Studies, 51*, 991-1006.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*, 701–717.

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes, 60*, 1-13.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108,* 183-203.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30,* 445-459.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1,* 49-75.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology, 5,* 61-77.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B, 64,* 583-639.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332-336). Santa Monica, CA: Human Factors and Ergonomics Society.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31,* 137-149.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied, 17,* 71-96.

Tanner, W. P., & Birdsall, T. G. (1958). Definitions of *d'* and η as psychophysical measures. *The Journal of the Acoustical Society of America, 30,* 922.

Tetlock, P. E. & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52,* 700-709.

Todd, P., & Benbasat, I. (1994). The influence of decision aids on choice strategies: An experimental analysis of the role of cognitive effort. *Organizational Behavior and Human Decision Processes, 60,* 36-74.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and psychology of choice. *Science, 211*, 453–458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business, 59*, S251–S278.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*, 1206-1220.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101-118.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213-217.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*, 348-365.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*, 135-138.

Walter, S., Wendt, C., Böhnke, J., Crawcour, S., Tan, J.-W., Chan, A., . . . Traue, H. C. (2014). Similarities and differences of emotions in human–machine and human–human interactions: What kind of emotions are relevant for future companion systems? *Ergonomics, 57*, 374–386.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting methods for the analysis of reliance on automation. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 287-291). Santa Monica, CA: Human Factors and Ergonomics Society.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors, 51*, 281-291.

Wells, K., & Bradley, D. A. (2012). A review of x-ray explosives detection techniques for checked baggage. *Applied Radiation and Isotopes, 70*, 1729-1746.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*, 291-298.

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). New York, NY: HarperCollins.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford.

Wickens, C., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors, 49*, 839-850.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*, 201-212.

Wickens, C. D., Gordon, S. E., & Liu, Y. (1997). *An introduction to human factors engineering*. Harlow, UK: Longman.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D., Lee, J. D., Liu, Y., & Becker, S. E. G. (2004). *An introduction to human factors engineering*. New Jersey, NJ: Prentice-Hall.

Wickens, C. D., Maver, A., Parasuraman, R., & McGee, J. (1998*). The future of air traffic control: Human operators and automation*. Washington, DC: National Academy Press.

Wickens, C. D., & McCarley, J. S. (2008). *Applied attention theory*. Boca Raton, CRC Press,

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors, 51*, 446–462.

Wickens, C. D., Thomas, L. C., & Young, R. (2000). Frames of reference for display of battlefield terrain and enemy information: Task-display dependencies and viewpoint interaction use. *Human Factors, 42*, 660-675.

Wiczorek, R. (2017). Investigating users' mental representation of likelihood alarm systems with different thresholds. *Theoretical Issues in Ergonomics Science*, *18,* 221-240.

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors, 56*, 1209-1221.

Wiczorek, R., Manzey, D., & Zirk, A. (2014). Benefits of decision-support by likelihood versus binary alarm systems: Does the number of stages make a difference? In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting* (pp. 380-384). Santa Monica, CA: Human Factors and Ergonomics Society.

Wiegmann. D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors, 44*, 44-50.

Wiegmann, D. A., McCarley, J. S., Kramer, A. F., & Wickens, C. D. (2006). Age and automation interact to influence performance of a simulated luggage screening task. *Aviation, Space, and Environmental Medicine, 77*, 825-831.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*, 352-367.

Wiener, E. L. (1981). Complacency: Is the term useful for air safety? In *Proceedings of the 26th Corporate Aviation Safety Seminar* (pp. 116–125). Denver, CO: Flight Safety Foundation.

Wiener, E. L. (1985). Beyond the sterile cockpit. *Human Factors, 27*, 75-90.

Wiener, E. L. (1988). Cockpit automation. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation* (pp. 433-461). San Diego, CA: Academic Press.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101–126.

Wong, C. Y., & Seet, G. (2017). Workload, awareness and automation in multiple-robot supervision. *International Journal of Advanced Robotic Systems*, 1-16.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*, 2371-2393.

Woods, D. D. (1996). Decomposing automation: Apparent simplicity, real complexity. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 1-17). Mahwah, NJ: Erlbaum.

Zanna, M. P., & Rempel, J. K. (1988). Attitudes: A new look at an old concept. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 315-334). New York, NY: Cambridge University Press.

Zhang, Y., Antonsson, E. K., & Grote, K. (2006). A new threat assessment measure for collision

avoidance systems. In *IEEE Intelligent Transportation Systems Conference* (pp. 968-975). Toronto, ON: IEEE.

Zuboff, S. (1988). *In the age of smart machines: The future of work technology and power*. New York, NY: Basic Books.