AN ABSTRACT OF THE THESIS OF

Fernando Munoz Gomez Andrade for the degree of Master of Science in Psychology presented on June 09, 2023.

Title: Not Good Enough: Ironic Efficiency in Automated-Aided Signal-Detection.

Abstract approved: _____

Jason S. McCarley

During applied signal-detection (e.g., airport-baggage screening) human operators can be assisted in their decision-making process by automated devices. Automation implementation is aimed at increasing performance relative to unaided levels. Generally, this intended effect is empirically observed. However, operators consistently fall short of optimal levels of aided performance, indicating suboptimal aid-use efficiency. Previous research suggests aid-use efficiency might vary depending on the sensitivity levels of each agent in the human + automation team. In the present research we manipulated Task Difficulty (easy vs. difficult) and Aid Reliability (low vs high) to examine how measures of sensitivity and aid-use efficiency vary across these factors. Participants completed a numerical signal-detection task with automated-support manipulated within-subjects. Bayesian inference analyses suggested higher sensitivity gains were achieved at higher levels of difficulty and aid reliability. Interestingly, however, aid-use efficiency was lower at these conditions. These findings replicate and expand previously observed ironic patterns of aided performance where operators fall shorter of optimal levels in conditions where empirical and potential levels of aid-benefit are higher. These findings provide valuable insight for system designers and highlight the need to better understand factors contributing to suboptimal human-automation interaction during aided signal-detection to procure safety and efficiency in naturalistic settings.

Not Good Enough:

Ironic Efficiency in Automated-Aided Signal-Detection.

by

Fernando Munoz Gomez Andrade

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 09, 2023
Commencement June 2024

Master of Science thesis of Fernando Munoz Gomez Andrade presented on June 09, 2023

APPROVED:

Major Professor, representing Psychology.

Head of School of Psychological Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Fernando Munoz Gomez Andrade, Author

ACKNOWLEDGEMENTS

I would like to express my appreciation to Jason McCarley, my advisor, for his support and advice on this project and throughout my education at Oregon State. We also thank Christopher Wickens and my committee for their feedback on this project.

TABLE OF CONTENTS

Section Page

1 Introduction..... 1

1.1 Signal-Detection Theory (SDT)..... 2

1.2 Automated Aids 6

1.3 Human-Automation Interaction 8

1.3.1 Optimal Human-Automation Interaction 9

1.4 Task Difficulty and Aid Reliability 11

1.5 The Present Research.....12

2 Method 14

2.1 Preregistration and Open Data 14

2.2 Participants.....14

2.3 Numeric Signal-Detection Task.....15

2.2 Task Difficulty & Aid Reliability.....16

2.3 Procedure.....18

3 Analysis.....19

3.1 Data Exclusions.....19

3.2 Analyses of Performance.....19

4 Results.....22

4.1 Raw Measures of Sensitivity (*d*).....23

4.2 Optimal vs. Empirical Sensitivity.....26

4.3 Aid-Use Efficiency.....29

TABLE OF CONTENTS (CONTINUED)

Section Page

5 Discussion..……...……..31

5.1 Implications for Human-Automation Interaction……...…….32

5.1 Applied Implications ……...……...…….33

5.1 Generality Constraints and Future Directions ……...…….34

5.2 Conclusion……...……....35

References…….……36

Appendices ……43

Appendix A: Summary Statistics for *d''*……...……...…….44

Appendix B: Summary Statistics for Aid-Effect and Efficiency..…..……....45

LIST OF FIGURES

Figure Page

1 Possible Decisions during Signal Detection……..……...3

2 The Equal-Variances Gaussian Model of SDT……..……4

3 Bias and Criterion Types in SDT……...5

4 Levels of Information Processing for Assistive Automation……..…7

5 Cue Types for Decision-Aids ……..…..….8

6 Optimal Aided Detection under the CC model…….……...10

7 Sample Aided Trial……..……16

8 Manipulation of Task Difficulty…….……..….17

9 Experimental Procedure Flow Chart……18

10 Visualization of Raw Sensitivity Measures…….……..….23

11 Visualization of Aid-Effect……....……24

12 Optimal vs. Empirical Aided Sensitivity……...….27

13 Aid-Use Efficiency ……...….29

LIST OF TABLES

Table Page

1 Classification of Bayes Factors……..……20

2 Model Comparison Summary for Bayesian ANOVA on *d'*…..……24

3 Analyses of Effects Output for ANOVA on aided *d'*….……26

4 Model Comparison Summary for Bayesian ANOVA on aided *d'*……28

5 Analyses of Effects Output for ANOVA on aided *d'*……....28

6 Model Comparison Summary for Bayesian ANOVA on η…..……30

7 Analyses of Model-Averaged Effects Summary for ANOVA on η…..31

Not Good Enough: Ironic Efficiency in Automated-Aided Signal-Detection.

Signal-detection tasks are ubiquitous in everyday and professional life. While driving, we monitor the environment for collision hazards such as vehicles present in our trajectory or blind spot during lane changing (e.g., Zhan et al, 2006). In a professional setting, air-traffic control operators monitor air vessel trajectories for dangerous collision patterns (Rovira & Parasuraman, 2007). As new technological developments arise, human operators might be more frequently assisted in such signal detection processes by automated aids. For example, a side collision warning system might alert drivers of a vehicle present in their blind spot or an automated warning device might signal to an air traffic controller that a conflict between planes is developing. Automated devices may also assist airport security personnel during baggage screening (e.g., Bartlett & McCarley, 2018; Wiegmann et al., 2006), financial advisors while examining records for financial fraud (Bell & Carcello, 2000; Glancy & Yadav, 2011), military personnel distinguish between friend from foe (Dzindolet et al., 2001; Neyedli et al., 2011), nuclear power plant operators monitor dangerous environments (Lee et al., 2007), and medics evaluating ECG feedback for cardiac anomalies (e.g., Bond et al, 2018; Novotny et al., 2017). Furthermore, assistive automation can be implemented in various other aspects within the fields of defense (e.g., Rovira et al., 2007; Yeh & Wickens, 2001), medicine (Carol et al., 103; Kunar, 2022), dentistry (e.g., Araki et al., 2010), transportation and aeronautics (e.g., Dixon & Wickens, 2006; Maltz & Shinar, 2004), etc.

The general purpose of automation support in these settings is to improve detection performance relative to that achieved by the unaided human operator. More specifically, automation implementation in these tasks aims to increase *sensitivity* (Hautus et al., 2021), the ability to distinguish between two mutually exclusive states of the world (noise and signal + noise). Empirically, as expected, automated aids generally improve the decision maker's performance, allowing aided sensitivity to surpass unaided levels. However, operators also consistently fail to achieve maximum potential levels of aid-benefit, letting empirical performance fall short of optimal (see Boskemper et al., 2022; Duncan-Reid & McCarley, 2021, 2022; Gyles & McCarley, 2019; Munoz Gomez Andrade et al., 2022; etc). This shortcoming translates to operators making unnecessary mistakes and a reduction in the benefit-cost ratio of developing and deploying automated aids. Psychologically, this shortcoming is indicative of

suboptimal human-automation interaction and highlights the need to identify factors that encourage efficient aid-use. In the present research, we explored how two factors, the individual sensitivities of the automation and the operator, affect various aspects of aided performance including aid-use efficiency. We seek to expand on previous research to best guide system designers and ultimately procure safety and efficiency in naturalistic settings. To measure performance in the present research, we relied on the framework of Signal-Detection Theory (SDT, Green & Swets, 1966).

Signal-Detection Theory

In a standard signal-detection task, human operators need to differentiate between two mutually exclusive states of the world based on probabilistic evidence. Specifically, they are required two differentiate between *noise* and *signal + noise* states, which we will hereafter refer to simply as signal (Hautus et al., 2023). For instance, during airport baggage screening, security operators evaluate pieces of baggage containing non-prohibited items (noise) for the presence of prohibited items (signal). Importantly, the evidence distinguishing noise and signal states is often ambiguous, making incorrect decisions inevitable. For example, some prohibited items might closely resemble the visual characteristics of non-prohibited items, or conversely, non-prohibited items may resemble weapons or other threats.

SDT (Hautus et al., 2023; Wickens, 2001) provides a two-stage framework for modeling and characterizing detection performance. In the first stage, operators assess the strength of evidence for or against the presence of a signal, encoding it as a scalar decision variable. Then, this decision variable is gauged against a criterion (λ) to decide if a signal state is present. Scalar values above λ result in a yes-judgment (signal-present) while those below λ result in a no-judgment (signal-absent). For each trial, operators may arrive to one of four decisions depending on the true state of the world and their detection judgment (see Figure 1). Correct decisions refer to judgments which match the true state of the world. When operators correctly arrive at a yes-judgment, their decision is termed a hit. Incorrect yes-judgments are termed false alarms. A signal-absent judgment is referred as either a correct rejection or a miss depending on the true state of the world. When the total number of signal and noise trials is known, hit rates (HR) and false alarm rates (FAR) can be calculated through the following formulas:

$$HR = \frac{\sum Hits}{\sum Signal\ Trials};\ FAR = \frac{\sum False\ Alarms}{\sum Noise\ Trials},$$

where the sum of hits or false alarms is divided by the sum of signal or noise trials, respectively. In SDT, these statistics allow for the estimation of relevant measures of performance.

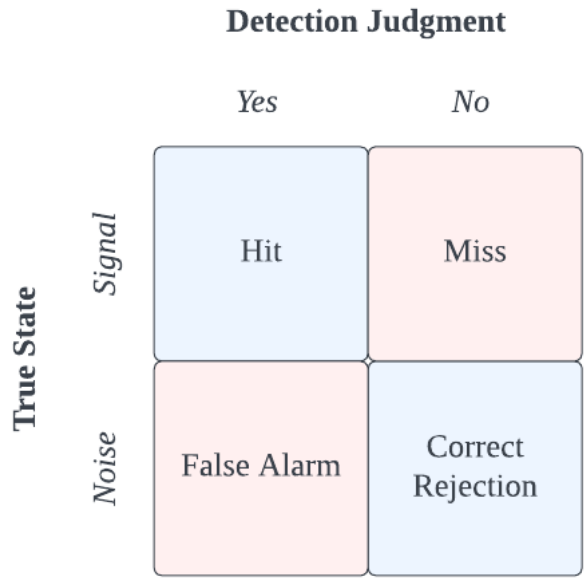

Figure 1. Possible decisions in yes or no signal-detection tasks. Correct decisions are depicted in a blue background, while erroneous decisions are depicted in a red background.

The standard form of the SDT framework is the equal-variance Gaussian model (Wickens, 2001), which assumes the evidence distributions for signal and noise states are normally distributed with a common standard deviation. By convention the noise distribution is assigned a mean of 0, and the mean of the signal distribution is assumed to be equal to or greater than 0. Figure 2 portrays a graphical representation of this model, with the horizontal bottom line corresponding to potential values for scalar decision variables.

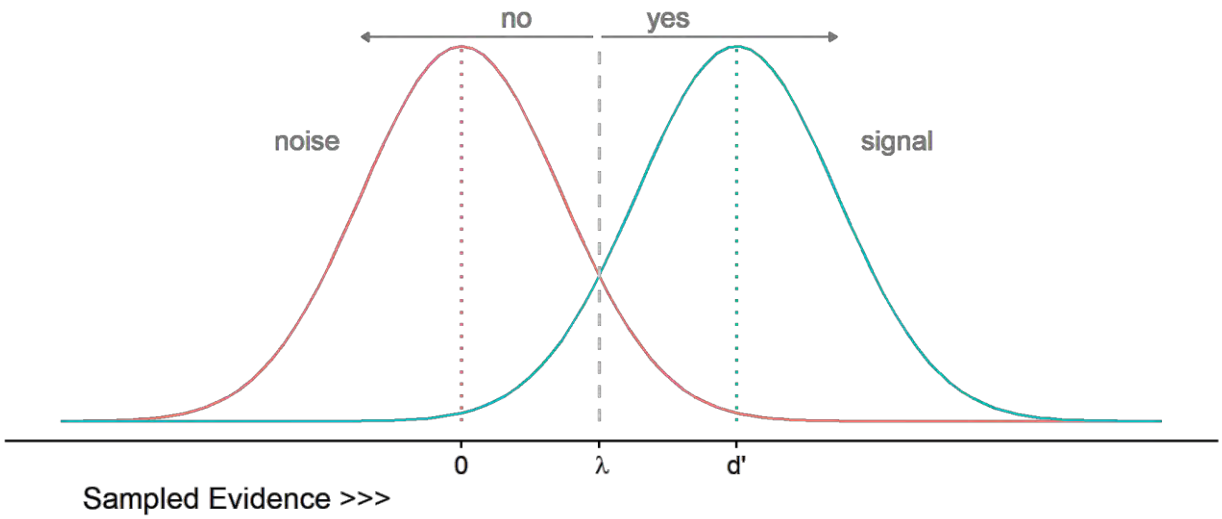

Figure 2. Equal-variances Gaussian model of signal detection. The horizontal line depicts potential values for scalar variables of encoded evidence. The evidence distribution for signal is depicted in blue, while that of noise is depicted in red. λ depicts operator's criterion to arrive to a yes judgment.

Notice that in Figure 2, the evidence distributions for signal and noise overlap. The area of overlap represents confusability between states of the world. If the curves overlapped perfectly, noise and signal states would be indiscernible. As overlap between distributions decreases, the sensitivity of detection agents increases. Under this model, sensitivity can be characterized by *d'*, the distance between the means of the two curves as measured in units of the common standard deviation. The value of *d'* is given by,

$$d' = Z(HR) - Z(FAR).$$

Performance can be further characterized by the operator's response bias, or their willingness to arrive at a yes judgment. Consider two hypothetical operators, one with relatively higher response rates, *HR* = 0.90 and *FAR* = 0.42, and another one with lower rates such that HR = 0.60 and FAR = 0.11. Applying the above formula gives the same value of sensitivity for both operators: *d'* = .48. What differs between these operators is their placement of λ, the criterion required for a yes judgment. Participants might adopt different criteria depending on the nature of a task and the payoffs associated with different responses.

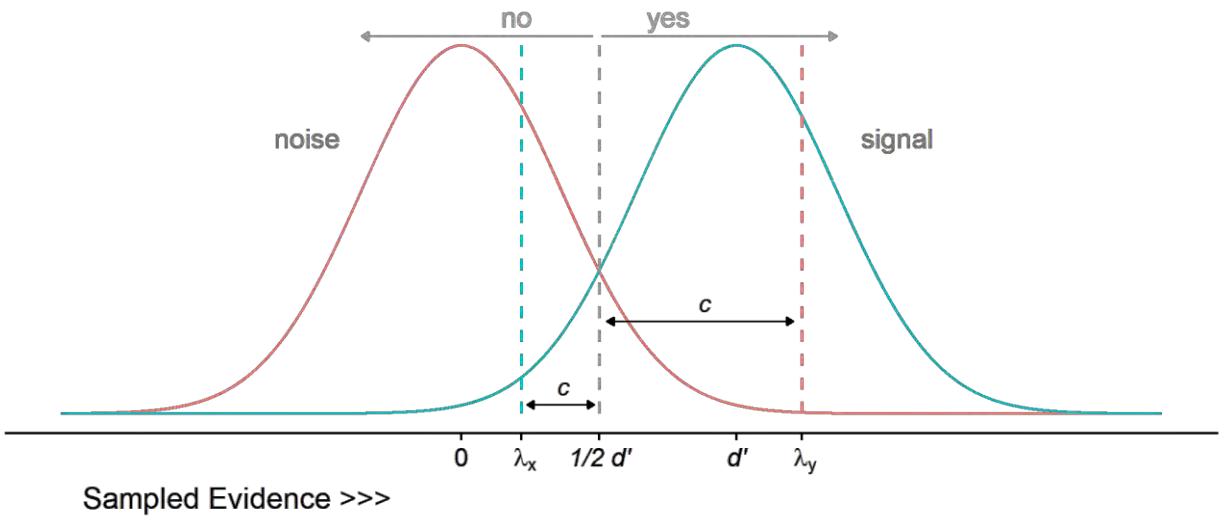

Figure 3. Equal-variances Gaussian model depicting different criterions. The blue dashed line at λ_x represents a liberal criterion, while the red line at λ_y represents a conservative criterion. The grey dashed line at 1/2 *d'* depicts an unbiased criterion.

Placement of response criterions (λ) can be described in three ways, depending on their position relative to ½ *d'*, the point equidistant from the means of the noise and signal distributions (see Figure 3). λ is said to be unbiased when it equals ½ *d'*. Liberal criteria are set below this point and increase the probability of arriving to a yes-judgment, thus resulting in more Hits but also more False Alarms. Conservative criteria are set above unbiased levels and decrease the probability of yes-judgments, leading to fewer False Alarms but also fewer Hits. In Figure 3, λ_x and λ_y depict liberal and conservative criteria, respectively. When signals are more common than noise events, or when Hits have high payoffs relative to the costs of False Alarms, a liberal λ is more appropriate. Conversely, when signals are rare or when rewards for Correct Rejections are high relative to the cost for Misses, a conservative λ is more convenient for operators.

The distance between ½ *d'* and λ is termed *c* and provides a measure of bias in the response strategies adopted in signal-detection tasks. Under the assumptions of the equal-variances Gaussian model, $c = \lambda - \frac{1}{2}d'$. Liberal criteria result in negative values of *c*, while conservative criteria result in positive values. Bias may also be measured by *β*, the likelihood ratio between the probability densities of signal and noise at λ. This measure is given by the formula $\beta = \frac{f_s(\lambda)}{f_n(\lambda)}$, where $f_s(\lambda)$ and $f_n(\lambda)$ denote the probability density functions for signal (*s*)

or noise (*n*) events at λ. In the Gaussian model of SDT, β corresponds to the relative height of the signal and noise curves at λ. When λ = ½ *d'*, the heights of the two curves are equal and $\beta = 1$. Liberal and conservative criterion shifts decrease or increase β, respectively.

Assuming equal response payoffs, optimal λ (λ^*) in unaided signal-detection is that which maximizes the probability of a correct response (Wickens, 2001). Under the equal-variances Gaussian SDT model, λ^* is determined by the prior probability of a signal occurring such that: $\lambda^* = \frac{1}{2}d' - \frac{logit(s)}{d'}$, where s denotes the probability of signal. λ^* corresponds to β^*, the point at which β is equal to $\frac{p(n)}{p(s)}$, with $p(.)$ denotes the probability of noise or signal events. When $p(n) = p(s)$, their ratio = 1, thus $\beta^* = \frac{1}{2}\ d'$ and λ^* is unbiased. When $p(n) > p(s)$, λ^* is liberal and $\beta^* < 1$, while a lower probability of signal relative to noise leads to a conservative λ^* and $\beta^* > 1$.

Automated aids

Human operators engaged in signal-detection tasks can be assisted by automation in the form of decision aids. Automation here is defined as the replacement of a process that could have been carried out by a human operator (Parasuraman et al., 2000). Automated aids can greatly vary in form from one task to another. In aviation, for instance, automation can assist operators through auditory collision alerts, but may also altogether takeover flight duties at times. We here offer a brief discussion of aid taxonomy as introduced in Parasuraman et al. (2000) to best delimit the form of automation the present research focuses on.

The taxonomic model introduced in Parasuraman et al. (2000), classifies automation based on two criteria: 1) its degree of autonomy relative to operators'; 2) the stage of information processing at which it offers assistance. According to this model, automation autonomy ranges from low to high across ten levels. At lowest levels, automation offers no assistance with all actions and decisions undertaken by the human operator. At highest levels, automation decides the course of action fully autonomously, ignoring human operators. Of particular interest here is level four, at which automation suggests a course of action amongst a set of alternatives but defers decision-making and action implementation to operators.

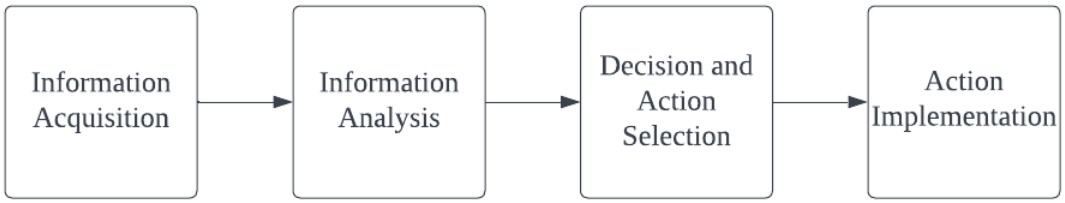

Figure 4. Stages of information processing from earliest to latest; adapted from Parasuraman et al., (2000).

In terms of information processing stages, Parasuraman et al. (2000) classified aids across four levels (see Figure 4). During the information acquisition stage, automation assists operators in sensory processing. For instance, during baggage screening, automation might highlight areas of an x-ray image for operator review (e.g., Huegli et al., 2020). During information analysis, automation assists in perceptual and working memory operations (e.g., Morphew & Wickens, 1998; Wickens & Colcombe, 2007). For example, drivers might be assisted by automaton that visualizes their trajectory while driving in reverse. At the decision and action selection stage, automation recommends a course of action based on the information assessed at previous levels. In aviation, for example, automation might suggest specific collision avoidant maneuvers to pilots. At the highest levels of information processing, automation might overtake several aspects of action implementation once a decision has been made. For instance, automation engrained in automotives might automatically break if an imminent collision is detected. In this research, we investigated variations in aided performance with automation that assists during the decision-making stage of information processing.

Across both dimensions of aid taxonomy, we focused on automation which assists during decision making but defers action selection and implementation to operators. Furthermore, we specifically focused on automation that assists during yes-or-no signal-detection processes, which are common in the literature and in applied settings (e.g., Elvers & Elrif, 1997; Yamani & McCarley, 2018). Such aids can provide feedback to operators through either binary or graded cues (see Figure 5). Graded automation communicates the strength of sampled evidence through, for example, confidence ratings or likelihood messages (e.g., Bartlett & McCarley, 2019; Sorkin et al., 1988; Wiczorek & Manzey, 2014). On the other hand, binary automation provides absolute judgments (e.g., “signal present”) without communicating the strength of sampled evidence (e.g., Meyer, 2001; Munoz Gomez Andrade et al., 2022). Graded automation allows for higher

potential aided sensitivity as greater information sharing in the human + automation system helps clarify evidence ambiguity (Robinson & Sorkin, 1985). However, empirical comparisons of performance between the two types of cues are mixed (see Bartlett & McCarley, 2019; Duncan-Reid & McCarley, 2022; Wiczorek & Manzey, 2014). For simplicity, in this research, we focus on aids that provide binary feedback to operators.

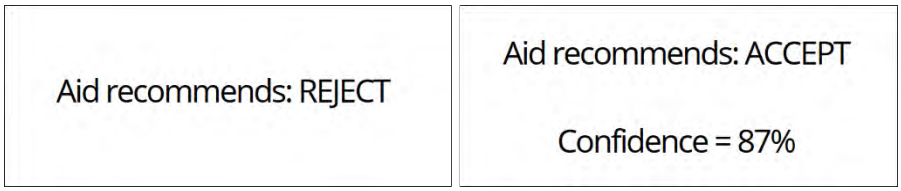

Figure 5. Sample cues for binary and graded automation. The binary aid on the left provides an absolute judgment, while the graded aid on the right communicates the strength of evidence assessed through a confidence rating.

Human-Automation Interaction.

Maximum sensitivity in aided tasks is a function of the individual sensitivity of each agent in the human + automation team (Robinson & Sorkin, 1985). However, it also varies with the patterns of human-automation interaction adopted by human operators. Optimal aided performance refers to human-automation interaction which allows for highest attainable team sensitivity given the individual sensitivities of each agent. However, although optimal human-automation is theoretical attainable to operators, empirical assessments of performance reveal widespread aid-use suboptimality.

Suboptimal human-automation interaction in terms of automation dependence is generally classified into one of two forms: *disuse* or *misuse* (Parasuraman & Riley, 1997). Ideally, dependence on automation will increase as the sensitivity of the automation (aid reliability) increases relative to that of the human operator (Sorkin & Dai, 1994). Operators engage in disuse when they fail to sufficiently rely on automated aids, placing too much weight on their own judgments. Automation disuse in aided signal-detection is commonly observed when aid reliability falls short of perfect (e.g., Rice & McCarley, 2011; Tikhomirov et al., *in press*). Automation misuse is a consequence of suboptimally high automation dependence. During misuse, operators uncritically rely on the diagnoses provided by automation, without checking them against the raw data. Misuse is most often observed when operators are assisted

by automated aids whose sensitivity is near-perfect (Parasuraman, 2000; Parasuraman et al., 1993).

During aided-signal detection, a variety of strategies for automation-interaction are available to human operators (see Bartlett & McCarley, 2017; Duncan-Reid & McCarley, 2022; Tikhomirov et al., *in press*). Such strategies vary in terms of the maximum levels of sensitivity they allow for the human + automation team. When aided by binary automation, the Contingent Cutoff (*CC*) model proposed by Robinson and Sorkin (1985; see Murrell, 1977) allows for best-attainable aided sensitivity if dependence is optimally calibrated. It thus provides a benchmark of optimal aided performance against which to assess empirical estimations (e.g., Bartlett & McCarley, 2017).

Optimal Human-Automation Interaction.

The *CC* model assumes the human + automation team engages in a typical process of signal-detection, gauging sampled evidence against a criterion (λ) to derive a yes or no judgment. Under this model, the criterion adopted by operators during aided detection is contingent on the feedback provided by the automated aid. When automation renders a positive (signal-present) cue, operators are assumed to adopt a more liberal λ, thus decreasing required evidence for a yes-judgment. Following a negative (signal-absent) cue, λ is assumed to become more conservative, increasing required evidence for yes-judgments. In a *CC* strategy, criterion shift ($\Delta\lambda$), the extent to which λ is shifted following an aid's cue, provides a measure of automation dependence. Best-attainable aided performance requires optimal calibration of $\Delta\lambda$. We will refer to empirical measures of *d'* for unaided operators as $d'_{\text{Unaided,}}$ while empirical aided sensitivity for the human + automation team is referred to as d'_{Aided}. Optimal levels of sensitivity are described as d'_{Optimal}.

Optimal λ^* values on aided trials depend on the aid's cue and predictive value. Assuming that signal and noise trials are equally common, when an aid renders a positive cue (+), the probability of signal is equal to $p(S|+) = \frac{p(+|s)p(s)}{p(+|s)p(s)+p(+|n)p(n)}$. Conversely, when the aid renders a negative judgment, the probability of a signal is $p(S|-) = \frac{p(-|s)p(s)}{p(+|s)p(s)+p(+|n)p(n)}$. Thus, two different values of λ^* should be adopted by operators, one for positive-cue trials (λ^*_{yes}) and one

for negative-cue trials (λ^*_{no}). When $p(s) = p(n)$ and the aid is unbiased, $p(s|+)$ reduces to the aid's overall reliability, and $p(s|-)$ reduces to one minus the aid's reliability such that:

$$\lambda^*_{yes} = \frac{1}{2}d' - \frac{logit(aid\ reliability)}{d'};\ \lambda^*_{no} = \frac{1}{2}d' - \frac{logit(1 - aid\ reliability)}{d'}.$$

In an ideal *CC* strategy, optimal criterion shift ($\Delta\lambda^*$) is calibrated to optimal measures of λ^* for positive and negative cue trials. Figure 6 depicts λ^* and $\Delta\lambda^*$ values under an optimal *CC* strategy for an aid of .9 reliability and an initial unbiased λ. Notice that λ^*_{yes} is liberal and captures .9 of the signal distribution, while λ^*_{no} is conservative and captures .9 of the noise distribution.

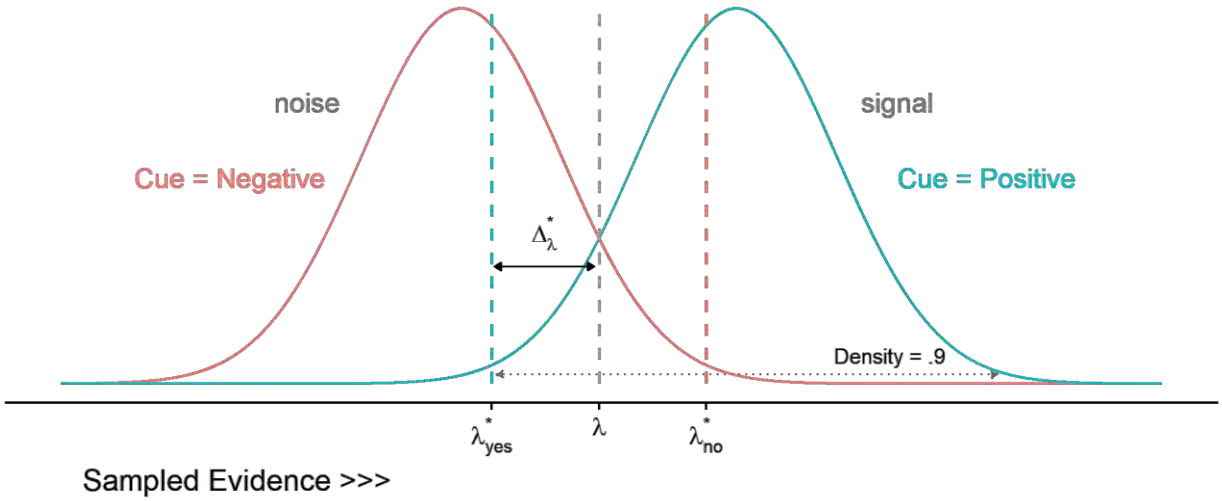

Figure 6. Optimal $\Delta\lambda$ for an aid of .9 reliability. The blue dash at λ^*_{yes} encompasses .9 of the signal distribution, while the red line at λ^*_{no} encompasses .9 of the noise distribution. Assuming an initial unbiased λ, $\Delta\lambda^*$ depicts optimal criterion shift for λ^*_{yes}.

When the frequency of signal trials and aid reliability are known, optimal λ^* for positive and negative cue trials can be calculated given an estimate of $d'_{Unaided}$. Measures of λ^* for positive and negative cue trials allow for an estimation of $d'_{Optimal}$. Furthermore, empirical and optimal estimates of sensitivity, in turn, allow for a quantification of the quality of operator's use as *efficiency* (η; Tanner & Birdsall, 1958):

$$\eta = \left(\frac{d'_{Aided}}{d'_{Optimal}}\right)^2.$$

Where η is a squared ratio of observed to optimal sensitivity in aided conditions. An efficiency value of 1 would indicate perfect efficiency.

Although a *CC* strategy allows for the human + automation team to attain optimal levels of aided *d'*, empirical estimations of $\Delta\lambda$ are more consistent of a sluggish *CC* strategy in which $\Delta\lambda$ is suboptimally low (e.g., Munoz Gomez Andrade et al., 2022; Robinson & Sorkin, 1985). These suboptimal low levels of dependence are symptomatic of suboptimal aid-use efficiency and, more concretely, of automation disuse. Importantly, however, extent of inefficiency displayed by operators has been shown to differ based on various task conditions, such as the sensitivity of individual agents in the human + automation team.

Task Difficulty and Aid Reliability.

Several task factors can affect performance and general aid-use efficiency during aided signal-detection. For instance, higher workloads might increase dependence on automation (McBride et al., 2011; Wickens & Dixon, 2007), whereas a bias toward false alarms might engender disuse (Dixon et al., 2007; Rice & McCarley, 2011). In the present research we focus on how two factors of interest, the individual sensitivity of the human operator and that of the automated device, affect performance during aided signal-detection. As discussed, these factors determine the upper bound on aided sensitivity and dictate the operator's optimal level of aid dependence. As the aid's sensitivity increases relative to operator's, the operator's optimal behavior is to increase automation dependence.

In the *CC* model, higher dependence equates to higher measures of $\Delta\lambda$, that is, higher adjustments of operator's criterion in response to the aid's diagnoses. To optimize d'_{Aided}, higher measures of $\Delta\lambda$ should be observed at increased levels of aid reliability or at decreased levels of unaided sensitivity. Recall that $\Delta\lambda^*$ is determined by λ^* values for positive and negative cue trials. Consider levels of λ^*_{no} when $d'_{Unaided} = 2$, aid reliability = .6, and $p(s) = p(n)$: $\lambda^*_{no} = \frac{1}{2}2 - \frac{logit(1-.6)}{2} = 1.2$. If other variables are held constant but aid reliability increases to .9, then $\lambda^*_{no} = 2.1$. Notice that as reliability goes up, $\Delta\lambda^*$ becomes more extreme (assuming an initial unbiased λ). Even higher measures of $\Delta\lambda$ would be required to match levels of λ^*_{no} if reliability is held constant at .9, but decreases such that $d'_{Unaided} = 1$. In such case, $\lambda^*_{no} = 2.7$. However, empirical estimations under a *CC* framework suggest operators consistently fail to appropriately calibrate dependence as each agents' sensitivity varies.

Unsurprisingly, more reliable aids tend to produce bigger sensitivity gains (e.g., Rice & McCarley, 2011, Rovira et al., 2007). A review of the literature suggests aids around 70% reliability or better tend to produce an aid-benefit in performance, while aid-benefit generally nullifies at lower reliabilities (Wickens & Dixon, 2007). Interestingly, however, Bartlett & McCarley (2021) compared aid-use efficiency measures across reliability levels through a model of optimal human-automation interaction with this kind of automation (see Bahrami et al., 2010; Sorkin et al., 2001). Analyses suggested that, contrary to aid-benefit, aid-use efficiency decreased as automation reliability increased. In recent research by Tikhomirov et al., (*in press*) higher aid-benefit was observed at higher levels of task difficulty, suggesting appropriate higher dependence at lower levels operator sensitivity. However, analyses revealed lack of enough evidence for or against an effect of task-difficulty on aid-use efficiency.

The Present Research

Previous research suggests both aid-reliability and task difficulty may affect performance during aided signal-detection. The literature consistently shows higher levels of aid-benefit as operator sensitivity decreases relative to aid reliability. However, more research is warranted regarding the effects of either factor on aid-use efficiency (η). Previous research by Bartlett & McCarley (2021) examining the effects of reliability on performance employed graded automation, thus highlighting the need for this effect to be replicated with binary automation. In addition, in Tikhomirov et al (*in press*) the data was unable to discern between a positive and a null effect of difficulty on efficiency. More research manipulating this factor may clarify the nature of the relationship between these two variables. Furthermore, no research till date has manipulated both factors within a single experimental design to explore potential interactions on measures of aided performance.

In the present research, we seek to address the above-mentioned gaps in the literature to better inform system designers and future research. To this end, we employed a numeric signal-detection task framed as a quality-control process (e.g., Botzer et al., 2013; 2015) during which the availability of an assistive Automation Aid was manipulated as within-subjects factor. Participants performed a block of trials unaided and a block of trials assisted in their decision-making process by a binary automated aid (e.g., “Aid recommends: Reject”). Additionally, Task

Difficulty (easy vs. difficult), and Aid Reliability (low vs. high) were manipulated as between subject factors, resulting in four different between-subject experimental conditions: Easy/Low Reliability; Easy/High Reliability; Difficult/Low Reliability, Difficult/High Reliability. We assessed the effect of these factors on measures of *d'*, and aid-use efficiency (η). Measures of *d'* were derived through the framework of the equal-variances Gaussian model of SDT discussed above (see signal-detection theory section). Aid-use efficiency (η) was estimated by fitting an ideal *CC* strategy to unaided measures of performance (see optimal human-automation interaction section).

We outlined three specific research objectives:

a) To compare sensitivity measures across levels of Automation Aid, Aid Reliability, and Task Difficulty. This allowed us to gauge the sensitivity of the human + automation team to that of each individual agent. More importantly, these comparisons allowed us to examine if and how aid-effect varies across levels of difficulty and reliability.
b) To contrast empirical vs. optimal measures of aided *d'* across conditions. This allowed us to examine if aid-use inefficiency was widespread or if empirical sensitivity levels approached optimal levels under any conditions.
c) To compare aid-use efficiency across experimental levels for a converging assessment of conditions in which participants came closest to optimal aided performance.

Based on previous literature we derived the following hypotheses for each research objective:

1. For objective a, an interaction between automation-aid and reliability will be observed, such that higher aid benefit will be observed at higher reliability levels. Similarly, an interaction between automation-aid and difficulty will be observed, such that higher aid benefit will be observed at higher difficulty levels.

2. For objective b, we expected a main effect for aided *d'* type such that empirical measures of aided sensitivity differed from optimal estimations across experimental levels.
3. For objective c, we expected lower aid-use efficiency at higher levels of reliability and of difficulty.

Due to a lack of previous research, we did not make specific predictions regarding interaction effects of Automation-Aid × Difficulty × Reliability on measures of *d'* or of an interaction effect of Difficulty × Reliability on aid-use efficiency.

Experimental design, objectives, and hypotheses were pre-registered and are publicly available (see https://osf.io/ztj9u/). An additional research objective included in the pre-registration regarding modeling of automation-use strategies was omitted from the present work and will be addressed in a future composition. Research objective b was derived based on preliminary data from Tikhomirov et al. (*in press*). Inferential analyses in that study suggested absence of evidence to distinguish between a null and a credible effect of task difficulty on efficiency. However, the original hypothesis was included above for transparency.

Method

Preregistration and Open Data

All experimental procedures including experimental hypotheses, sampling plan, exclusionary criteria, and analysis plan were pre-registered prior to commencement of data collection. The pre-registration for this project, raw data for all practice and experimental blocks, and analytic code are publicly available for review and download at the Open Science Framework (see https://osf.io/ztj9u/).

Participants

Participants ($N = 156$) were recruited online through Prolific Academic (www.prolific.co; $n_{Prolific} = 125$) and the student participant pool at pacific-northwestern public university ($n_{SONA} = 31$). After recruitment, participants were redirected to the Gorilla platform (www.gorilla.sc; Anwyl-Irvine et al., 2020) for data collection, which took place exclusively online. Participants were compensated either with half an hour of course credit (SONA participants) or with

monetary compensation ($5 USD; Prolific participants). We pre-specified an initial target sample of 100 participants from Prolific Academy (n = 25 per condition). Recruitment in Prolific was increased following initial data exclusions (see analysis) and attrition to achieve an n for this platform close to the initial target. The student participant pool was open for a period of 18 days to increase statistical sensitivity. No initial target n was preregistered for the student pool. Rather, as many participants were collected as signed up during the recruitment period.

Numeric Signal-Detection Task

Participants completed a numeric signal detection task (see Healy & Kubovy, 1981) framed as a quality control process (see Munoz et al., 2022; Tikhomirov et al., *in press*). Participants were instructed to imagine they were chemists charged with determining if batches of a chemical compound are contaminated. Each trial, they were presented with four readings sampled from one of two distributions (noise or noise + signal). They were informed an average reading higher than 500 is generally indicative of contamination and should lead to rejection, with about ½ of batches expected to be rejected. However, to communicate the ambiguity between noise and noise + signal in the sampled evidence, they were also told mean readings were highly variable with some mistakes being therefore unavoidable. Figure 7 provides an example of an aided trial. Unaided trials were similar to aided trials, except for the absence of feedback provided by automation. Trials were untimed, but participants were instructed to complete as many as possible during each block without sacrificing accuracy or rushing. The time for each block was fixed such that participants could not progress through the study by increasing the speed of their decision-making process. Participants made their responses via button presses. The experiment was set to run on desktop devices only, excluding mobile and tablet devices.

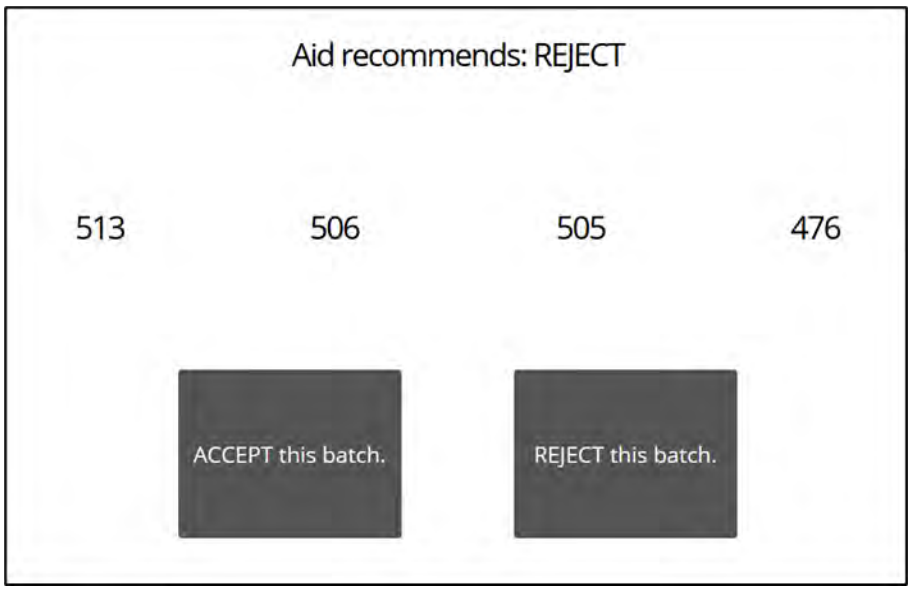

Figure 7. Sample aided trial for the Difficult/High Reliability condition. Unaided trials did not include automation binary feedback visible on the top-center portion.

Task Difficulty & Aid Reliability

In line with previous research (Tikhomirov et al., *in press*), Task Difficulty was manipulated by adjusting the degree of overlap between the noise and noise + signal distributions (see Figure 8). While all distributions had a standard deviation of $\sigma = 20$, μ values for the Easy conditions were closer to each other, resulting in a higher degree of overlap between the distributions. For the Easy conditions, $\mu = 485$ for the noise distribution, and $\mu = 515$ for the noise + signal distribution. For the Difficult conditions, $\mu = 495$ for the noise distribution and $\mu = 505$ for the signal distribution. The ideal strategy for unaided detection was to base decisions on the average of all four readings. This strategy allowed for a maximum $d'_{Unaided} = 3.0$ for the Easy condition and of $d'_{\text{Unaided}} = 1.0$ for the Difficult condition. A strategy in which judgments were based on one randomly selected reading would have yielded maximum d'_{Unaided} values of 1.5 and .5 for the Easy and Difficult conditions, respectively. Previous research employed a similar manipulation and observed a credible effect of Task Difficulty on measures of *d'* (Tikhomirov et al., *in press*).

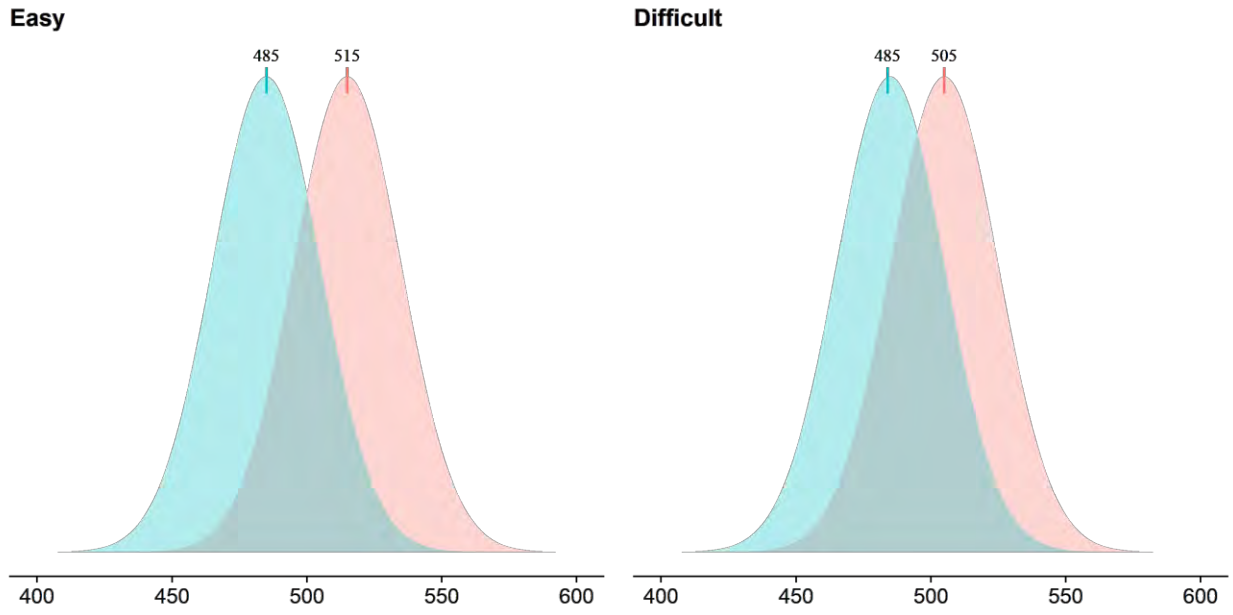

Figure 8. Noise and noise + signal distributions for stimuli used in the Easy (left) and Difficult (right) conditions. Notice that the shaded area depicting the degree of overlap between the two distributions is higher for the Difficult condition, resulting in higher evidence ambiguity for this condition.

Based on previous research (Bartlett & McCarley, 2021; Wickens & Dixon, 2007), aid reliability was set to 77% (d' = 1.5) for the low reliability conditions and to 93% (d' = 3.0) for the high reliability conditions. Aid reliability levels were disclosed to participants prior to completion of aided blocks. Previous research suggests disclosure of automation reliability levels encourages more appropriate response bias (Avril, 2023; Wang et al., 2009). Aid Reliability and Task Difficulty were manipulated as between-subjects rather than as within-subject factors to avoid carryover effects on measures of performance.

Procedure

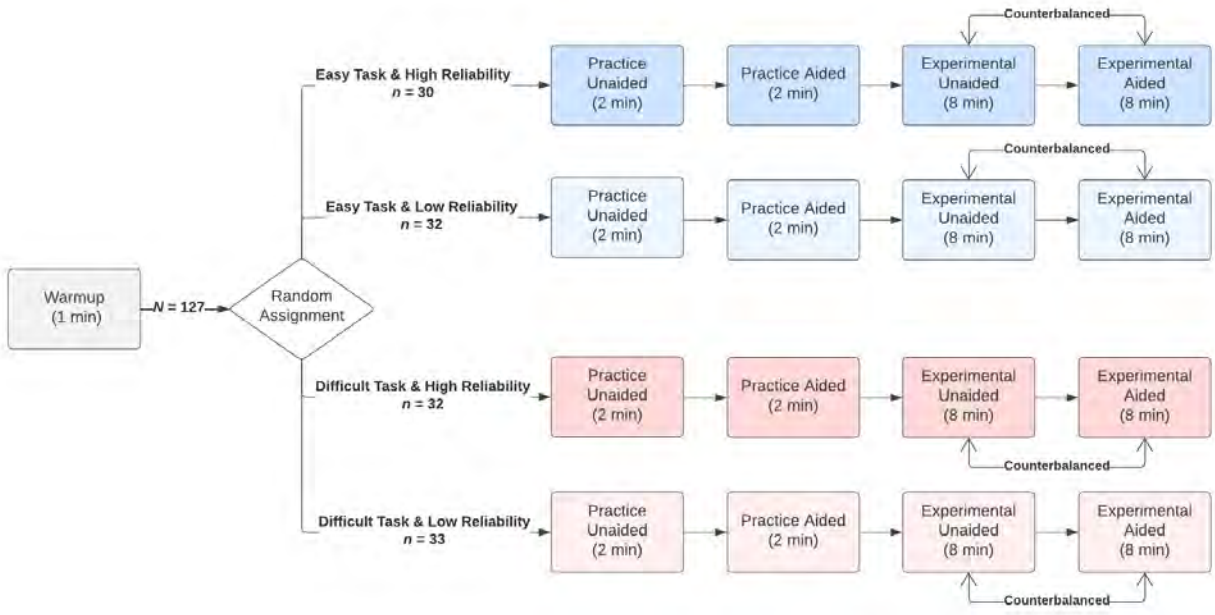

Figure 9. Flow chart of sequence of events for experimental procedures following informed consent and prior to debrief.

A flow chart of the course of events for this experiment is depicted on Figure 9. After providing informed consent, participants received general instructions and completed a block of unaided warmup trials (1 min) for which Task Difficulty was lower than all other conditions. Then, they were randomly assigned to one of two levels of each between-subjects factor: a) Easy/High Reliability; b) Easy/Low Reliability; c) Difficult/Low Reliability; d) Difficult/High Reliability. Participants in the Easy conditions completed all subsequent practice and experimental blocks at a lower difficulty than those in the Difficult conditions. For each condition, levels of Task Difficulty remained stable for all subsequent practice and experimental blocks.

Following random assignment, participants completed two practice blocks of 2 minutes each. They completed the first block while unaided, which allowed them to familiarize themselves with their assigned level of Task Difficulty. They then received instructions on how to complete aided trials (see Figure 7). They were informed that for the next set of practice trials they would perform the same task but would be assisted by a computer aid which would make a

judgment about whether each batch should be accepted or rejected. They were further informed that the aid's judgment might sometimes not match their own. Participants in the High-Reliability conditions were informed that tests have shown the aid is correct about 93% of the time, while those in the Low-Reliability conditions were informed tests have shown that the aid is correct about 77% of the time. All participants were further informed that they should use the aid to help make their decisions, but that they were free to disagree with it if they wished. They then completed a second practice block while aided by automation.

Following completion of the practice blocks, participants completed two experimental blocks of trials (8 min each) with automation aid manipulated within-subjects. The order of experimental blocks for each condition was counterbalanced across participants. For all warmup and practice blocks, participants received performance feedback for correct and incorrect response (e.g., "Oops, that batch should have been rejected"; "Good judgment"). The purpose of this was to allow participants to familiarize themselves with the ambiguous nature of sampled evidence from noise and noise + signal distributions. For experimental blocks, performance feedback was removed to better resemble task conditions in naturalistic settings.

Analysis

Data exclusions

Exclusionary criteria for data analysis were preregistered and consisted of: a) failure to complete the experiment within the pre-established time of 30 mins; b) failure to complete a minimum of 32 trials during either experimental block; c) failure to achieve a *d'* greater than or equal to 0.25 in any experimental block. The threshold for exclusionary criterion b was determined by the quantity corresponding to two *SD* below the mean number of trials completed per person in a similar experiment (see Munoz et al., 2022). Due to a failure to meet criterion a, 25 participants were excluded from analysis. No participants were excluded due to the threshold imposed in criterion b. An additional four participants were excluded per criterion c for a final sample of $N = 127$ ($n_{\text{prolific}} = 98$).

Analysis of performance and efficiency

All analyses on measures of performance were submitted to default Bayesian ANOVAS in JASP (JASP Team, 2019; see van den Bergh et al., 2020). Bayesian hypothesis tests are

comparative analyses in which various models, including prior parameter distributions, are assessed for their ability to account for empirical data (see Ly et al., 2018; Rouder et al., 2009; 2017). In a default Bayesian ANOVA, the null model includes only one parameter representing the grand mean of scores. Competing models include additional parameters for main effects and interactions. In JASP models are excluded when they violate the principle of marginality, which prohibits interaction terms to be present without constituent main effects (van den Bergh et al., 2020). Default analyses place Cauchy priors on the normalized effect sizes (Rouder et al., 2012). This approach specifies equal probability for all competing models prior to data assessment. MCMC effective sample size (ESS), the effective length of MCMC chains, was in line with the recommendations outlined in Kruschke (2021; ESS ≥ 10,000). Visual inspection Q-Q plots for all analyses conducted indicated no obvious deviations from normality.

We report several summary outputs for all primary analyses conducted. For model comparisons, we report the posterior model probability, $P(Mi|\text{data})$, of each model *i* under comparison. Posterior model probabilities were calculated under the assumption that prior probabilities of all models were uniform. We also report *Bayes Factors* (*BF*), which are likelihood ratios that quantify the predictive performance of the competing models: $\frac{p(D|M_0)}{p(D|M_1)}$. A BF_{01} of 10 indicates that the likelihood of the reference model (M_0) was 10x the likelihood of the comparison model (M_1). Unless otherwise specified, the reference model (M_0) for each analysis was that which showed best predictive performance amongst all compared. Traditionally M_0 depict the null model, however we here seek to adhere with the symbology employed by JASP analyses. We interpreted the strength of evidence communicated by *BF*s based on the classification system suggested by Jefreys (1961) and Wetzels et al. (2011; see Table 1).

Table 1

Classification of Bayes factors, adapted from Wetzels et al., (2011).

BF_{01}	Evidence for H_0	BF_{01}	Evidence for H_1
≥ 100	Decisive	≤ 1/100	Decisive
30 – 100	Very Strong	1/30 – 1/100	Very Strong
10 – 30	Strong	1/10 – 1/30	Strong
3 – 10	Substantial	1/3 – 1/10	Substantial
1 -3	Anecdotal	1 – 1/3	Anecdotal

1	No evidence	1	No evidence

We also provide model-averaged analyses that assess the plausibility of each main effect and interaction individually, considered across the full set of models under comparison. This output is useful when the number of models under comparison is high. Model-averaged measures include the prior inclusion probability for each effect, the posterior probability of inclusion, and the Bayes Factor for inclusion of each effect (BF_{incl}). This last statistic quantifies the change in prior to posterior inclusion probability, or how effect plausibility changed after data assessment. Following the guidelines outlined in van der Bergh et al. (2020), analyses of effects followed the matched models approach which results in smaller prior probability differences between interaction effects and main effects.

Primary analyses examined empirical sensitivity scores for unaided and aided conditions, as measured by *d'* (Hautus et al., 2022); ideal *d'* scores for aided conditions as determined by participant's unaided sensitivity and an optimal *CC* automation-use strategy; and efficiency (η). To calculate empirical levels of *d'*, we first calculated hit rates (HR) and false alarm rates (FAR) using Hautus's (1995) transformation to correct for extreme values:

$$HR = \frac{.5+n\ hits}{n\ signal\ trials+1};\ FAR = \frac{(.5+n\ false\ alarms)}{(n\ noise\ trials+1)}.$$

For each participant, $d'_{unaided}$ and d'_{aided} were derived from the corrected hit and false alarm rates. Empirical *d'* scores for unaided conditions were transformed to optimal aided scores through the framework of the *CC* model. Specifically, we used participants estimation of $d'_{Unaided}$ to calculate what would have been optimal values of λ for positive and negative trials using the following formulas:

$$\lambda^*_{yes} = \frac{1}{2}d'_{Unaided} - \frac{logit(aid\ reliability)}{d'_{Unaided}};\ \lambda^*_{no} = \frac{1}{2}d' - \frac{logit(1-aid\ reliability)}{d'}.$$

Estimated λ^* measures were, in turn, employed to estimate optimal HR and FAR for each participant, which ultimately allowed for an estimation $d'_{Optimal}$ through the traditional formula for *d'*. Optimal scores represent best-attainable levels of sensitivity for automation-aided operators, conditional on operators' unaided sensitivity levels and the aid's reliability. Finally,

empirical, and ideal aided scores for each participant were used to calculate efficiency of automation-use (η, Tanner & Birdsall, 1958).

Estimated measures of performance were submitted to three default Bayesian ANOVAs in JASP:

1. A mixed factors 2 (Task Difficulty: Easy vs. Difficult) × 2 (Aid Reliability: Low vs. High) × 2 (Automation-Aid: Unaided vs. Aided) factorial analysis of empirical *d'* scores. This analysis aims to determine the conditions under which assistance from the automated aid produces the largest gains to sensitivity.
2. A mixed factors 2 (Task Difficulty: Easy vs. Difficult) × 2 (Aid Reliability: Low vs. High) × 2 (Score Type: Empirical vs. Ideal) analysis of *d'* scores for aided conditions. This analysis aims to identify the conditions under which empirical and optimal measures differ from one another.
3. A between-subjects 2 (Task Difficulty: Easy vs. Difficult) × 2 (Aid Reliability: Low vs. High) factorial analysis of efficiency. This analysis aims to identify the conditions under which the quality of automation use falls farthest from ideal levels.

Note that the second and third tests have similar purposes but will characterize the quality automation use in different ways, either in terms of raw *d'* (#2) or in terms of the squared ratio of empirical to ideal *d'* (#3). Analyses assumed uniform prior probabilities for all models, such that models compared are deemed equally probable prior to assessing the data.

Results

Descriptive statistics summaries are provided for various measures of performance (see appendices section). Mean and standard deviations for parameter of interest describe the empirical distribution of raw data. 95% Bayesian Credible Intervals (BCI) comprise the values between the 2.5% and 97.5% quantiles of posterior distributions of parameter values. Model comparison and analyses of effects summaries for all default Bayesian ANOVAs conducted on measures of performance are likewise provided. Model comparison summaries include only the top 5 best-performing models. Null models were not included amongst these for any of the analyses here conducted.

Raw Measures of Sensitivity (*d'*).

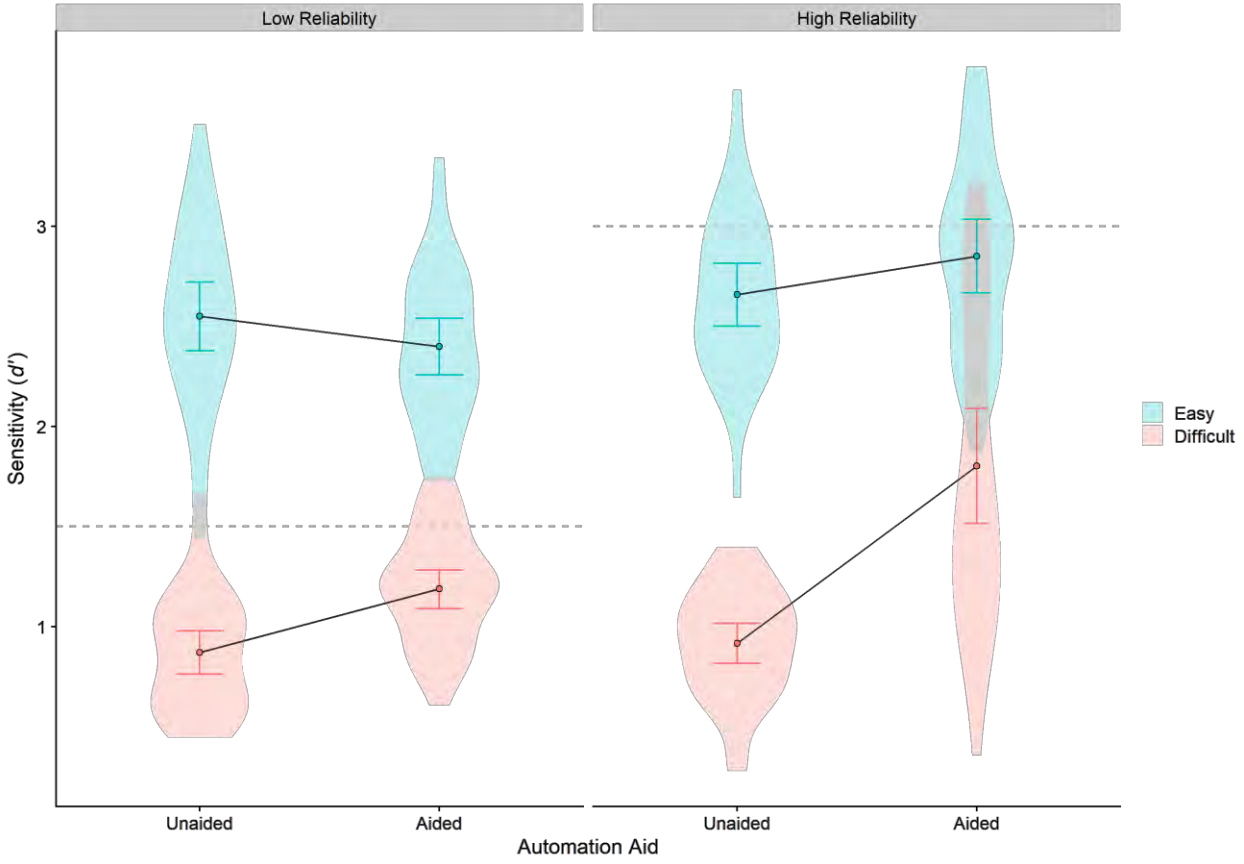

Figure 10. Measures of unaided and aided *d'* across levels of Aid Reliability and Task Difficulty. Violin plots visualize the distribution of empirical scores for each condition. Empirical mean sensitivity scores are represented by colored circles. Error bars represent 95% Bayesian Credible Intervals (BCIs) for posterior distributions. Dashed grey lines depict aid reliability levels.

Measures of *d'* across unaided and aided blocks for each condition are visualized in Figure 4 (see Appendix A for a summary of descriptive statistics). Violin plots depict the distributions of empirical scores. The dashed grey line in each panel represents *d'* scores for automated aids. Notice that mean levels of sensitivity in the difficult conditions were lower than aid sensitivity, even at low-reliability levels. Similarly, for easy conditions mean sensitivity levels failed to surpass the sensitivity of the highly reliable aid. These patterns suggest suboptimal automation-use. In this figure, the difference between mean *d'* for unaided and aided blocks is visualized through connecting solid black lines. To best visualize this aspect of the data, we plotted measures of aid effect directly in Figure 11. Aid effect was derived through the following formula: $d'_{Aided} - d'_{Unaided}$.

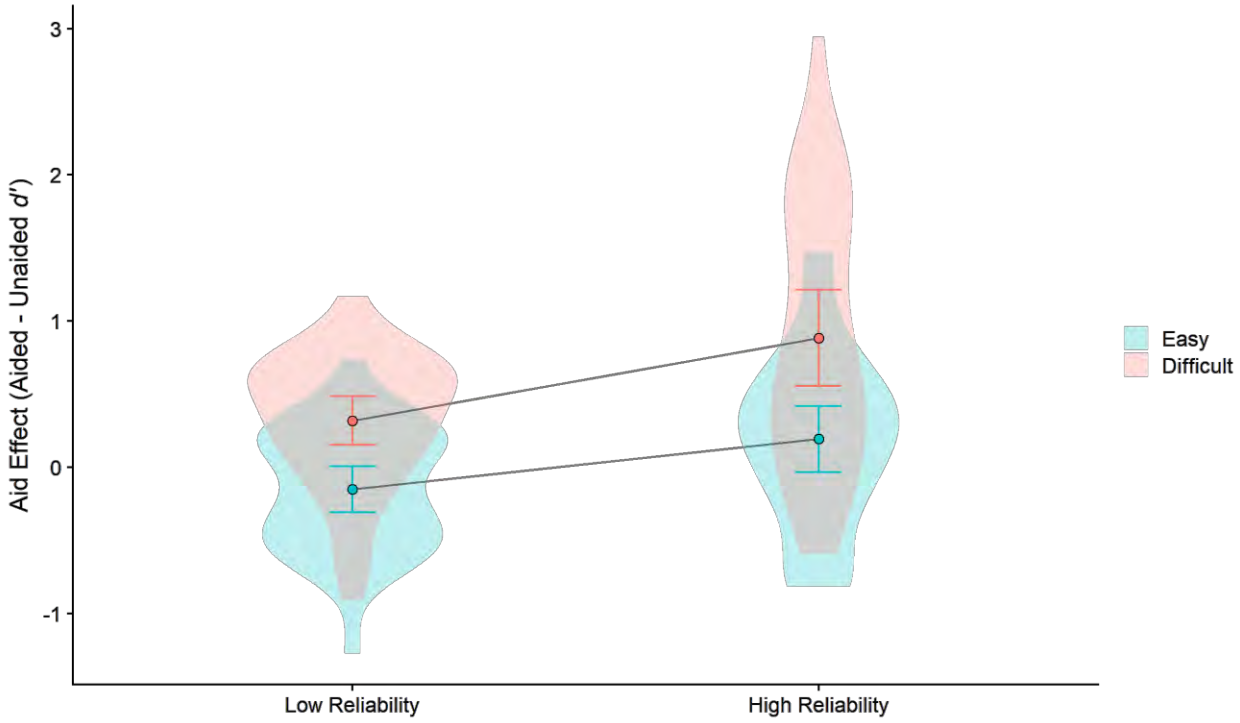

Figure 11. Aid-effect measures across levels of Task Difficulty and Aid Reliability. Violin plots depict distributions for empirical data. Empirical mean measures are depicted by color circles. Errorbars indicate 95% BCIs for posteriors distributions.

Table 2

JASP summary output of Bayesian ANOVA on d'.

	Summary output			
Model	**P(M)**	**P(M|data)**	**BF_{01}**	**Error %**
Aid + Diff + Rel + (Aid × Diff) + (Aid × Rel)	0.053	0.747	1.0	
Aid + Diff + Rel + (Aid × Diff) + (Aid × Rel) + (*Diff* × *Rel*)	0.053	0.182	4.1	5.365
Aid + Diff + Rel + (Aid × Diff) + (Aid × Rel) + (*Diff* × *Rel*) + (*Aid* × *Diff* × *Rel*)	0.053	0.070	10.67	4.674
Aid + Diff + Rel + (Aid × Diff)	0.053	0.001	747	4.521
Aid + Diff + Rel + (Aid × Diff) + (Diff x Rel)	0.053	2.446×10^{-4}	3053	4.283

Note. Effects absent from best model are italicized. Only the four best models out of 19 total models compared are presented here (Automation-Aid = Aid; Reliability = Rel.; Difficulty = Diff.).
P(M) = Model prior probability.
P(M|data) = Model posterior probability.
BF_{01} = Bayes Factor relative to the best-fitting model.

Sensitivity scores were submitted to a mixed factorial Bayesian ANOVA with Task Difficulty (Easy vs. Difficult) and Aid Reliability (Low vs. High) as between-subjects factors, and Automation-Aid (Unaided vs. Aided) as a within-subjects factor. Table 2 shows model

comparison summary output for the top five best performing models. The best-performing model included all three main effects along with two-way interactions for Automation-Aid × Reliability and Automation-Aid × Difficulty. These results are consistent with Hypothesis 1, which predicted higher aid-benefit at higher levels of Task Difficult and Aid Reliability. In addition, these results are consistent with discernible empirical patterns. Notice how the difference between Unaided and Aided sensitivity in Figure 10 appears steeper for the Difficult conditions in red and at High Reliability levels on the right panel. Congruently, in Figure 11, the parallel mean-connecting lines for Task-Difficulty levels across levels of Aid Reliability are suggestive of additive effects of these factors on measures of Aid-Benefit.

Model comparisons between the best-performing model and the second-best selected model, which included an additional interaction term between Difficulty and Reliability, indicated substantial evidence in favor of the best-performing model ($BF_{01} = 4.1$). The data also indicated strong evidence in favor of the best model compared to a model that includes all possible two-way and three-way interactions ($BF_{01} = 10.67$). All subsequent model comparisons, including that against a null model, indicated decisive evidence in favor of the best model. The uncertainty of the comparisons between the top-three best-performing models is clarified by model-averaged analyses of effects reported in Table 3. Data decisively favored the inclusion of all three main effects along with interaction terms between Automation-Aid and each between-subjects factors (all $BF_{Incl} \geq 677$), consistent with the best-fitting model. On the other hand, data indicated substantial evidence against an interaction between reliability and difficulty ($BF_{Incl} = \frac{1}{4.1}$) and anecdotal evidence against a three-way interaction ($BF_{Incl} = \frac{1}{2.62}$).

Table 3

Model-averaged analyses of effects for Bayesian ANOVA on d'.

	Model-Averaged Summaries		
Effect	**P(incl)**	**P(incl|data)**	BF_{incl}
Automation Aid	0.263	3.902×10^{-8}	49191.376
Reliability	0.263	0.001	4082.696
Difficulty	0.263	$9.712\times10\text{-}^{6}$	$8.249\times10^{+43}$
Automation Aid × Reliability	0.263	0.929	677.748
Automation Aid × Difficulty	0.263	0.930	78422.445
Reliability × Difficulty	0.263	0.183	1 / 4.1
Automation Aid × Reliability × Difficulty	.053	0.070	1 / 2.62

P(incl) = Prior probability of inclusion.
P(incl|data) = Posterior probability of inclusion.
BF_{incl} = Bayes Factor for probability of inclusion.

Optimal vs. Empirical Aided-Sensitivity.

Measures of empirical and optimal aided *d'* across levels of Difficulty and Reliability are visualized in Figure 12 (see Appendix B for a summary of descriptive statistics). Note that empirical values are replotted from Figure 10. Here we see that the distribution of scores for optimal measures in the Difficult conditions is narrow compared to other conditions, particularly in the Difficult/High reliability condition, where the gap between individual agent sensitivity was highest. This pattern results from the highly superior sensitivity of the automated aid relative to participants in these conditions. In effect, the aid dominated the human operator in the difficult conditions, meaning that optimal *d'* in the aided conditions was very similar to the aid's *d'*.

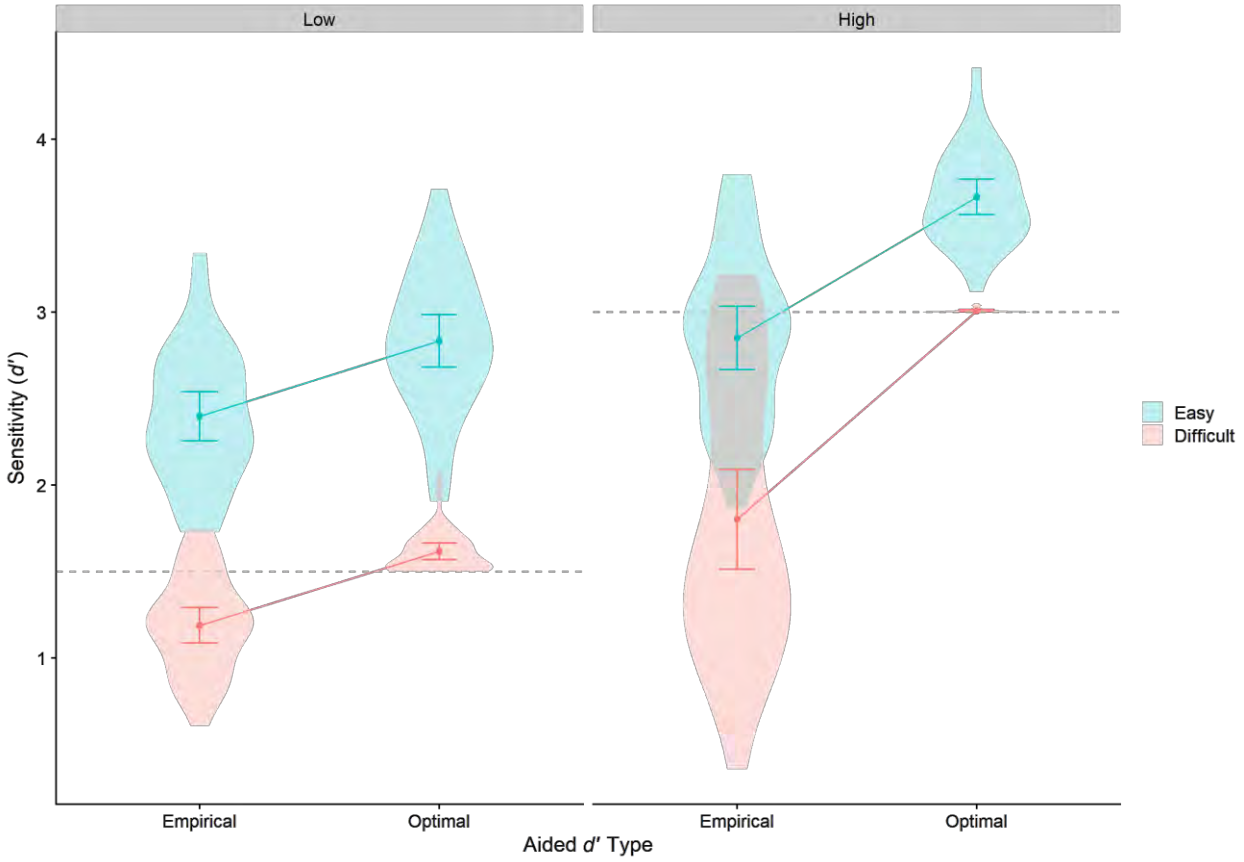

Figure 12. Measures of empirical and optimal aided *d'* across levels of Reliability and Difficulty. Violin plots visualize the distribution of empirical scores for each condition. Empirical mean sensitivity scores are represented by colored circles. Errorbars represent 95% BCIs for posterior distributions. Dashed lines depict automation *d'*.

Measures of empirical and optimal aided *d'* were submitted to a mixed Bayesian ANOVA with Automation Type (Empirical vs. Optimal) as a within-subjects factor, and Difficulty (Easy vs. Difficult) and Reliability (Low vs. High) as between-subjects factors. Table 4 shows model comparison summary output for all models considered. The best performing model included all possible effects considered including a three-way interaction for Type × Reliability × Difficulty. However, the data was highly undecisive between this model and the subsequent two best performing ones. The best-performing model was only anecdotally favored over the second best performing one which did not include a three-way interaction (BF_{01} = 1.63). Similarly, only anecdotal support was observed for the best model compared to the third best performing one which only included interaction effects for Type × Reliability, and for Reliability × Difficulty (BF_{01} = 1.67). Uncertainty in model comparisons is clarified by model-averaged analyses of effects reported in Table 5. Here we see decisive evidence for all three main effects and an

interaction for Type × Reliability (all $BF_{Incl} \geq 1000$), followed by strong evidence for an interaction for Reliability × Difficulty ($BF_{Incl} = 15.49$), but only anecdotal evidence for an interaction term for Type × Difficulty ($BF_{Incl} = 1.033$) and a three-way interaction ($BF_{Incl} = 1.63$).

Table 4

JASP summary output of Mixed Bayesian ANOVA on aided d'.

Model	Summary output			
	P(M)	P(M|data)	BF_{01}	Error %
Type + Rel + Diff + (Type × Rel) + (Type × Diff) + (Rel × Diff) + (Type × Rel × Diff)	0.053	0.437	1	
Type + Rel + Diff + (Type × Rel) + (Type × Diff) + (Rel × Diff)	0.053	0.268	1.63	7.486
Type + Rel + Diff + (Type × Rel) + (Rel × Diff)	0.053	0.261	1.67	10.498
Type + Rel + Diff + (Type × Rel) + (Type × Diff)	0.053	0.018	24.28	6.093
Type + Rel + Diff + (Type × Rel)	0.053	0.016	27.31	5.394

Note. Best five out of 19 total models compared are summarized here (Aided *d'* Type (optimal vs. empirical) = Type; Reliability = Rel.; Difficulty = Diff.).
P(M) = Model prior probability.
P(M|data) = Model probability given the data.
BF_{01} = Bayes Factor relative to best betting model.

Table 5

Summary output of model averaged analyses of effects aided d' measures.

Effect	Model-Averaged Summaries		
	P(incl)	P(incl|data)	BF_{incl}
Automation Aid	0.263	1.862×10^{-7}	$1.392\times10^{+28}$
Reliability	0.263	3.024×10^{-8}	$1.214\times10^{+24}$
Difficulty	0.263	0.016	$9.601\times10^{+32}$
Automation Aid × Reliability	0.263	0.563	$1.639\times10^{+6}$
Automation Aid × Difficulty	0.263	0.286	1.033
Reliability × Difficulty	0.263	0.529	15.487
Automation Aid × Reliability × Difficulty	.053	0.437	1.627

P(incl) = Prior probability of inclusion.
P(incl|data) = Posterior probability of inclusion.
BF_{incl} = Bayes Factor relative to the best-fitting model.

Results from analyses of effects are consistent with empirical data patterns visualized in Figure 12. Here we see optimal estimates of aided d' generally outperformed empirical values.

This pattern is consistent with Hypothesis 2, which predicted widespread suboptimal aid-use interaction. In addition, an interaction between Type and Reliability is apparent in Figure 12. Notice how the connecting lines visualizing aided *d'* difference are steeper at higher reliability levels in the right panel. This pattern suggests empirical aided *d'* fell shorter from optimal at higher reliability levels. However, the difference between these measures does not seem to be affected by difficulty levels, indicating similar suboptimality in aid-use across easy and difficult conditions. An interaction between Reliability and Difficulty is harder to discern. However, d'_{Aided} measures for difficult and easy conditions appear closer together at high reliability levels, perhaps due to aid dominance for d'_{Optimal} estimations in the Difficult/High reliability condition.

Aid-Use Efficiency.

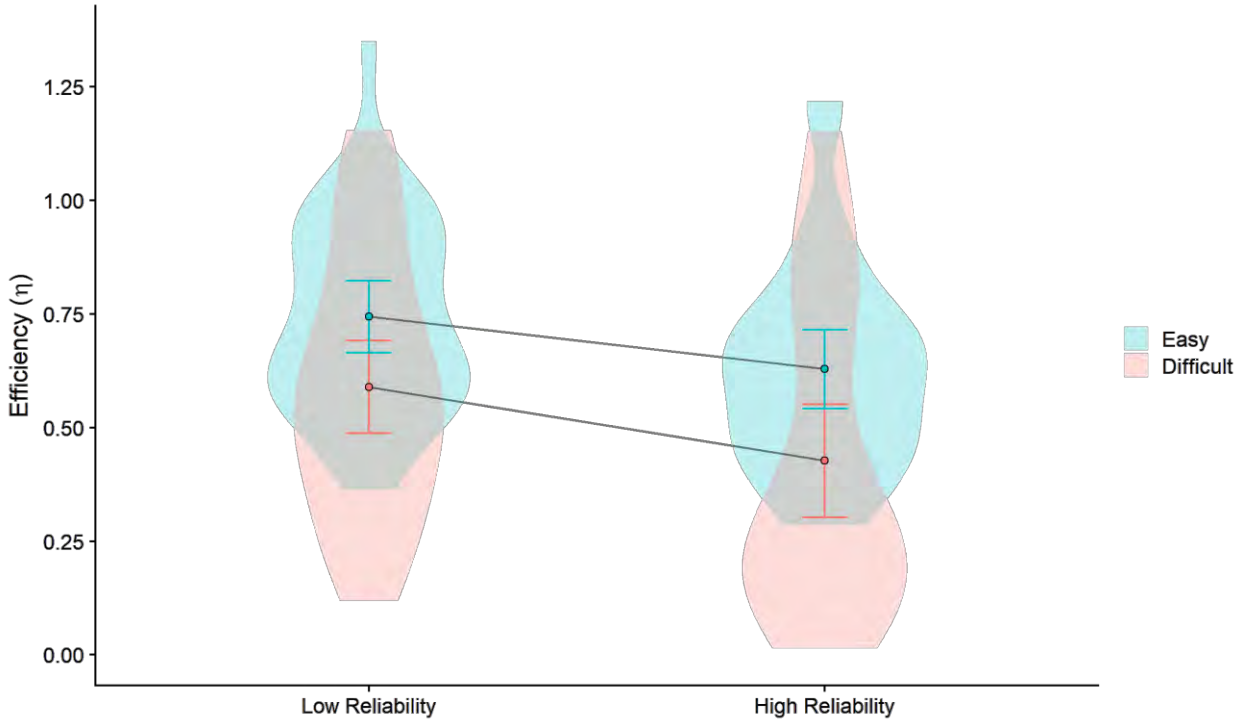

Figure 13. Measures of efficiency across levels of Reliability and Difficulty. Violin plots depict the distributions of empirical scores. Empirical mean scores are visualized by colored circles. Errorbars represent 95% BCIs for posterior distributions.

Measures of efficiency across levels of difficulty and reliability are visualized in figure 13 (see Appendix B for descriptive statistics summaries). As discussed, efficiency is calculated as the squared ratio of empirical optimal aided performance, such that measures of 1 are indicative

of best attainable performance. However, note that occasional efficiency scores higher than one are observed as a result of random variance within subjects.

Table 6

JASP summary output of Bayesian ANOVA on efficiency.

Model	Summary output			
	P(M)	P(M|data)	BF_{01}	Error %
Difficulty + Reliability	0.2	0.69	1	
Difficulty + Reliability + (Difficulty × Reliability)	0.2	0.19	3.63	2.08
Difficulty	0.2	0.1	6.9	0.7
Reliability	0.2	0.01	69	0.7
Null Model	0.2	0.002	345	0.7

Note. A summary for all models compared is included here.
P(M) = Model prior probability.
P(M|data) = Model probability given the data.
BF_{01} = Bayes Factor relative to best-fitting model.

Measures of efficiency were submitted to a Bayesian ANOVA (see Table 4) with Difficulty (Easy vs. Difficult) and Reliability (Low vs. High) as between-subjects factors. Model comparisons favored a model which included only main effects for both factors. This model is consistent with Hypothesis 3 which predicted lower aid-use efficiency at high levels of Reliability and Task Difficulty. Furthermore, this model is echoed in discernible patterns in empirical data visualized in Figure 13. The parallel lines mean-connecting lines for easy and difficult conditions across levels of reliability are suggestive of additive effects for these factors on aid-use efficiency. Efficiency was lower for the difficult conditions on red and at higher reliability levels on the right side of the x-axis. These patterns are reversed from measures of aid-effect and suggest participants fall shorter of potential aided performance in conditions where they exhibit higher aid-benefit. Interestingly, it is also in these conditions that participants could benefit most from automation.

Model comparisons indicated only substantial evidence in favor of the best-fitting model compared to the second best-fitting including an interaction effect between Difficulty and Reliability (BF_{01} = 3.63). In addition, the data only substantially favored the best-fitting model over the third best-fitting one which includes only a main effect for Difficulty (BF_{01} = 6.9). Analyses of effects in Table 7 offer some clarification to this uncertainty. Here, we see data

indicate very strong evidence for a main effect of Difficulty (BF_{incl} = 60.39), followed by substantial evidence for an effect of reliability (BF_{incl} = 6.9), but substantial evidence against an interaction term (BF_{incl} = 1 / 3.7).

Table 7

Model averaged analyses of effects for Bayesian ANOVA on efficiency.

	Model-Averaged Summaries		
Effect	**P(incl)**	**P(incl|data)**	***BF*$_{incl}$**
Reliability	0.400	0.709	6.955
Difficulty	0.400	0.798	60.386
Reliability × Difficulty	0.200	0.189	1 / 3.7

P(incl) = Prior probability of inclusion.
P(incl|data) = Posterior probability of inclusion.
BF_{incl} = Bayes Factor for inclusion.

Discussion

The present research examined the effects of Automation Aid (unaided vs. aided), Aid Reliability (low vs. high), and Task Difficulty (easy vs. difficult) on performance in a numeric signal detection task. In line with previous research, measures of sensitivity generally improved as a result of automation implementation. However, the extent of aid benefit increased at higher levels of Task Difficulty and Aid Reliability. Thus, participants benefitted most from the aid in conditions where their sensitivity was lower compared to that of automation. Interestingly, however, the sensitivity of the human + automation team did not always surpass that of automation. Empirical mean levels of sensitivity for the High Reliability conditions fell below automation sensitivity levels (d' = 3.0). Similarly, for the Low Reliability condition, mean sensitivity levels for participants completing a difficulty task failed outperformed automation sensitivity (d' = 1.5). These team shortcomings in raw performance are indicative of insufficient aid-use.

Participants empirical aided performance generally differed from optimal estimations, indicating widespread suboptimal automaton use, although this difference increased at high reliability levels. Interestingly, the difference between empirical and optimal measures was invariant across levels of task difficulty, even though optimal aided sensitivity was closer to unaided levels at easier difficulties. This pattern suggests participants could have benefitted from

higher aid-use even when completing a relatively easier task. Congruent with comparisons of empirical and optimal aided sensitivity, measures of aid-use efficiency (η) fell below 1.0 across conditions. In direct contrast with measures of aid-benefit, however, efficiency was lower at high levels of task difficulty and aid reliability. This inverse pattern suggests an ironic effect of these factors on efficiency where the human + automation team fell shorter of potential sensitivity gains in conditions where they showed most improvement relative to unaided performance. Interestingly, these conditions are also those in which operators could have benefited most from automation assistance.

The present research replicates and extends the ironic effects of aid reliability on aid-benefit and aid-use efficiency previously observed with graded automation (Bartlett & McCarley, 2021). In addition, the observed effects of difficulty on both these measures help clarify data uncertainty in previous research (Tikhomirov, *in press*). No decisive evidence was observed for interaction effects between task difficulty and reliability on measures of performance, rather, aid benefit and efficiency measures were best predicted by additive effects of these factors.

Implications for Human-Automation Interaction.

The shortcomings in aid-use efficiency and aided performance observed in this research indicate suboptimal human-automation interaction. Under a *CC* strategy framework, these patterns can be explained by suboptimal calibration of dependence across conditions. Recall that an optimal *CC* strategy requires criterion shift to calibrate with cue-contingent ideal values of λ^*. For instance, consider the formula for ideal criterions for positive cue trials: $\lambda^*_{yes} = \frac{1}{2}d' - \frac{logit(aid\ reliability)}{d'}$. Applying this simplified formula to sensitivity levels in the unaided Easy/Low reliability condition ($M\ d'_{\text{Unaided}} = 2.55$; aid $d' = 1.5$) yields a λ^*_{yes} of .84, while at higher difficulty in the Difficult/Low reliability condition ($M\ d'_{\text{Unaided}} = .871$; aid $d' = 1.5$) $\lambda^*_{yes} = -.85$. Finally, at higher reliability levels in the Difficult/High reliability condition ($M\ d'_{\text{Unaided}} = .917$; aid $d' = 3$) λ^*_{yes} is smallest at -2.36. Here we see values of λ^*_{yes} decrease as reliability and difficult increase, indicating higher measures of $\Delta\lambda$ would be necessary to achieve optimal performance. However, consistent with prior literature (e.g., Munoz Gomez Andrade et al., 2022; Robinson & Sorkin, 1985), empirical measures of aided sensitivity were more

compatible with a sluggish *CC* strategy in which operators fail to sufficiently increase dependence as automation sensitivity increases relative to their own.

Although insufficient dependence under a *CC* framework provides a potential explanation for the observed shortcomings in aided performance, suboptimal human-automation interaction may also be explained by recruitment of alternative suboptimal aid-use strategies. Previous research proposed a Discrete Deference (*DD*) model of human-automation interaction in which, rather shifting their λ contingent on an aid's cue, participants simply differ to the aid's judgment in a subset of trials (Tikhomirov et al., *in press*). Although this strategy might alleviate cognitive load, it also results in lower highest potential values of d'_{Aided} due to an overall loss of system information. Previous research comparing the fit of the *DD* and *CC* models, as well as a Mixture model of both strategies suggested least predictive performance for a *CC* strategy (Duncan-Reid & McCarley, 2022; Munoz & McCarley, 2023; Tikhomirov et al., *in press*). However, although suboptimality in aid-use efficiency is necessarily indicative of suboptimal human-automation interaction, analyses of mean sensitivity values are unable to discern between the plausibility of different strategies to account for the data. Future research may employ cognitive modeling (see Lee & Wagenmakers, 2014) to better understand how aid-use strategies and dependence might vary across levels of reliability and task difficulty.

Applied Implications.

Various aspects of the present findings speak to the design of automated decision aids. For instance, although aid-benefit was highest for the difficult condition, this effect was present even at low levels of task difficulty. This pattern suggests automation implementation might be valuable whenever operator sensitivity is limited by inherent ambiguity in sampled evidence (see Metz & Shen, 1992). Second, even in circumstances where a large aid-benefit is observed, as in the difficulty conditions, performance might still fall short of optimal levels. These findings are consistent with a growing body of literature documenting widespread suboptimality in aid-use (Chi & Drury, 1998; Duncan-Reid & McCarley, 2021, 2022; Neyedli et al., 2011; etc), and suggest the expected value and cost-efficiency of automation implementation might be undercut in these conditions.

Generality Constraints and Future Directions.

The findings of the present research should be interpreted with caution when generalizing across populations and task settings. Here we recruited novice subjects who may behave different than field experts or subjects with more intensive training (see Arkes et al., 1986; Berner and Graber, 2008; Liang et al., 2022). Future research might recruit professional or trained populations to best inform system-design (e.g., Araki et al., 2010; Novotny et al., 2017; Rovira & Parasuraman, 2007). In addition, the present research employed a very specific form of signal-detection task, which might not be fully representative of the tasks performed in many applied settings. To examine the generalizability of the present findings, researchers might employ paradigms that more closely resemble applied tasks (e.g., Hutchinson et al., 2022; McCarley, 2009). Finally, participants completed the present task remotely from desktop devices, which, again, may not best-represent the nature of naturalistic tasks. Environments closer to naturalistic settings (e.g., Morphew & Wickens, 1998) might better resemble arousal and motivation levels in applied settings and provide valuable input for design.

As discussed, suboptimal aid-use efficiency is symptomatic of poor human-automation interaction. Future research might aim to uncover interventions that promote more efficient automation-use through the adoption of more appropriate aid-use strategies and dependence. For instance, providing precise information regarding aid reliability (see Avril, 2023; Wang et al., 2009) and extended training with decision aids (see Liang et al., 2022) might increase automation reliance. However, to improve signal-detection performance, future research might also compare aid-benefits produced by other kinds of automation, including those which provide graded cues (e.g., Bartlett & McCarley, 2021); provide assistance at different stages of information processing (e.g., Liechty, 2019); and function at higher levels of automation autonomy. Previous research suggests automation and a reliance in actuarial information can match and, sometimes, outperform the ability of experts to gauge diagnostic information and efficiently weigh multiple information sources (see Ægisdóttir et al., 2006; Dawes et al., 1989; Metz & Shen, 1992; Yaniv, 2004).

Conclusion.

Automation implementation aims to increase signal-detection performance to reduce human error. Although this intended effect was empirically observed in the present research, participants ironically fell shorter of optimal performance under conditions at which they showed highest aid-benefit and in which they could benefit most from aid assistance. This pattern highlights the need to better understand human-automation interaction across task settings to increase the value and cost-efficiency of automation implementation.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Araki, K., Matsuda, Y., Seki, K., & Okano, T. (2010). Effect of computer assistance on observer performance of approximal caries diagnosis using intraoral digital radiography. *Clinical Oral Investigations*, *14*(3), 319–325. https://doi.org/10.1007/s00784-009-0307-z

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110. https://doi.org/10.1016/0749-5978(86)90046-4

Avril, E. (2023). Providing different levels of accuracy about the reliability of automation to a human operator: Impact on human performance. *Ergonomics*, *66*(2), 217–226. https://doi.org/10.1080/00140139.2022.2069870

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(6), 881–900. https://doi.org/10.1177/0018720817700258

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. https://doi.org/10.1504/IJHFE.2018.092227

Bartlett, M. L., & McCarley, J. S. (2019). No Effect of Cue Format on Automation Dependence in an Aided Signal Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(2), 169–190. https://doi.org/10.1177/0018720818802961

Bartlett, M. L., & McCarley, J. S. (2021). Ironic efficiency in automation-aided signal detection. *Ergonomics*, *64*(1), 103–112. https://doi.org/10.1080/00140139.2020.1809716

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*(1), 169–184. https://doi.org/10.2308/aud.2000.19.1.169

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5), S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Bond, R. R., Novotny, T., Andrsova, I., Koc, L., Sisakova, M., Finlay, D., Guldenring, D., McLaughlin, J., Peace, A., McGilligan, V., Leslie, S. J., Wang, H., & Malik, M. (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology*, *51*(6), S6–S11. https://doi.org/10.1016/j.jelectrocard.2018.08.007

Boskemper, M. M., Bartlett, M. L., & McCarley, J. S. (2022). Measuring the Efficiency of Automation-Aided Performance in a Simulated Baggage Screening Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *64*(6), 945–961. https://doi.org/10.1177/0018720820983632

Botzer, A., Meyer, J., & Parmet, Y. (2013). Mental effort in binary categorization aided by binary cues. *Journal of Experimental Psychology: Applied*, *19*(1), 39–54. https://doi.org/10.1037/a0031625

Botzer, A., Meyer, J., Borowsky, A., Gdalyahu, I., & Shalom, Y. B. (2015). Effects of cues on target search behavior. *Journal of Experimental Psychology: Applied*, *21*(1), 73–88. https://doi.org/10.1037/xap0000035

Carroll, A. E., Bauer, N. S., Dugan, T. M., Anand, V., Saha, C., & Downs, S. M. (2013). Use of a Computerized Decision Aid for ADHD Diagnosis: A Randomized Controlled Trial. *Pediatrics*, *132*(3), e623–e629. https://doi.org/10.1542/peds.2013-0933

Chi, C.-F., & Drury, C. G. (1998). Do people choose an optimal response criterion in an inspection task? *IIE Transactions*, *30*(3), 257–266. https://doi.org/10.1080/07408179808966456

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

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: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

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: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 564–572. https://doi.org/10.1518/001872007X215656

Duncan-Reid, J. (2022). *Automation-Aided Collaborative Strategies in Signal Detection Tasks* [Unpublished doctoral dissertation]. Oregon State University.

Duncan-Reid, J., & McCarley, J. S. (2021). Strategy Use in Automation-Aided Decision Making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 96–100. https://doi.org/10.1177/1071181321651259

Duncan-Reid, J., & McCarley, J. S. (2022). *Automation-Aided Collaborative Strategies in Signal Detection Tasks*.

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*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Elvers, G. C., & Elrif, P. (1997). The Effects of Correlation and Response Bias in Alerted Monitor Displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(4), 570–580. https://doi.org/10.1518/001872097778667960

Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, *50*(3), 595–601. https://doi.org/10.1016/j.dss.2010.08.010

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* John Wiley.

Gyles, S. P., & McCarley, J. S. (2019). Metacognition, numeracy, and automation-aided decision-making [Unpublished masters dissertation]. Oregon State University

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection Theory: A User's Guide* (3rd ed.). Routledge. https://doi.org/10.4324/9781003203636

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354. https://doi.org/10.1037/0278-7393.7.5.344

Huegli, D., Merks, S., & Schwaninger, A. (2020). Automation reliability, human–machine system performance, and operator compliance: A study with airport security screeners supported by automated explosives detection systems for cabin baggage screening. *Applied Ergonomics*, *86*, 103094. https://doi.org/10.1016/j.apergo.2020.103094

Hutchinson, J., Strickland, L., Farrell, S., & Loft, S. (2022). The Perception of Automation Reliability and Acceptance of Automated Advice. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082110629. https://doi.org/10.1177/00187208211062985

JASP Team (2019). JASP (Version 0.9.2)[Computer software]. https://jasp-stats.org/.

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Kunar, M. A. (2022). The optimal use of computer aided detection to find low prevalence cancers. *Cognitive Research: Principles and Implications*, *7*(1), 13. https://doi.org/10.1186/s41235-022-00361-1

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lee, S. J., Mo, K., & Seong, P. H. (2007). Development of an Integrated Decision Support System to Aid the Cognitive Activities of Operators in Main Control Rooms of Nuclear Power Plants.

2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, 146–152. https://doi.org/10.1109/MCDM.2007.369429

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive Research: Principles and Implications*, *7*(1), 14. https://doi.org/10.1186/s41235-022-00364-y

Liechty, Molly M. (2019). The Effect of Differing Degrees of Automation and Reliability on Simulated Luggage Screening Performance [Doctoral dissertation]. Old Dominion University

Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian Reanalyses From Summary Statistics: A Guide for Academic Consumers. *Advances in Methods and Practices in Psychological Science*, *1*(3), 367–374. https://doi.org/10.1177/2515245918779348

Maltz, M., & Shinar, D. (2004). Imperfect In-Vehicle Collision Avoidance Warning Systems Can Aid Drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 357–366. https://doi.org/10.1518/hfes.46.2.357.37348

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the Effect of Workload on Automation Use for Younger and Older Adults. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(6), 672–686. https://doi.org/10.1177/0018720811421909

McCarley, J. S. (2009). Response Criterion Placement Modulates the Benefits of Graded Alerting Systems in a Simulated Baggage Screening Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(17), 1106–1110. https://doi.org/10.1177/154193120905301713

Metz, C. E., & Shen, J.-H. (1992). Gains in Accuracy from Replicated Readings of Diagnostic Images: Prediction and Assessment in Terms of ROC Analysis. *Medical Decision Making*, *12*(1), 60–75. https://doi.org/10.1177/0272989X9201200110

Meyer, J. (2001). Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 563–572. https://doi.org/10.1518/001872001775870395

Morphew, M. E., & Wickens, C. D. (1998). Pilot Performance and Workload Using Traffic Displays to Support Free Flight. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(1), 52–56. https://doi.org/10.1177/154193129804200113

Munoz Gomez Andrade, F., Duncan-Reid, J., & McCarley, J. S. (2022). The Effect of Correlated Observations on Human-Automation Interaction in a Signal Detection Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 2062–2066. https://doi.org/10.1177/1071181322661290

Munoz Gomez Andrade, F., & McCarley, J.S. (2023). The Effect of Higher Information Redundancy in Aided Signal Detection. Manuscript in preparation [Manuscript in preparation]. School of Psychological Science, Oregon State University.

Murrell, G. A. (1977). Combination of evidence in a probabilistic visual search and detection task. *Organizational Behavior and Human Performance*, *18*(1), 3–18. https://doi.org/10.1016/0030-5073(77)90015-0

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond Identity: Incorporating System Reliability Information Into an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 338–355. https://doi.org/10.1177/0018720811413767

Novotny, T., Bond, R., Andrsova, I., Koc, L., Sisakova, M., Finlay, D., Guldenring, D., Spinar, J., & Malik, M. (2017). The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. *International Journal of Medical Informatics*, *101*, 85–92. https://doi.org/10.1016/j.ijmedinf.2017.02.007

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics*, *43*(7), 931–951. https://doi.org/10.1080/001401300409125

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

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 - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

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*(4), 320–331. https://doi.org/10.1037/a0024243

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, Netherlands: North-Holland.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–321. https://doi.org/10.1037/met0000057

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*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rovira, & Parasuraman. (2007). *Effects of Imperfect Automation Support on Air Traffic Controller (ATCO) Performance, Mental Workload, and Attention Allocation: Miss vs. False Alarm Prone Automation.*

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 76–87. https://doi.org/10.1518/001872007779598082

Sorkin, R. D., & Dai, H. (1994). Signal Detection Analysis of the Ideal Group. *Organizational Behavior and Human Decision Processes*, *60*(1), 1–13. https://doi.org/10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. https://doi.org/10.1037/0033-295X.108.1.183

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. Human Factors: The Journal of the Human Factors and Ergonomics Society, 30(4), 445–459. https://doi.org/10.1177/001872088803000406

Tanner, W. P., & Birdsall, T. G. (1958). Definitions of *d ′* and η as Psychophysical Measures. *The Journal of the Acoustical Society of America*, *30*(10), 922–928. https://doi.org/10.1121/1.1909408

Tikhomirov, L., Bartlett, M. L., Dr, Duncan-Reid, J., & McCarley, J. S. (*in press*). When the Going Gets Tough: The Efficiency of Automation-Aided Signal Detection Declines with Task Difficulty. https://doi.org/10.31219/osf.io/tu7qpti

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharský, Š., & Gupta, A. R. K. N. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique*, *120*, 73.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and Reliance on an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(3), 281–291. https://doi.org/10.1177/0018720809338842

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*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Wickens, C., & Colcombe, A. (2007). Dual-Task Performance Consequences of Imperfect Alerting Associated With a Cockpit Display of Traffic Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 839–850. https://doi.org/10.1518/001872007X230217

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments: Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(7), 1209–1221. https://doi.org/10.1177/0018720814528534

Wiegmann, D., 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*(8), 825–831.

Yamani, Y., & McCarley, J. S. (2018). Effects of Task Difficulty and Display Format on Automation Usage Strategy: A Workload Capacity Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(4), 527–537. https://doi.org/10.1177/0018720818759356

Yaniv, I. (2004). The Benefit of Additional Opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yeh, M., & Wickens, C. D. (2001). Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(3), 355–365. https://doi.org/10.1518/001872001775898269

Zhang, Y., Antonsson, E. K., & Grote, K. (2006). A new threat assessment measure for collision avoidance systems. *2006 IEEE Intelligent Transportation Systems Conference*, 968–975. https://doi.org/10.1109/ITSC.2006.1706870

Appendices

Appendix A

Table

Descriptive statistics for measures of d'.

				95% BCI					95% BCI	
Difficulty	***M***	***SD***	***n***	***Low***	***High***	***M***	***SD***	***n***	***Low***	***High***
					Raw d'					
		Unaided					*Aided*			
High Reliability										
Difficult Task	0.917	0.277	32	0.817	1.017	1.802	0.795	32	1.515	2.089
Easy Task	2.658	0.417	30	2.502	2.814	2.851	0.492	30	2.668	3.035
Low Reliability										
Difficult Task	0.871	0.305	33	0.763	0.979	1.190	0.288	33	1.088	1.292
Easy Task	2.550	0.476	32	2.379	2.722	2.398	0.393	32	2.257	2.540
					Aided d'					
		Empirical					*Optimal*			
High Reliability										
Difficult Task	1.802	0.795	32	1.515	2.722	3.009	0.013	32	3.004	3.014
Easy Task	2.851	0.492	30	2.668	2.814	3.667	0.275	30	3.564	3.769
Low Reliability										
Difficult Task	1.190	0.288	33	1.088	0.979	1.618	0.134	33	1.57	1.665
Easy Task	2.398	0.393	32	2.257	1.017	2.834	0.419	32	2.683	2.985

M = mean.
SD = standard deviation.
BCI = Bayesian Credible Interval.

Appendix B

Table

Descriptive statistics for measures of aid-effect and aid-use efficiency.

				95% BCI	
Condition	***M***	***SD***	***n***	***Low***	***High***
			Aid Effect		
High Reliability					
Difficult Task	0.885	0.913	32	0.555	1.214
Easy Task	0.193	0.607	30	-0.033	0.420
Low Reliability					
Difficult Task	0.318	0.467	33	0.153	0.484
Easy Task	-0.152	0.434	32	-0.308	0.005
			Efficiency		
High Reliability					
Difficult Task	0.427	0.346	32	0.302	0.552
Easy Task	0.629	0.233	30	0.542	0.716
Low Reliability					
Difficult Task	0.590	0.287	33	0.488	0.692
Easy Task	0.744	0.219	32	0.665	0.823

M = mean.
SD = standard deviation.
BCI = Bayesian Credible Interval.