Not all information is equal: Effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming

Na Du
Industrial and Operations Engineering, University of Michigan, Ann Arbor

Kevin Y. Huang
Industrial and Operations Engineering, University of Michigan, Ann Arbor

X. Jessie Yang
Industrial and Operations Engineering, University of Michigan, Ann Arbor

***Human Factors**, Accepted for Publication on June 14, 2019*

Manuscript type: *Research Article*

Running head: *Disclosing likelihood information*

Corresponding author: X. Jessie Yang, 1205 Beal Avenue, Ann Arbor, MI48109, Email: xijyang@umich.edu

Acknowledgment: We would like to thank Kevin Li and Benjamin Pinzone for programming the simulator.

Précis: We examined the effects of disclosing likelihood information on trust, compliance and reliance, and task performance. Results indicate that operators informed of the predictive values or the overall likelihood value, rather than the hit and correct rejection rates, relied on the decision aid more appropriately and obtained higher task scores.

Topic: Human-Computer Interaction, Computer Systems

Abstract

Objective: The study examines the effects of disclosing different types of likelihood information on human operators' trust in automation, their compliance and reliance behaviors, and the human-automation team performance.

Background: To facilitate appropriate trust in and dependence on automation, explicitly conveying the likelihood of automation success has been proposed as one solution. Empirical studies have been conducted to investigate the potential benefits of disclosing likelihood information in the form of automation reliability, (un)certainty, and confidence. Yet, results from these studies are rather mixed.

Method: We conducted a human-in-the-loop experiment with 60 participants using a simulated surveillance task. Each participant performed a compensatory tracking task and a threat detection task with the help of an imperfect automated threat detector. Three types of likelihood information were presented: overall likelihood information, predictive values, and hit and correct rejection rates. Participants' trust in automation, compliance and reliance behaviors, and task performance were measured.

Results: Human operators informed of the predictive values or the overall likelihood value, rather than the hit and correct rejection rates, relied on the decision aid more appropriately and obtained higher task scores.

Conclusion: Not all likelihood information is equal in aiding human-automation team performance. Directly presenting the hit and correct rejection rates of an automated decision aid should be avoided.

Application: The findings can be applied to the design of automated decision aids.

Keywords: Human-robot interaction, trust in automation, likelihood alerts, Bayesian inference, base rate fallacy

1. INTRODUCTION

Automated decision aids have been used in a wide array of domains, including military operations, medical diagnosis, transportation safety administration (TSA) among others. As automation becomes more capable in perception, planning, learning and action execution, it is expected to significantly enhance the human-automation team performance. However, issues arise when human agents place unjustified trust in and dependence on automation or when they do not display enough trust and dependence (Dixon, Wickens, & McCarley, 2007; Du et al., 2019; Lee & See, 2004; Parasuraman & Riley, 1997; Petersen, Robert, Yang, & Tilbury, 2019; Yang, Unhelkar, Li, & Shah, 2017).

To facilitate appropriate trust in and dependence on automation, explicitly conveying the likelihood of automation success has been proposed as one solution. Empirical studies have investigated the potential benefits of disclosing likelihood information in the form of automation reliability, (un)certainty, and confidence. Among existing studies, few were based upon specific computational algorithms, for instance, the neural network used in a study by McGuirl and Sarter (2006). Not surprisingly, to model the performance of the automation, the majority of existing studies employed the signal detection theory (SDT) (Macmillan & Creelman, 2005; Tanner & Swets, 1954), based on which the likelihood information is calculated. Yet, results from these studies seem to be inconsistent. Some studies revealed that the likelihood information significantly helped human operators calibrate their trust, adjust their reliance and compliance behaviors, and enhance human-automation team performance (McGuirl & Sarter, 2006; Walliser, de Visser, & Shaw, 2016; Wang, Jamieson, & Hollands, 2009). Other studies, however, reported that human operators did not trust or depend on automated decision aids appropriately even when the likelihood information was disclosed (Bagheri & Jamieson, 2004; Fletcher, Bartlett, Cockshell, & McCarley, 2017). A close examination of existing literature suggests that studies employ different methods to calculate the likelihood information, which potentially contribute to the mixed results.

SDT models the relationship between signals and noise, as well as the automation's ability to detect signals among noise. The state of the world is characterized by either "signal present" or "signal absent", which may or may not be identified correctly by the automation. The combination of the state of the world and the automation's detection results in four possible states: hit, miss, false alarm (FA) and correct rejection (CR).

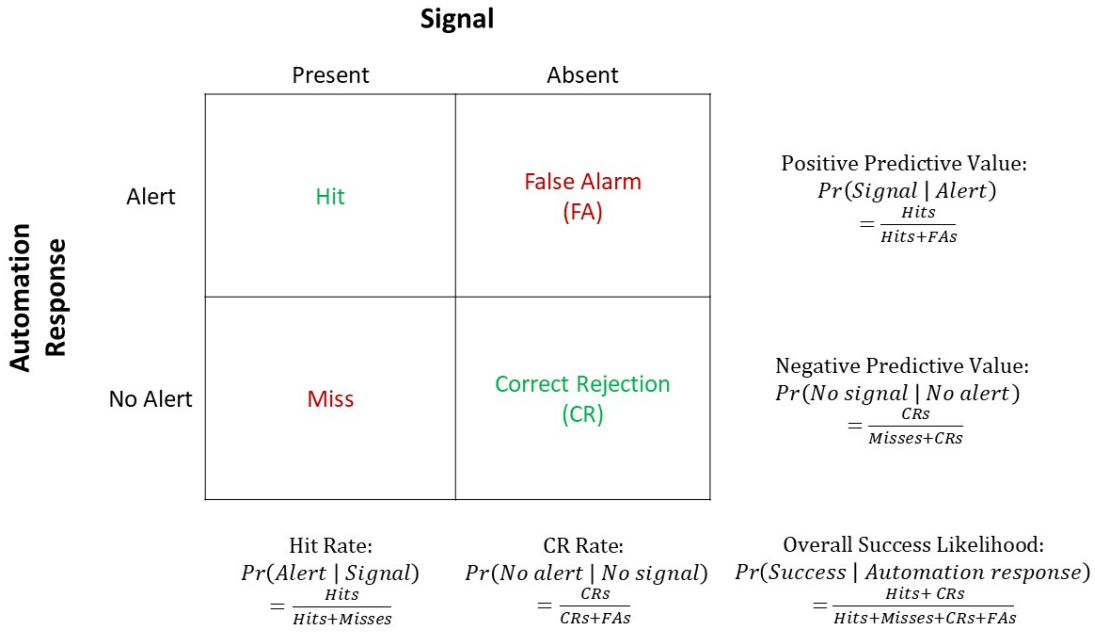


Figure 1. Signal Detection Theory (SDT) and Calculations of Hit Rate, CR Rate, Positive Predictive Value, Negative Predictive Value and Overall Success Likelihood

Based on the framework of SDT, the calculation of automation likelihood information can be broadly classified into three categories. The first category of likelihood information is the automation's overall likelihood of success regardless of hits or CRs, calculated as $Pr(Success \mid Response) = \frac{Hits\ +\ CRs}{Hits\ +\ Misses\ +\ FAs\ +\ CRs}$. For example, Dzindolet, Pierce, Beck, and Dawe (2002) examined how revealing the number of errors an automated decision aid made affected the perceived performance of and reliance on the automated aid. In their study, participants viewed 200 slides of displaying pictures of a military terrain and indicated whether or not a solider in camouflage was in the slide with the help from either an automated decision aid or a human decision aid. After

200 trials, half of the participants were provided with the reliability of the decision aid (total number of errors) and the other half not. The participants then rated the decision aid's performance and indicated whether to rely on the aid for the target detection task in 10 trials randomly chosen from the past 200 trials. Results showed that both types of decision aids were rated more favorably when its reliability was disclosed. More recently, Walliser et al. (2016) conducted a study where participants interacted with four unmanned aerial vehicles (UAVs) that utilized automated target recognition (ATR) systems to identify targets as enemy or friendly. Results showed that when participants were informed of the overall success likelihood information ("corrected identification rate" in the article), participants tended to apply a more appropriate strategy when interacting with the automation, resulting in better task performance.

The second type of likelihood information is the predictive value, calculated as $\frac{Hits}{Hits\ +\ FAs}$ or $\frac{CRs}{Misses\ +\ CRs}$. The positive predictive value means the probability of having a true signal given an automation alert, $Pr(Signal \mid Alert)$, and the negative predictive value the probability of not having a signal when the automation is silent, $Pr(No\ signal \mid No\ alert)$. Along this line of research, Wang et al. (2009) examined the effects of presenting the positive predictive value on human operators' belief, trust and dependence using a combat identification (CID) task. In the study, participants distinguished friend from foe with the aid from an imperfect CID. More specifically, due to its working mechanism, once the CID identified a soldier as friendly, it was always correct. However, when the CID identified a soldier as "unknown", the solider could be "friendly", "hostile" or "neutral". Half of the participants were informed of the positive predictive value and the other half not. Results of their study revealed that disclosing the positive predictive value to users positively influenced trust and reliance. In a follow up study, Neyedli, Hollands, and Jamieson (2011) developed four visual displays for presenting predictive values in the CID task. Display type (pie, random mesh) and display proximity (integrated, separated) of likelihood information were manipulated in the experiment. The results revealed that participants relied on the automation more appropriately and had greater sensitivity with the integrated display and random mesh

display. Studies on likelihood alarms also shed light on the effects of disclosing the predictive values. Likelihood alarms, in contrary to traditional binary alarms, integrate both state information and likelihood information by dividing a state into two or more graded levels. For instance, "warning" and "caution" could both indicate the presence of a target, with "warning" indicating a higher probability. Although not explicitly stated, these studies manipulated the positive and negative predictive values to represent the varying likelihood of true positives and true negatives given the automation responses, showing that in general human operators demonstrated higher trust in and dependence on alerts with higher likelihood (Sorkin, Kantowitz, & Kantowitz, 1988; Wiczorek & Manzey, 2014; Yang et al., 2017). Despite the above-mentioned positive evidence, Fletcher et al. (2017) asked participants to view a series of simulated sonar returns and decide whether a target was present or not. However, the display of rings indicating the likelihood that a target was present, given a return signal, did not seem to improve the overall ability of participants to distinguish targets from noise.

The third type of likelihood information is the hit rate and correct rejection rate, calculated as $\frac{Hits}{Hits\ +\ Misses}$ and $\frac{CRs}{FAs\ +\ CRs}$. Hit rate is the probability of automation issuing an alarm or an alert given a true signal, $Pr(Alert \mid Signal)$, and CR rate is the probability of automation silence when there is no signal, $Pr(No\ alert \mid No\ signal)$. It is important to distinguish the predictive values from the hit/CR rates. In fact, the positive/negative predictive values and the hit/CR rates are ***inverse*** conditional probabilities of each other. The two predictive values can be derived from the hit/CR rates using the Bayes Theorem (Please see the Present Study section for details). Utilizing the Multi-Attribute Task Battery (MAT; Comstock and Arnegard 1992), Bagheri and Jamieson (2004) examined the effect of providing operators with information about the context-related nature of automation reliability. Participants performed three tasks simultaneously: tracking, fuel management, and system monitoring. The monitoring task was automated and a gauge showing abnormal numbers would automatically reset its value. However, sometimes the automation would fail (miss) to correct the value and the human operator should intervene.

Automation reliability, essentially hit rate, ("Slightly under 100%" for high hit rate or "Slightly above 50%" for low hit rate) was disclosed to the participants. Comparing to a previous study where participants were unaware of the likelihood information, there seemed to be no evidence on any beneficial effects of disclosing hit rate on trust in automation or task performance.

2. THE PRESENT STUDY

The above-mentioned studies on likelihood information suggest that disclosing the overall likelihood information could increase preference and task performance. In addition, in general there is positive evidence supporting that presenting predictive values could help human operators calibrate their trust and adjust their dependence behaviors, leading to better performance. In contrast, revealing hit/CR rates does not seem to be beneficial. Despite the inconsistent results, there is little, if not no, research directly comparing the effects of revealing different types of likelihood information.

In the present study, we aimed to investigate if and how different methods of calculating likelihood information affect operators' trust in and dependence on automation, and task performance. We argue that the beneficial effects of disclosing the likelihood information are influenced by, at least, two factors. The first factor is information granularity – the extent to which the likelihood information represents probabilistic information specific to certain conditions. The overall success likelihood, $Pr(Success \mid Automation\ response)$, is less fine-grained compared to the predictive values and the hit/CR rates, as it represents an aggregated probability regardless what the automation response is (alert or no alert). The second factor is information directness – the extent to which the likelihood information can be directly used to guide people's behaviors without the need to estimate or integrate other information. The predictive values are the most direct in guiding people's compliance and reliance behaviors. The positive predictive value, $Pr(Signal \mid Alert) = x\%$, indicate that when the automation's alert or alarm goes off, there is $x\%$ chance that there is a true signal. Probabilistically speaking, if the automation's alarm goes off 100 times, there would be

x true alarms and $100 - x$ false alarms. And an optimal decision maker should only check the x number of true alarms and save his or her time and resources when false alarms happen. The same logic applies to the negative predictive value. On the other hand, the hit/CR rates, $Pr(Alert \mid Signal)$ and $Pr(No\ alert \mid No\ signal)$, are less usable, because the human operator cannot directly use the probabilities to guide their behaviors. Instead, the hit/CR rates need to be integrated with the base rate in order to generate useful information in guiding behaviors. And this particular integration process, known as Bayesian inference, is very difficult (Kahneman, 2011). To better illustrate the idea of Bayesian inference, consider the following scenario:

> An airport security officer detects threats with the help of an nearly perfect decision aid. The alarm of the decision aid goes off if it recognizes a threat. The security officer could also manually check any luggage. The decision aid is correct 95 percent of the time. In other words, if there is a threat, the decision aid recognizes it with a 95 percent probability (Hit rate = $Pr(Alarm \mid Threat) = 95\%$), and if there is no threat, the aid shows no threat with a 95 percent probability (CR rate = $Pr(No\ alarm \mid No\ threat) = 95\%$). Suppose threats are rare in the airport, on average occurring only 1 percent of the time. If an alarm went off, should the officer panic and what would be the chance that there was actually a threat?

In the example, the hit rate is 95%. However, it does not mean that when an alarm goes off, there is 95% chance that there would be a threat. To answer the question correctly, we need to apply the Bayes' rule to calculate the positive predictive value, mathematically the ***inverse*** of the hit rate:

$$Pr(Threat \mid Alarm) = \frac{Pr(Alarm|Threat)Pr(Threat)}{Pr(Alarm)} =$$

$$\frac{Pr(Alarm|Threat)Pr(Threat)}{Pr(Alarm|Threat)Pr(Threat)+Pr(Alarm|No\ threat)Pr(No\ threat)} = \frac{95\%\times 1\%}{95\%\times 1\%+5\%\times 99\%} = 16\%$$

The probability of a true threat is only 16%! If we do not consider the payoff structure associated with the task (i.e. high cost if missing a threat), the result

indicates that probabilistically the officer only needs to manually check 16 luggage out of 100 alarms, and could invest his or her time on other tasks 84% of the time when alarms go off.

Prior research shows that it is cognitively demanding to use the Bayes' rule (Bar-Hillel, 1980; Cosmides & Tooby, 1996; Goodie & Fantino, 1996; Kahneman, 2011), because of several reasons. First, the base rate may not be readily available and an operator needs to estimate it. Second, when making a probabilistic judgment, an operator may neglect the base rate of $Pr(Threat)$, that threats only occur 1% of the time (Kahneman, 2011). Third, a person might be confused about $Pr(Alarm \mid Threat)$ and its ***inverse***, $Pr(Threat \mid Alarm)$, as both are related to the probability of an accurate threat identification (Bar-Hillel, 1980). Due to the difficulty in performing Bayesian inference, we speculate that the hit/CR rates are the least direct.

The overall success likelihood, $\Pr(Success \mid Automation\ response)$, represents the probability of a true state (Hit or CR) given an automation response (Alert or No alert), and a higher probability means an operator should follow the automation more overall. The overall success likelihood alone only guides human operators' behaviors at an aggregated level – if the automation overall success likelihood is 80%, when the automation issues 100 suggestions (regardless what the suggestion is), 80 suggestions are correct. Despite the lack of granularity, we speculate that the overall likelihood information is more direct than the hit/CR rate, as it can be easily used to guide overall human behaviors.

Due to the influence of the two factors, we predicted that there would be significant differences in participants' trust, dependence and dual-task performance when presented with different types of likelihood information. In particular, disclosing hit/CR rate would be the least beneficial in fostering proper trust and dependence, and would lead to the worst task performance. Revealing the predictive values, in contrast, would be the most beneficial.

3. METHOD

This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board at the University of Michigan.

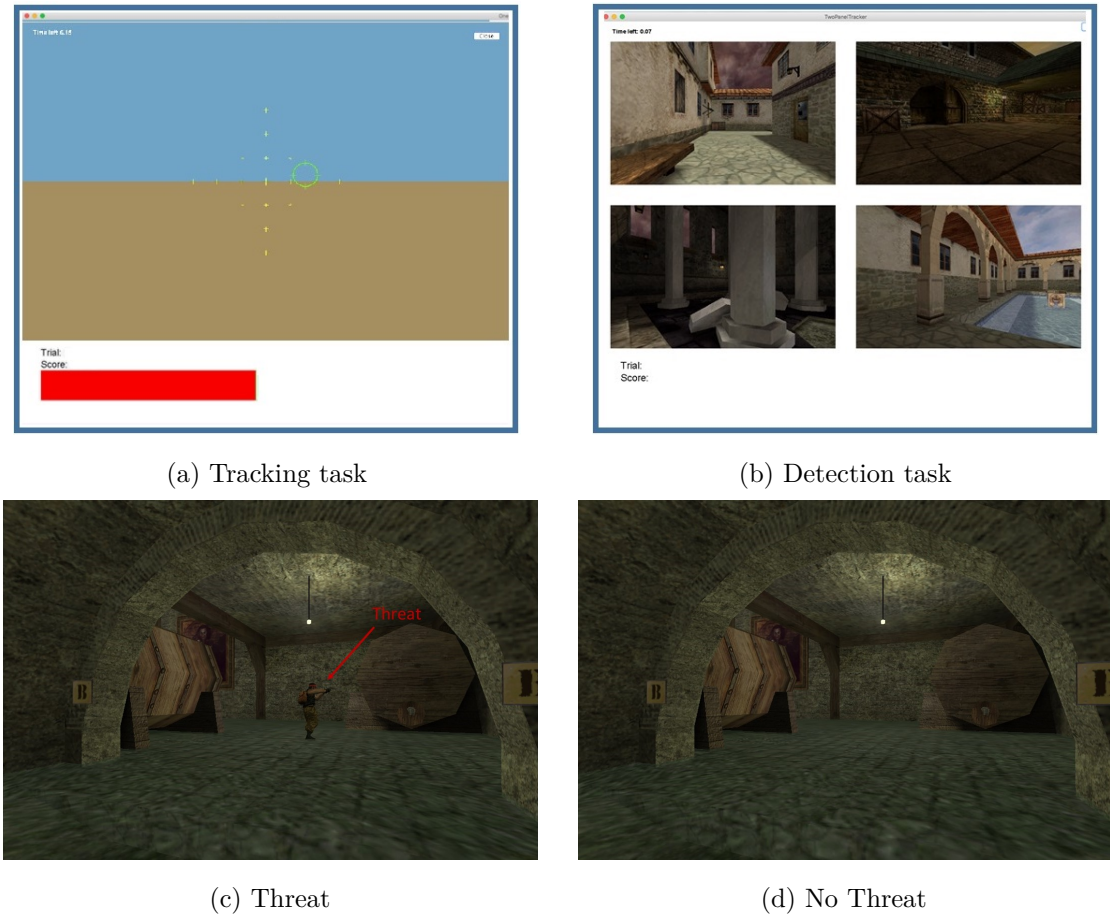
3.1 Participants

A total of 25 male and 36 female university students (average age = 22.28 years, SD = 4.88) with normal or corrected-to-normal vision participated in the experiment. Participants were compensated with $10 upon completion of the experiment. In addition, there was a chance to obtain an additional bonus of 1 to 5 dollars based on their performance.

3.2 Apparatus and stimuli

We used a simulated surveillance task in the experiment. In the experimental task, participants were asked to control the level of flight of a simulated swarm of drones, essentially a compensatory tracking task, and simultaneously detect potential threats in photo feeds from the drones (Figure 2). Participants were only able to access the display for either the tracking task or the detection task at any time and needed to toggle between the two displays. The simulated surveillance task was programmed using Java and the experiment was run on a 24 inch monitor.

Tracking task. Each trial started on the tracking display and lasted 10 seconds. The tracking task was programmed based on the PEBL (The Psychology Experiment Building Language) compensatory tracker task (http://pebl.sourceforge.net/battery.html). Participants used a joystick to move a randomly-drifting green circle to a crosshair located at the center of the screen – i.e. minimize the distance between the green circle and the crosshair as shown in Figure 2 (a). When a trial started, the green circle started at the centre of the crosshair. The position of the circle is a function of its previous position, its velocity, and the actions of three forces. The first force is the user input. The second force is a buffeting force composed of six sine waves at different amplitudes, frequencies and phase angles. The



(a) Tracking task (b) Detection task

(c) Threat (d) No Threat

Figure 2. Dual-task environment in the simulation testbed

third force simulates the force of gravity that causes the circle to slip on an unseen slippery surface. As a result of the buffeting force and the gravitational force, the circle drifts randomly. The performance of the tracking task is measured by two metrics: the Root Mean Square of the tracking errors (RMSE) and the tracking score ranging 0-10. The tracking error – the distance in pixels between the location of the circle and the crosshair, was measured at a frequency of 20Hz. The RMSE was calculated as $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(Tracking\ Error)^2}$, where $n = 200$. The tracking score was calculated using a 10-bin histogram of the RMSE distribution based on a dataset collected in a prior study (Yang et al., 2017).

Detection task. Besides the tracking task, every trial participants received a new set of four images from the simulated drones and were responsible for threat detection. The four images were static during every trial as shown in Figure 2(b). The

threat was a person as shown in Figure 2(c) and only one threat would present in one of the four images. There was no distractor in the four images and the participants did not determine if the person was a friend or a foe. The distribution of the threats across the four images followed a uniform distribution. Participants performed the detection task with the help of an imperfect automated threat detector. If the detector recognized a threat, the alert "Danger" went off immediately when a trial started in both visual and auditory modalities. The visual red alert was only shown on the tracking display (Figure 2(a)) and the auditory notification was a synthetic sound of "Danger". Participants were expected to identify the presence of the threat by pressing the "Report" button on the joystick as accurately and quickly as possible. The participant could follow the decisions of the threat detector blindly, or check the images in person and make his or her own decisions. If the detector identified no threat, the alert was silent. Participants did not report the absence of threat, i.e., participants were expected to perform no action when there was no threat. The performance of the detection task is measured by detection time, detection accuracy and detection score (Please refer to the Scoring System section).

Toggle between two displays. Every trial started on the tracking display. Participants were only able to access one display at a time, and needed to toggle between the displays of the tracking and the detection tasks using a "Switch" button on the joystick. There was a 0.5-second time delay every time they toggled between the displays, simulating the time for computer processing and loading the displays. The time stamp and the number of occurrences of participants pressing the "Switch" button were tracked automatically by the program.

Scoring system. In the experimental task, participants performed the tracking task and the detection task simultaneously, and needed to make a trade-off decision on which task to perform at any time, i.e. if they decide to check the four images, they would probably earn more points in the detection task but fewer points in the tracking task, and vice versa. Therefore, a pay-off structure has to be determined to eliminate potential bias toward either the tracking or the detection task by ensuring that the

potential gain in one task is approximately equal to the opportunity cost in the other task. A pilot study was conducted to determine the payoff structure (Please refer to Appendix A for more details). As a result, every trial participants could obtain 0-10 points for the tracking task and 0-5 points for the detection task.

$$Detection\ score = \begin{cases} 0 & Detection\ is\ wrong \\ 5 - 5 \times \frac{Detection\ Time}{10000\ milliseconds} & Detection\ is\ correct: Hit \\ 5 & Detection\ is\ correct:\ CR \end{cases}$$

3.3 Experimental Design

The experiment adopted a 2 (automation reliability: low vs high) ×3 (likelihood information: overall success likelihood, predictive values, and hit/CR rates) mixed design with automation reliability as the within-subjects factor and likelihood information as the between-subjects factor.

The reliability of the automated threat detector was configured based on SDT. In the present study, the criterion c was set at -0.25 and sensitivity d' at 1.5 or 3, resulting in automation with low and high reliability (Table 1). Benching marking prior literature (McBride, Rogers, & Fisk, 2011; Wiczorek & Manzey, 2014; Yang et al., 2017), we set the base event rate at 30%. Based on the preset c, d' and base rate, the number of occurrences of hits, misses, FAs and CRs were calculated and rounded to integers.

TABLE 1: *Corresponding hits, misses, false alarms and correct rejections*

Reliability	c	d'	Alert	Threat	No threat
Low	-0.25	1.5	Danger	13	11
			Clear	2	24
			Alert	Threat	No threat
High	-0.25	3	Danger	14	4
			Clear	1	31

Different likelihood information was calculated as follows.

Overall success likelihood = $\frac{Hits + CRs}{Hits + Misses + FAs + CRs}$ = 74% or 90%

Positive predictive value = $\frac{Hits}{Hits + FAs}$ = 54% or 78%

Negative predictive value = $\frac{CRs}{Misses + CRs}$ = 92% or 97%

Hit rate = $\frac{Hits}{Hits + Misses}$ = 87% or 93%

Correct rejection rate = $\frac{CRs}{FAs + CRs}$ = 69% or 89%

3.4 Measures

Trust. We measured participants' subjective rating of trust using a visual analog scale (Wiczorek & Manzey, 2014). The leftmost anchor of the the trust scale indicated, "I don't trust the detector at all" and the rightmost anchor "I absolutely trust the detector". The visual analog scale was then converted to a 0-100 scale. As part of the testbed design, in addition to trust ratings, participants needed to report their self-confidence in performing the task without the decision aid and perceived reliability of the decision aid. As the two measures were less relevant to this study, we did not report the data analysis results.

Compliance and Reliance. We also assessed participants' compliance and reliance behaviors. Compliance and reliance were operationally defined as the possibility of a participant blindly following the recommendation given by the automated threat detector without crosschecking the detection display. In particular, compliance was calculated as the possibility that a participant blindly reports a threat upon receiving a "Danger" alert without crosschecking the detection display, and reliance the possibility that the participant neither reports nor crosschecks when the detector is silent.

Compliance = $Pr(\,Report\ AND\ not\ crosschecking \mid\ Alert)$

Reliance = $Pr(\,Not\ reporting\ AND\ not\ crosschecking \mid\ No\ alert)$

Performance. The performance of the detection task was measured by the detection accuracy and detection time, as well as the detection score. The performance of the tracking task was calculated using the RMSE and the tracking score. The combined performance of both task was calculated as the sum of the detection score and the tracking score.

3.5 Experimental procedure

Upon arrival, participants provided informed consent and filled out a demographics form. Afterward, participants received practice on the tasks. The practice session consisted of a 30-trial block with the tracking task only and an 8-trial block of combined tasks, where participants experienced 2 hits, 2 misses, 2 false alarms and 2 correct rejections. Participants were informed that the automated threat detector used in the practice was just for illustration purpose. Afterward, they were randomly assigned to one of the three likelihood information conditions. A table similar to Figure 1 was then shown to the participants. Based on the condition a participant was assigned to, the definition, the meaning and the calculation of a particular likelihood information were introduced to the participant. In order to ensure that participants understood the likelihood information, the participants were given an example with different number of hits, misses, FAs and CRs, and were asked to calculate the likelihood information themselves. If a participant had difficulty doing so, the verbal definitions were reiterated and shown again to the participant, with potential further clarification on specific terms, until the correct answer was reached by the participant.

The experiment consisted of two 50-trial blocks with different automation reliability. The order of automation reliability was counterbalanced. Participants were verbally informed of the values of the likelihood information prior to the experiment. A text message showing the probability was also present throughout the experiment. Prior to the onset of each trial, there was a splash screen with a 3-second countdown timer. After every trial, participants were informed of the detection accuracy, the tracking score and the detection score they obtained in this trial and the accumulative scores they had obtained so far. After every 5 trials, participants indicated their trust. Participants were told that their subjective ratings should be based on all the trials they have completed so far, instead of just the previous 5 trials.

4. RESULTS

Data from one participant were excluded from analysis as his tracking task performance was below three standard deviations from the mean. All hypotheses were tested using data from the remaining 60 participants. We used mixed design analysis of covariance (ANCOVA) to analyze the relationship between independent variables and dependent variables. Participants' tracking task performance (last ten trials) in the practice session, was used as the covariate for analysis. The α level was set at .05 for all statistical tests. All post hoc comparisons utilized a Bonferroni α correction.

TABLE 2: *Mean and Standard Error of dependent variables in each condition*

	Low Reliability, $c = -0.25$, $d' = 1.5$			High Reliability, $c = -0.25$, $d' = 3$		
	Overall success likelihood	Predictive values	Hit/CR rates	Overall success likelihood	Predictive values	Hit/CR rates
Trust	53.15 ± 4.98	53.21 ± 3.70	58.82 ± 4.22	76.53 ± 2.79	71.60 ± 2.57	69.55 ± 4.92
Compliance (%)	31.67 ± 7.27	17.92 ± 6.24	33.13 ± 7.58	58.33 ± 9.30	47.78 ± 7.51	54.72 ± 9.03
Reliance (%)	46.54 ± 6.67	83.85 ± 5.36	37.69 ± 9.07	79.69 ± 5.92	90.78 ± 5.01	59.84 ± 8.68
Detection time (ms)	2518.90 ± 228.54	2623.29 ± 272.74	2275.49 ± 217.99	1655.10 ± 242.21	2002.61 ± 173.16	1736.20 ± 236.60
Detection accuracy (%)	84.10 ± 2.17	86.00 ± 1.68	85.10 ± 2.11	89.80 ± 0.93	91.30 ± 0.87	90.40 ± 0.72
Detection score	3.91 ± 0.10	4.03 ± 0.06	3.96 ± 0.08	4.27 ± 0.04	4.33 ± 0.04	4.30 ± 0.03
Tracking error	60.30 ± 3.28	58.37 ± 4.23	67.89 ± 3.70	47.10 ± 3.31	46.78 ± 4.04	55.20 ± 3.53
Tracking score	6.66 ± 0.27	6.95 ± 0.34	5.97 ± 0.32	7.87 ± 0.28	7.91 ± 0.32	7.08 ± 0.32
Total score	10.58 ± 0.24	10.97 ± 0.29	9.93 ± 0.27	12.15 ± 0.27	12.22 ± 0.32	11.38 ± 0.32

4.1 Subjective trust

Trust. Participants had higher trust in the automated threat detector as automation reliability increased ($F(1, 56) = 7.533$, $p = .008$). However, the effect of likelihood information was non-significant.

4.2 Compliance and Reliance

Figure 3 shows the participants' compliance and reliance behaviors.

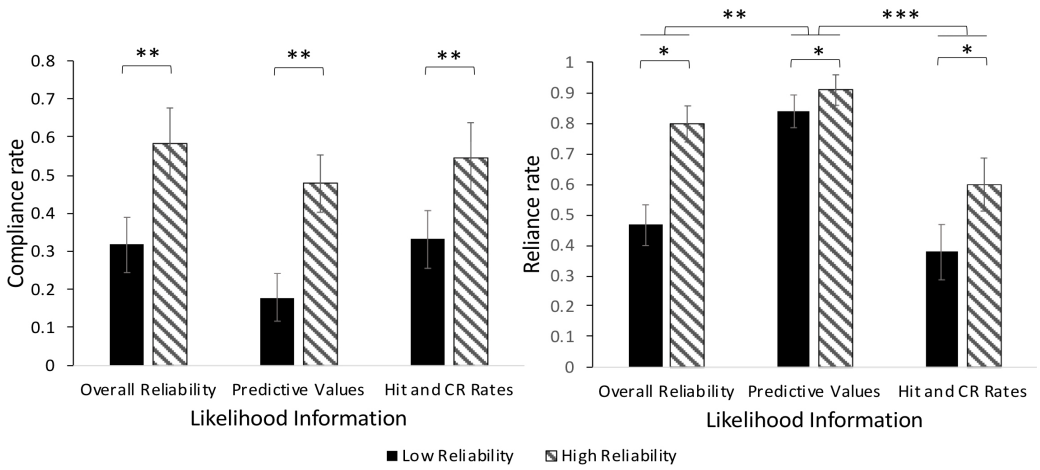


Figure 3. Compliance with and reliance on automated threat detector

*** Difference is significant at the 0.001 level; ** Difference is significant at the 0.01 level; * Difference is significant at the 0.05 level (2-tailed).

Compliance. Higher automation reliability led to higher compliance rate on the automated threat detector ($F(1, 56) = 7.196$, $p = .01$). The effect of likelihood information was non-significant.

Reliance. Automation reliability ($F(1, 56) = 5.905$, $p = .018$) and likelihood information ($F(2, 56) = 10.752$, $p < .001$) significantly affected reliance rate. Higher automation reliability led to higher reliance. Moreover, providing participants with predictive values led to higher reliance on the automated threat detector, compared to the overall success likelihood condition ($p = .009$) and the hit/CR rates condition ($p < .001$). There was also a significant two-way interaction between automation reliability and likelihood information ($F(2, 56) = 4.807$, $p = .012$). When automation reliability was low, participants relied on the automated threat detector the most when they were informed of the predictive values (predictive values > overall success likelihood: $p < .001$; predictive values > hit/CR rates: $p < .001$). As reliability increased, the reliance rate was significantly higher when participants were provided with predictive values relative to the hit/CR rates ($p = .004$).

4.3 Performance

Detection performance. As depicted in Figure 4, participants detected threats more accurately ($F(1, 56) = 9.702$, $p = .003$) and faster ($F(1, 56) = 8.659$, $p = .005$)

and gained higher scores ($F(1, 56) = 14.633$, $p < .001$) when automation reliability increased. However, the effect of likelihood information was not significant.

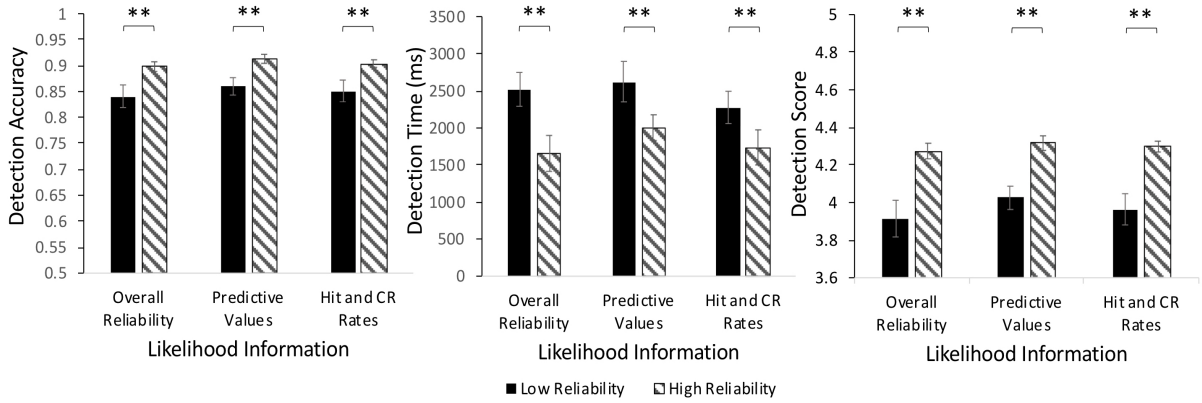


Figure 4. Detection task performance

Tracking performance. As shown in Figure 5, there were significant main effects of automation reliability ($F(1, 56) = 4.37$, $p = .041$) and likelihood information ($F(2, 56) = 5.381$, $p = .007$) on tracking score. Post hoc analysis indicated that when participants were presented with hit/CR rates, they had the lowest tracking score (hit/CR rates < predictive values: $p = .038$; hit/CR rates < overall success likelihood: $p = .011$).

Additionally, there was a significant effect of likelihood information ($F(2, 56) = 4.311$, $p = .018$) on RMSE. When participants were presented with hit/CR rates, they had the a higher RMSE (hit/CR rates > overall success likelihood: $p = .019$). The main effect of automation reliability on RMSE was not significant.

Combined performance. The main effects of automation reliability ($F(1, 56) = 10.744$, $p = .002$) and likelihood information ($F(2, 56) = 6.293$, $p = .003$) were significant (Figure 6). Participants obtained higher combined scores as automation reliability increased. There was also a difference among the three types of likelihood information. Post hoc analysis revealed that participants informed of overall success likelihood or predictive values, instead of the hit/CR rates, had higher total scores (overall success likelihood > hit/CR rates: $p = .008$; predictive values > hit/CR rates: $p = .014$).

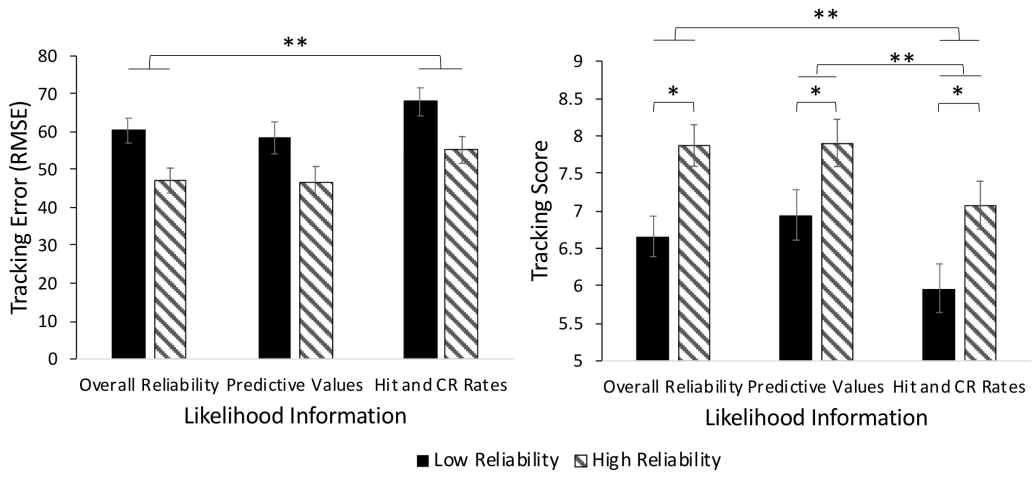


Figure 5. Tracking task performance

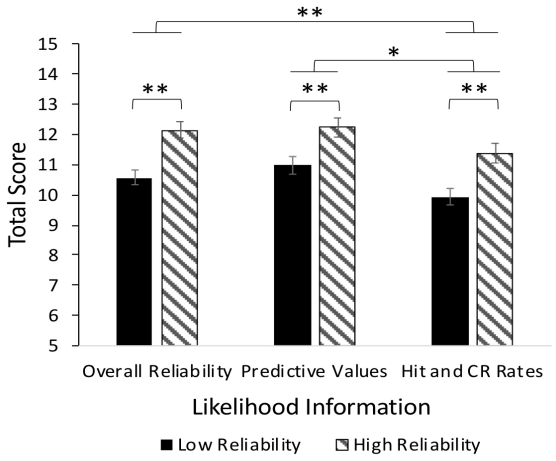


Figure 6. Total task performance

5. DISCUSSION

In the present study, we predicted that there would be significant differences in participants' trust, dependence and dual-task performance when presented with different types of likelihood information. In particular, disclosing hit/CR rate would be the least beneficial in fostering proper trust and compliance and reliance behaviors, and would lead to the worst task performance. Revealing the predictive values, in contrast, would be the most beneficial. We discuss how the results support our prediction.

Trust in automation

Our results indicate a non-significant difference on trust between the three types of likelihood information. The lack of significance might have been due to two reasons.

First, the sensitivity of a uni-dimensional trust scale might not be as high as that of a multi-dimensional scale. A uni-dimensional scale has the advantage of easy implementation. However, it might not be able to capture the different dimensions underlying the concept of trust compared to the multi-dimensional scales. Two widely used multi-dimensional scales (Jian, Bisantz, & Drury, 2000; Muir & Moray, 1996) have 12 and 7 questions, respectively. Second, the reliability of the threat detector was consistent across the different types of likelihood information. Therefore, the judgment of trust might be largely based on the true performance of automated detector instead of the likelihood information presentation. Further research is needed to systematically examine potential differences between uni-dimensional and multi-dimensional trust scales.

Compliance and Reliance behaviors

We found a significant difference in reliance and a non-significant difference in compliance between the three types of likelihood information. Disclosing the predictive values led to higher and more appropriate reliance, compared to the overall success likelihood condition and the hit/CR rates condition. We argue that the predictive values can be considered as the gold standard of optimal behaviors. The negative predictive value, $\Pr(No\ signal \mid No\ alert) = x\%$, means that when the automation is silent, there is $x\%$ chance that a site is clear. Therefore, probablistically speaking, if the threat detector is silent for 100 cases, the human operator only need to check $100 - x$ sites in person.

In our study, the negative predictive value was 97% for the high reliability automation, and 92% for the low reliability automation. Therefore, a rational strategy for the human operator is to cross-check only a small number of sites, and to allocate more resource on the tracking task. When presented with negative predictive values, participants' reliance rates were 90.8% and 83.8%, respectively (see Figure 3 and Table 2), which were fairly close to the optional values of 97% and 92%. When informed of the overall likelihood, the observed reliance values were 79.7% and 46.5%, further away

from the optimal values; When presented with the hit/CR rates, the observed reliance values were 59.8% and 37.7%, furthest away from the optimal values. In the present study, the base rate was set to be 30%. In real life, bases rates of critical events are usually much lower (Parasuraman & Riley, 1997). With a lower base rate, most of the time the automated decision aid would be silent, and the benefit of presenting the predictive values would be further enhanced, as the predictive values promote proper reliance behaviors.

We failed to find a significance in participant's compliance behaviors. This lack of significance might have resulted from participants' strategies between the detection and the tracking tasks. The positive predictive values were 78% for the high reliability automation and 54% for the low reliability automation. However, across all the likelihood conditions, the compliance rates were considerably lower than the optimal values (see Figure 3 and Table 2). This suggests that participants cross checked the detection display much more frequently than they should have done. This is further supported by our observation: participants mentioned that the tracking task was fairly boring and they preferred to cross-checking the detection display even if the strategy was not optimal. The unnecessary cross-checking behaviors allowed the participants to detect threats that the automated detector failed to recognize and contributed to a similar performance in the detection task.

Performance

Our results indicate a significant difference in tracking task and non-significant difference in detection task. The tracking performances in the predictive value condition and the overall likelihood condition were better than that in the hit/CR rate condition. Such results are attributable to participants' reliance and compliance behaviors. When presented with hit/CR rates, participants' reliance behaviors were the least optimal, which means they cross-checked much more frequently than they should have done. Every time a cross-checking was performed, participants could not access the tracking display, hurting the tracking performance. In addition, as mentioned before, the similar

compliance behaviors resulted in the similar performance in the detection task.

The observed pattern on tracking and detection performance suggests that the automated threat detector was largely used as a tool for attention management in multitask environments, benefiting the continuous unaided task (i.e. the tracking task), rather than a tool directly benefiting the aided task (i.e. the detection task). The result support the findings of Wiczorek and Manzey (2014).

In addition, we also observed a difference in the combined task performance. Disclosing predictive values and overall likelihood information led to higher combined performance than the hit/CR rates condition. We note the importance of obtaining an explicit pay-off structure with the same unit of measurement. Most of the prior literature did not report the combined task performance, largely because different tasks were measured in different units and a combined task performance score was impossible to obtain.

At last, consistent with findings from previous studies, our results showed that as the automated threat detector became more reliable, participants' trust in and dependence on the threat detector increased, and their dual task performance improved. (Neyedli et al., 2011; Walliser et al., 2016; Wang et al., 2009).

6. CONCLUSION

Although disclosing likelihood information has been proposed as a design solution to promote proper trust and dependence, and to enhance human-automation team performance, prior studies showed mixed results (Bagheri & Jamieson, 2004; Dzindolet et al., 2002; Fletcher et al., 2017; Walliser et al., 2016; Wang et al., 2009). The goal of this study was to experimentally examine the effects of presenting different types of likelihood information. Based on the framework of SDT, we categorized likelihood information calculated in prior literature into three types: overall success likelihood, predictive values, and hit/CR rates.

The present study offered a framework to summarize existing literature pertaining to disclosing likelihood information. Our results showed that not all likelihood

information is equally useful. Simply presenting the hit/CR rates should be avoided. Our findings can be applied to a wide array of domains such as urban search and rescue (USAR), medical diagnosis and TSA, where the hit/CR rates are often readily available but not the predictive values and overall likelihood information. Hit/CR rates, also known as sensitivity and specificity (Altman & Bland, 1994), are referred to as the diagnostic information (Please note that the sensitivity as hit rate is different from the sensitivity d' in SDT). Often, the diagnostic information is more accessible to people. For instance, physicians are often provided with the diagnostic information when a new test is introduced: The HIV test is 99% accurate – if a patient is infected by HIV, there is 99% chance the test will show a positive result; if a patient is healthy, there is 99% chance the test will show a negative result.

Efforts should be made to clarify the meanings of different types of likelihood information when an automated decision aid is introduced. Prior research indicates that people could be confused about predictive values and hit/CR rates (Bar-Hillel, 1980). In real life, bases rates of critical events are usually very low (Parasuraman & Riley, 1997). With a lower base rate, for instance, 1% in the airport security officer example, the discrepancies between the predictive values and the hit/CR rates would be even larger. Mis-attributing hit/CR rates as predictive values would lead to more detrimental outcomes.

The findings should be viewed in light of the following limitations. First, consistent with prior research, we did not provide participants with the base rate and they had to estimate it by themselves. Future study can present base rate to participants and examine whether people can utilize hit/CR rates more appropriately. Base rate can also be manipulated in further research to examine the effects of likelihood information when base rate is extremely low. Second, we used probabilities instead of natural frequencies to present likelihood information. Previous studies have shown that reasoning with natural frequencies results in more accurate inference (Gigerenzer & Hoffrage, 1995; Hoffrage, Hafenbrädl, & Bouquet, 2015; Mandel, 2014). A future study could compare the differences of presenting probabilities and natural

frequencies. Third, the criterion c in this study was set to be liberal, which led to more false alarms than misses. Future studies should examine the effects of likelihood information with different d' and c.

Keypoints

- We proposed a framework to summarize existing literature pertaining to disclosing likelihood information and classified the calculation of likelihood information into three categories: overall likelihood value, predictive values, and hit and correct rejection (CR) rates.
- Human operators informed of the overall likelihood value or the predictive values, rather than the hit and correct rejection (CR) rates, relied on the decision aid more appropriately.
- Human operators informed of the overall likelihood value or the predictive values, rather than the hit and correct rejection (CR) rates, performed better on the tracking task and obtained higher combined task scores.
- As automation reliability increased, trust, compliance, reliance and performance increased accordingly.

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ*, *308*(6943), 1552.

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics* (pp. 212–217).

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233.

Comstock, J. R., & Arnegard, R. J. (1992). *The Multi-attribute Task Battery for Human Operator Workload and Strategic Behaviour Research* (Tech. Rep. No. 104174). National Aeronautics and Space Administration.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*(4), 564–572.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, *44*(1), 79–94.

Fletcher, K. I., Bartlett, M. L., Cockshell, S. J., & McCarley, J. S. (2017). Visualizing probability of detection to aid sonar operator performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, pp. 302–306).

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, *102*(4), 684.

Goodie, A. S., & Fantino, E. (1996, mar). Learning to commit or avoid the base-rate error. *Nature*, *380*, 247.

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*, 642.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Faraus, Straus and Girous.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Macmillan, N. A., & Creelman, C. D. (2005). Mahwah, NJ: Lawrence Erlbaum.

Mandel, D. R. (2014). The psychology of bayesian reasoning. *Frontiers in Psychology*, *5*, 1144.

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011, November). Understanding the Effect of Workload on Automation Use for Younger and Older Adults. *Human Factors*, *53*(6), 672–686.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665.

Muir, B. M., & Moray, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors*, *53*(4), 338–355.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253.

Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. (2019). Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance.

SAE International Journal of Connected and Automated Vehicles, *2*(2).

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors*, *30*(4), 445-459.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

Walliser, J. C., de Visser, E. J., & Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, pp. 133–137).

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*(3), 281-291.

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors*, *56*(7), 1209–1221.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)* (pp. 408–416).

Biographies

Na Du is a Ph.D. student in the Department of Industrial and Operations Engineering at the University of Michigan Ann Arbor. She completed a B.S. in Psychology in Zhejiang University, China.

Kevin Y. Huang is an undergraduate student studying Industrial and Operations Engineering at the University of Michigan Ann Arbor.

X. Jessie Yang is an Assistant Professor in the Department of Industrial and Operations Engineering and an affiliated faculty at the Robotics Institute, University of Michigan Ann Arbor. She obtained her PhD in Mechanical and Aerospace Engineering (Human Factors) from Nanyang Technological University Singapore in 2014.

Appendix A: Pilot study

We conducted a pilot study to create a scoring system for the experiment. In the experimental task, participants performed the tracking task and the detection task simultaneously. Participants were required to make a trade-off decision on which task to perform at any time, i.e. if they decided to check the four images, they would probably earn more points in the detection task but fewer points in the tracking task, and vice versa.

Therefore, a pay-off structure has to be determined to eliminate potential bias toward either the tracking or the detection task by ensuring that the potential gain in one task is approximately equal to the opportunity cost in the other task. In order to determine the parameters of the scoring system, first we set the tracking task score on a scale from 0 to 10, based on the distance of the green circle from the center of the crosshair. Next, we defined the detection task score as a function of the detection accuracy and time, i.e. $a + b \times \frac{Time}{10000\ milliseconds}$. To determine a and b, a total of 10 participants between the age of 19 and 23 participated in the pilot study. They performed a tracking task only block and a combined task block, each with 50 trials, with a 5-minute break in between. In the combined task block, participants performed both tasks and were instructed that two tasks were equally important. They could optimize their performance by minimizing the distance between the green circle and the center of the display, and by detecting the threats as accurately and as quickly as possible. The block order was counterbalanced. One participant's data was removed from data analysis due to his significantly poor performance on the tracking task. The results showed that when doing both tasks concurrently, participants lost on average a score of 3.7 points on their tracking tasks (Tracking task only block: $M = 8.8$, $SD = 1.2$; Combined task block: $M = 5.1$, $SD = 1.1$). We then varied a and b to make sure they would gain approximately 3.7 points with a similar SD from the detection task. As a result, $5 - 5 \times \frac{Detection\ Time}{10000\ milliseconds}$ was set to be the scoring scheme of the detection task. In each combined task trial, it is possible to obtain a maximum score of 15 points, 10 points from the tracking task and 5 points from the detection task.