Adapting to Imperfect Automation: The Impact of Experience on Dependence Behavior and Response Strategies

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

This study examined the impact of experience on individuals' dependence behavior and response strategies when interacting with imperfect automation. 41 participants used an automated aid to complete a dual-task scenario comprising of a compensatory tracking task and a threat detection task. The entire experiment was divided into four quarters and multi-level models (MLM) were built to investigate the relationship between experience and the dependent variables. Results show that compliance and reliance behaviors and performance scores significantly increased as participants gained more experience with automation. In addition, as the experiment progressed, a significant number of participants adapted to the automation and resorted to an extreme use response strategy. The findings of this study suggest that automation response strategies are not static and most individual operators eventually follow or discard the automation. Understanding individual response strategies can support the development of individualized automation systems and improve operator training.

INTRODUCTION

Automated/autonomous technology has become an integral part of our daily lives, from self-driving cars and drones to intelligent personal assistants and robots. If used properly, they offer many benefits such as improved safety (Gombolay et al., 2018; Luo et al., 2021) and performance (Luo, Du, & Yang, 2022; Wickens & Dixon, 2007), and reduced operator workload (Lu, Zhang, Ersal, & Yang, 2019; Luo et al., 2021).

Since the seminal paper by Parasuraman and Riley (1997), research has extensively explored aspects of automation use and the impacts on human-automation team performance. These aspects range from general, such as synthesizing the literature to estimate when automation performs well enough for the benefits to outweigh the costs (Wickens & Dixon, 2007), to the specific, by conducting a laboratory experiment to examine how automation error types affect performance (Dixon, Wickens, & McCarley, 2007). In the former, they found that performance is sensitive to the automation imperfections and estimated that automation performing correctly less than 70% of the time may not be worth using. In the latter, they found that human-automation team performance deteriorates differently based on the automation error type; false-alarm prone automation has been considered more damaging to performance than miss-prone automation. Others mentioned that this performance detriment is furthered when operators place undue trust and dependence on the automation (Lee & See, 2004; Parasuraman & Riley, 1997). Individual traits and task specific factors have also been explored, such as the effects of workload and age on automation dependence (McBride, Rogers, & Fisk, 2011). They found that the performance and compliance of younger adults was affected by the workload, while older adults did not change as the workload increased.

However, there is a gap regarding how automation use changes as operators repeatedly use automation. To address this gap, we conducted a laboratory experiment investigating how human operators change their decision-making and response strategies with increased use of imperfect diagnostic automation.

Dealing with uncertainty

Human interaction with diagnostic automation, a form of automation that analyzes raw information and infers the status of the world (Wickens & Dixon, 2007), is considered decision-making under uncertainty (Manzey, Gérard, & Wiczorek, 2014; Meyer, 2004; Sorkin & Woods, 1985; Wang & Yang, 2022; Yang, Unhelkar, Li, & Shah, 2017). As the most basic form of diagnostic automation, binary diagnostic automation categorizes the world into 'signal present' and 'signal absent' and alerts the human (often in the form of an alarm) once a signal is detected. The human operator is typically responsible for the final decision concerning if a signal is *actually* present.

Compliance and reliance refer to how an operator responds to automation alarms/non-alarms (Meyer, 2001, 2004). Compliance quantifies the human's response during an automation alarm and reliance quantifies the human's response when the automation is silent (i.e., non-alarm). The availability of alarm validity information (AVI) adds complexity to the operator's decision-making process (Allendoerfer, Pai, & Friedman-Berg, 2008; Sorkin & Woods, 1985). AVI is defined as information that can be used to validate the alarm or non-alarm; it reduces uncertainty in the operator's decision-making by informing about the ground truth. Compliance and reliance can be further divided, based on availability of AVI, into blind compliance and reliance, and verified compliance and reliance. The blind responses are categorized by when the operator blindly follows the automation's alarms/non-alarms. Verified responses are categorized by when the operator verifies the automation's alarms/non-alarms with additional information, such as AVI. For example, if a smoke detector started ringing, someone could comply with the alarm and evacuate immediately (i.e., blind compliance) while others could search the area for fire or smoke before deciding to exit (i.e., verified compliance).

Extreme response strategy

Prior research analyzed the occurrence of blind compliance and reliance and identified the use of the extreme response strategy. An extreme response strategy occurs when the opera-

tor either uses or discards the automated aid (Bliss, Gilson, & Deaton, 1995). Theoretically, the rational operator could use information about automation performance and situational specific consequences (costs and benefits) to model the outcome. They could then calculate a threshold and decide to always use or ignore the alarm to maximize their performance. This strategy was noted in the findings of Bliss (1993), where a small number of participants (just under 10%) resorted to an extreme response strategy. The participants appeared to be sensitive to alarm performance: they applied an extreme disuse strategy to automation that was 25% correct and an extreme use strategy to automation that was 75% correct. In a subsequent review, Bliss (2003), compared extreme response strategies of groups informed about collective alarm performance to groups that could access AVI (trial specific information to validate the alarm). They found that the participants who could not validate alarms were much more likely to resort to an extreme use response strategy. Building upon the findings of those studies, Manzey et al. (2014) did a controlled laboratory study with 4 experiments and examined the effects of having AVI, effort to validate alarms, and workload on response strategies. In addition to finding that AVI directly impacted operator response strategies, there was a reduced amount of cry wolf effect (coined from Breznitz (1984)) and AVI access increased operator sensitivity to non-alarms.

Our study differs from previous work by examining how people adapt their strategies to the automation throughout the experiment. We hypothesized that participants would adapt to the automation and adjust their strategies as they increase their experience with the automated aid.

METHODS

Participants

A total of 41 (25 males, 15 females, and 1 person who did not disclose gender information) university students (mean age = 24.05 years old, SD = 3.49) with normal or corrected to normal vision participated in the experiment. Upon completion of the experiment, the participant received US$20 with an opportunity to earn a performance bonus ranging from $1 to $5.

Apparatus and stimuli

The experimental displays were presented on a HP 24 in. monitor, with a 1920 x 1200 resolution. Participants used a Logitech Extreme 3D Pro joystick to interact with a simulated surveillance drone. Each participant performed 100 trials (each lasting 10 seconds) of a dual task scenario, which consisted of two simultaneous tasks: a compensatory tracking task and a threat detection task (Du, Huang, & Yang, 2020; Yang et al., 2017). Participants could only see one task display at a time and every trial began on the tracking task display. During the trial, participants could choose to toggle between task displays; this would occur a 0.5 second delay. An imperfect diagnostic automated aid would provide a binary recommendation about status of threats. Both displays can be seen below in Figure 1.

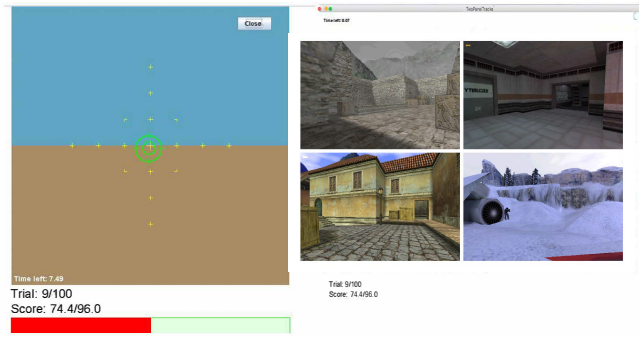

Figure 1. The tracking task is on the left and 4 static photos are on the right, only one display could be seen at a time.

Tracking Task. The tracking task was programmed based on the Psychology Experiment Building Language (PEBL) (Mueller & Piper, 2014). Participants used a joystick to center a randomly drifting green circle over a static, central cross-hair. During each trial, 10 points could be earned based on tracking task performance. Tracking error was defined as the pixel distance between the cross-hair and circle and it was collected at a frequency of 20 Hz. We calculated the RMSE from each trial's tracking error. The RMSE distribution from a previous study (Yang et al., 2017) was used to calculate a 10-bin histogram and determined the corresponding tracking score (out of 10).

Threat Detection Task. During each trial, participants were provided with a new set of 4 static photos, which could potentially display threats. Threats were presented as human figures and only 1 threat was shown at a time. There were no distraction stimuli and participants were instructed to treat every human figure as a threat. Threats were uniformly distributed across the four images. Participants could earn up to 5 points for each threat detection task, with a linear time penalty for threat detection time. An participant's incorrect detection (false alarm or miss) would result in a threat detection score of 0 for that trial. If participants correctly determined no threat, it would result in 5 detection task points - as no action was required. The following calculation was used during hits: Detection Score = 5- 5*(detection time/ total trial time).

Automated Aid. For the threat detection task, participants were supported by an imperfect automated aid, which would make a binary recommendation on the status of status of threats ("safe" or "danger") through audio and visual alarms. However, the participants were ultimately responsible for choosing to report a perceived threat, which was done via the joystick's trigger. We set the base rate to 30% (i.e., the amount of true signals) after bench-marking prior literature (Du et al., 2020; McBride et al., 2011; Wiczorek & Manzey, 2014; Yang et al., 2017). Of the 100 trials, there were 21 hits, 59 correct rejections, 11 misses, and 9 false alarms.

Procedure. Participants were recruited via email and invited to the lab. Upon arrival, they completed an informed consent form and begin training to become familiar with the scenario. During training, participants completed 30 practice trials of only the tracking task, followed by 8 practice trials of the dual task scenario. No automation likelihood information was disclosed to participants. A 3-second countdown preceded each trial and performance feedback was provided immediately after

each trial. A required 5-minute break was given after the 50th trial.

Independent and dependent variables

This study focused on the effects of repeated automation use on operators' compliance and reliance behaviors. Therefore, we divided the experimental session (100 trials) into 4 quarters (Q): Q1 = trials 1-25, Q2 = trials 26-50, Q3 = trials 51-75, and Q4 = trials 76-100.

The first set of dependent variables were dual task performance scores and the operator's compliance and reliance rates. In particular, we were interested in the operator's blind compliance and blind reliance behaviors, defined as the probability that the human follows the automation's recommendation without cross-checking the raw information and calculated using the following equations.

$$\begin{aligned} &\textit{Blind Compliance} = \\ &P(\textit{report and not cross checking}|\textit{automation alert}) \\ &\textit{Blind Reliance} = \\ &P(\textit{not report and not cross checking}|\textit{automation slience}) \end{aligned} \tag{1}$$

An additional dependent variable of interest was the operator's *extreme* automation use/disuse behaviors. According to prior research (Manzey et al., 2014), the participant applied an extreme strategy when he/she blindly complied with or relied on automation more than 90% or less than 10% of the time. In this study, the term extreme use refers to extreme high blind compliance and extreme high blind reliance (>90%), the term extreme disuse refers to extreme low blind compliance and extreme low blind reliance (>10%). Finally, we examined the proportion of time spent on the tracking task display (tracking task time/ total trial time) for each quarter.

Data Analysis. We used multi-level modeling (MLM) to analyze the relationship between the independent variable and dependent variables, using the 'nlme' package in R (version 4.2.2). Following the standard procedure for MLM building (Hofmann, Griffin, & Gavin, 2000), we started simple and gradually added complexity to our models. Next, we used the Analysis of Variance (ANOVA) test to compare the Akaike information criterion (AIC) scores between the simpler and more complex model, which determined whether the latter model was needed. Quarters of experiment were nested within participants, as this was a repeated measures design. In addition to the MLMs, the McNemar's test (Eliasziw & Donner, 1991) was performed to determine whether the proportions of extreme use/disuse strategy users was equal in the beginning (Q1) compared to the end (Q4) and a continuity correction was applied. The alpha for all statistical tests was set to 0.05.

RESULTS

Performance. The experimental quarter had a significant effect on the total performance score, t (122) = 8.14, p<0.001. The random-slope random-intercept model was used and β_1 = 14.75. When separating performance scores by task, the threat detection scores remained stable throughout the experiment while tracking task scores improved (details in Table 1).

Table 1. *Mean and SD values for performance scores, reliance and compliance rates.*

	Tracking Task Score	Detection Task Score	Compliance Rates	Reliance Rates
Q1	156.17 (48.71)	102.34 (12.17)	0.29 (0.36)	0.44 (0.36)
Q2	181.78 (43.81)	98.64 (13.57)	0.42 (0.41)	0.57 (0.35)
Q3	199.83 (42.12)	99.96 (12.70)	0.53 (0.44)	0.63 (0.36)
Q4	202.15 (38.94)	99.10 (13.18)	0.54 (0.46)	0.65 (0.39)

Blind Compliance. There was a significant effect of experimental quarter on operators' blind compliance rates t(122) = 3.99, p<0.001. The random-slope random-intercept model was used and β_1 = 0.09. As the experiment progressed, the blind compliance rates increased from 0.29 to 0.54, while the SD increased from 0.36 to 0.46. Box-plots with compliance rates in each quarter can be seen below in Figure 2.

Blind Reliance. There was a significant effect of experimental quarter on operators' blind reliance rates t (122) = 3.76, p<0.001. The random-slope random-intercept model was used and β_1 = 0.07. As the experiment progressed, the mean blind reliance rates increased from 0.44 to 0.65, while the SD remained fairly stable (it shifted from 0.36 to 0.39). Box-plots with reliance rates in each quarter can be seen below in Figure 3.

Extreme Use/Disuse. As the experiment progressed, the count of participants applying an extreme use strategy increased while the count for extreme disuse decreased (see Table 2). A significant difference in extreme high blind compliance counts was found between the beginning and end of the experiment $\tilde{\chi}$ (1) =7.56, p= 0.01. A significant difference was found in extreme high blind reliance counts between the beginning and end of the experiment $\tilde{\chi}$ (1) =10.08, p <0.01. No significant differences were found for the extreme low blind compliance (p=0.39) or low blind reliance (p=0.51).

The largest shift of participants (n = 7) to an extreme high compliance strategy occurred during Q3. Participants steadily shifted to an extreme high reliance strategy (n = 5) during Q2, Q3 and Q4. By quarter 4, the number of participants resorting to an extreme high compliance (n = 15) and extreme low compliance (n = 16) strategy were roughly equal. The number of participants applying an extreme high reliance strategy more than doubled from the beginning (quarter 1 = 8) to the end (quarter 4 = 20) of the experiment. Participants could resort to an extreme strategy during only alarms, only non-alarms, or they could have applied the extreme strategy to both; which is why the Q4 counts in Table 2 sum up to 57 for the 41 participants. The box-plots (Figures 2 and 3) display the group's increasing median compliance and reliance rates. Proportion of time spent on the tracking task also illustrated the group's eventual collective extreme use/disuse strategy. Participants increased the proportion of time spent on the tracking during the first half of the experiment: 89% in Q1, 92% in Q2, 94% in Q3, 94% in Q4.

Table 2. *Counts of participants with extreme automation use/disuse in response to alarm or non-alarm states.*

	Alarms		Non-Alarms	
	Extreme high blind compliance (# out of 41)	Extreme low blind compliance (# out of 41)	Extreme high blind reliance (# out of 41)	Extreme low blind reliance (# out of 41)
Q1	4	19	8	9
Q2	8	16	10	6
Q3	15	14	15	5
Q4	16			

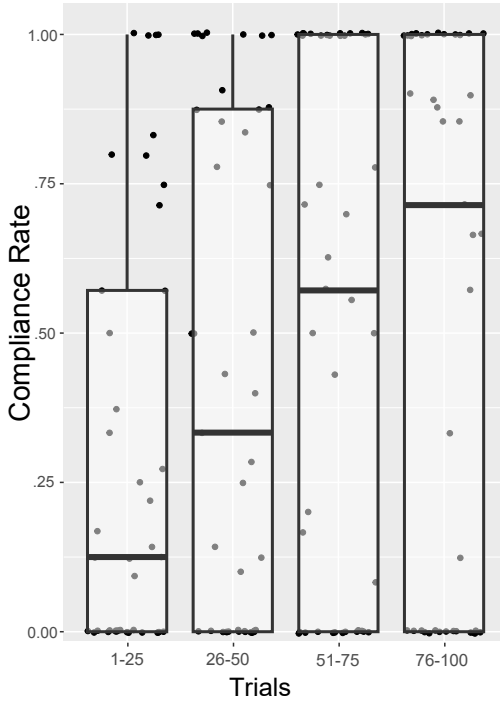

Figure 2. Box-plots illustrating operators' blind compliance rates (Y axis), divided into quarters of the experiment.

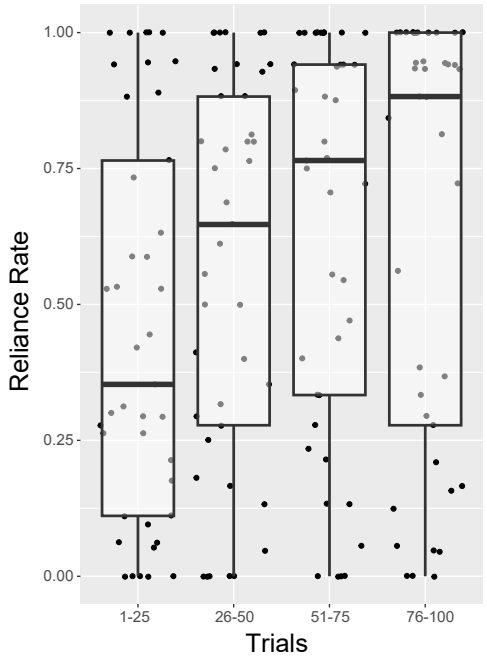

Figure 3. Box-plots illustrating operators' blind reliance rates (Y axis), divided into quarters of the experiment.

DISCUSSION

We predicted that participants would adjust their automation response strategy throughout the experiment. To test this, we examined operators' performance, compliance, and reliance across the 4 quarters of the experiment. Our results showed that compliance and reliance significantly increased and performance significantly improved. When specifically examining the performance score by task, the tracking task performance improved throughout the experiment while the threat detection task performance remained the same. The tracking improvement likely stems from participants spending more time focusing on that task. This is similar to other research Du et al. (2020), they noted how automation was used as an attention management tool instead of a tool that directly improved the task associated with the automated aid.

When examining compliance and reliance rates for the group, we see that operators generally increased their compliance and reliance rates throughout the experiment. Similar to prior results (Du et al., 2020; Manzey et al., 2014; Schuler & Yang, 2022; Wickens & Dixon, 2007), the changes in compliance and reliance behaviors were not random and appear to be systematic responses to alarm characteristics. In this study, the average blind compliance and blind reliance rates gradually became closer to the automation alarm and non-alarm performance as the experiment went on; the group appeared to use what has been referred to as a probabilistic matching strategy. The probabilistic matching strategy is a heuristic where the human's blind compliance and reliance rates roughly mirror the automation alarm/non-alarm performance (Bliss, 2003; Bliss et al., 1995; Dorfman, 1969).

When examining individual operator strategies, we see that the majority of individuals were not using a probabilistic matching strategy, but instead adapted an extreme response strategy. Our results show a significant change in the counts of extreme use strategy, people were increasing the amount of extreme high compliance and reliance. The extreme automation use strategy seems logical since the automation's performance was above the threshold calculated by Wickens and Dixon (2007), where automation benefits outweigh the costs of errors. Our results also show that participants using the extreme disuse strategy did not significantly change during the experiment. During non-alarms, extreme use (n=20) was eventually preferred over extreme disuse (n=6). The number of participants using a extreme alarm use and disuse strategy was roughly equal by the last quarter. The differences in alarm and non-alarm response strategies are likely due to the differences in automation's alarm and non-alarm performances. More people adapted an extreme use strategy when the automation was clearly performing well (during non-alarms it was 84% correct) than when it was unclear (during alarms it performed 70% correctly). The shift to extreme strategies appears to vary between alarm and non-alarm states and more research is warranted to explore this further. For extreme alarm responses, there was one fairly large shift of 7 participants to extreme alarm use in the third quarter. We can see a more gradual shift in participants who eventually chose an extreme non-alarm use strategy, with about 5 people changing to extreme non-alarm use in each of the last 3 quarters. Partici-

pants eventually may have found that extreme response strategy removed the burden of using AVI to validate automation's recommendation, allowing participants to maximize their attention resources and performance in dual task scenarios (Bliss, 1993; Du et al., 2020; Wiczorek & Manzey, 2014).

This study has the following limitations: first, no automation likelihood information was provided to participants, each participant was required to estimate how well the automation performed. Second, we only examined the decision-making and response strategies for one automation performance level. Future research studies should manipulate the automation's performance during alarms and non-alarms to examine how operators adjust their strategies.

CONCLUSION

Our study aim was to examine how decision-making and response strategies change as operators increase their experience with imperfect diagnostic automation. Our results show that mean performance, compliance, and reliance significantly increased during the experiment. Individual operators shifted their strategies and the majority eventually resorted to extreme automation use or disuse. Future work should expand to examine how automation performance affects individual decision-making and response strategies. Understanding the individual's strategy supports the development of individualized automation systems and can improve operator training.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 2045009.

REFERENCES

Allendoerfer, K. R., Pai, S., & Friedman-Berg, F. J. (2008). The complexity of signal detection in air traffic control alert situations. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, pp. 54–58).

Bliss, J. P. (1993). The cry-wolf phenomenon and its effect on operator responses. *Unpublished doctoral dissertation, University of Central Florida, Orlando*.

Bliss, J. P. (2003). An investigation of extreme alarm response patterns in laboratory experiments. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 47, pp. 1683–1687).

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38*(11), 2300–2312.

Breznitz, S. (1984). *Cry wolf: The psychology of false alarms*. Psychology Press.

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.

Dorfman, D. D. (1969). Probability matching in signal detection. *Psychonomic Science*, *17*(2), 103–103.

Du, N., Huang, K. Y., & Yang, X. J. (2020). 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. *Human factors*, *62*(6), 987–1001.

Eliasziw, M., & Donner, A. (1991). Application of the mcnemar test to non-independent matched pair data. *Statistics in medicine*, *10*(12), 1981–1991.

Gombolay, M., Yang, X. J., Hayes, B., Seo, N., Liu, Z., Wadhwania, S., ... Shah, J. (2018). Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research*, *37*(10), 1300–1316.

Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (2000). *The application of hierarchical linear modeling to organizational research.* Jossey-Bass.

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

Lu, S., Zhang, M. Y., Ersal, T., & Yang, X. J. (2019). Workload Management in Teleoperation of Unmanned Ground Vehicles: Effects of a Delay Compensation Aid on Human Operators' Workload and Teleoperation Performance. *International Journal of Human-Computer Interaction*, *35*(19), 1820–1830.

Luo, R., Du, N., & Yang, X. J. (2022). Evaluating Effects of Enhanced Autonomy Transparency on Trust, Dependence, and Human-Autonomy Team Performance over Time. *International Journal of Human–Computer Interaction*, *38*(18-20), 1962–1971.

Luo, R., Weng, Y., Wang, Y., Jayakumar, P., Brudnak, M. J., Paul, V., ... Yang, X. J. (2021). A workload adaptive haptic shared control scheme for semi-autonomous driving. *Accident Analysis & Prevention*, *152*, 105968.

Manzey, D., Gérard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, *57*(12), 1833–1855.

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*, *53*(6), 672–686.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human factors*, *43*(4), 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human factors*, *46*(2), 196–204.

Mueller, S. T., & Piper, B. J. (2014). The psychology experiment building language (pebl) and pebl test battery. *Journal of neuroscience methods*, *222*, 250–259.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230–253.

Schuler, P. T., & Yang, X. J. (2022). Human operators' blind compliance, reliance, and dependence behaviors when working with imperfect automation: A meta-analysis. In *2022 ieee 3rd international conference on human-machine systems (ichms)* (pp. 1–6).

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

Wang, Y., & Yang, X. J. (2022). Humans working with un-reliable automation: Reverse psychology versus disuse Model. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 707–710. Retrieved from https://doi.org/10.1177/1071181322661452 (_eprint: https://doi.org/10.1177/1071181322661452) doi: 10.1177/1071181322661452

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.

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* (pp. 408–416).