Degradations of Trust in Automation Associated with Repeated Monitoring Checks

Addison B. Bright, Jenna E. Cotter and Nathan L. Tenhundfeld
Department of Psychology,
The University of Alabama in Huntsville, Huntsville, USA
abb0040@uah.edu, jc0277@uah.edu, nlt0006@uah.edu

***Abstract*— Automated systems, or systems that execute, either partially or fully, a function previously completed by a human operator, have become widely incorporated in society today. As the use of automated systems becomes more prevalent, over-trust and subsequent over-reliance on those systems can occur and users can become complacent in monitoring. To ensure the user is correctly monitoring a system, many companies have incorporated monitoring checks to increase situational awareness (SA) and performance; however, there has not been much research evaluating the consequences of repeated monitoring checks. These checks could have consequences that negatively affect the user's trust, such as the "cry-wolf" effect, because the user is repeatedly reminded of imperfections in the system resulting in disuse of the system. In contrast, these checks could help calibrate trust, increase SA and performance, and promote appropriate use of the system. Understanding these consequences is essential in evaluating a user's trust in the system. To test this, an experiment was designed that explored the impact of monitoring checks and reliability on the level of trust individuals report in the system. In this study, the 60s and 90s frequency of monitoring check conditions had the highest reported trust. However, not all significant results were consistent with previous knowledge of the impacts of reliability and the "cry-wolf" effect. The general pattern of results suggests that there is merit in highly reliable systems especially over intermediately reliable systems. Additionally, the results suggest a complex relationship between trust and the frequency of monitoring checks especially at the intermediate frequency groups.**

Keywords—Automation, Trust, Reliability, Monitoring Checks, Situational Awareness, Autonomous Vehicles.

I. Introduction

In recent decades, automation has become increasingly prevalent with advancements in technology. Automation is defined as a system that executes a task, either partially or fully, that was previously carried out by a human operator [1]. As humans, we interact with automation daily in our homes, our jobs, and in our cars (amongst others). One such example of automation in many individuals' daily lives is the advanced driver assistance system which is used to correct the driver's steering and monitor the lanes the vehicle must remain in [2]

A user's reliance on the system is related to the user's trust in the system's capabilities to complete the task. Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [3]. When exploring trust, three major categories have been defined: dispositional, situational, and learned trust [4]. Dispositional trust is based on an individual's predisposition to trust. Situational trust is dependent on both internal factors, or characteristics of the user, such as the user's self-confidence, expertise, and mood [5] as well as external factors such as task difficulty, workload, and perceived risks and benefits [6], [7]. Finally, learned trust can be split into two categories: initial learned and dynamic learned trust [4]. Initial learned trust represents preexisting knowledge gained prior to any interaction with the system which influences the user's reliance on the system from the start. Dynamic learned trust occurs during the interaction and further influences reliance on the system. As the user interacts with a system, its performance will impact the dynamic learned trust. Because of these factors, it is important to distinguish increasing a user's trust to drive system use, from calibrating the user's trust in the system [8], [9]. Calibrating trust means matching the perceived trustworthiness with the actual trustworthiness of the system ensuring that the user appropriately relies on the system [8]. If the trust in the system is not appropriately calibrated, different consequences can occur.

There are two categories of miscalibrated trust: over-trust and under-trust [8]. Under-trust refers to the user perceiving the trustworthiness of the system as lower than the system's actual capabilities. When a user under-trusts a system they often will disuse or not rely on the system (e.g. disabling alarms), leading to an over-reliance on their own abilities [3], [8], [10]. In contrast, over-trust refers to the user perceiving the trustworthiness of the system to be higher than the system's actual capabilities. When the user over-trusts a system, this can lead to complacency, over-reliance, and degradations in situational awareness (SA).

SA is defined as having three levels: perception of the environment, comprehension of content, and projection of future conditions [11]. At level one, the user is perceiving the status attributes, and dynamics of relevant elements in the environment. At level two the user is understanding the significance of the elements recognized in level one. At level three, the user has the ability to project future actions of the elements in the environment after perceiving and understanding them. High SA is necessary for high performance and safety in complex and simple tasks.

Several factors have been found to affect SA such as diverted attention and task-related and task-unrelated mind wandering [12]–[14]. Over-trust and complacency can cause degradations in SA due to increased support by the automation leading to complacency for the user [15]. As such, the user fails to adequately monitor the situation, which leads to failure to perceive the environment even adequately. The Lumberjack Analogy establishes one consequence of degradations in SA [17]. The analogy stems from the tree-falling notion of "the higher the trees, the harder they fall". When a machine with higher degrees of automation (DOAs) has an automation failure, it is likely that the user is complacent and therefore has diminished SA which leads to more significant consequences than automation failures with lower DOAs. These system failures can either manifest as system errors, or simply the system transferring control back to the user in anticipation of not being able to complete the task.

When automation is implemented in circumstances like self-driving vehicles, humans must remain attentive in order to regain control in the event of a system failure. Maintaining sustained attention in a supervisory control tasks for long periods of time can be difficult and users will often engage in secondary tasks like using a cell phone or eating which leads to further degradation of SA [16]. As such, self-driving vehicle companies have sought to ensure that human drivers are maintaining supervisory control over the vehicle in a number of ways.

These designed and proposed 'attention check' solutions include eye-tracking and behavioral attention checks[17]. Behavioral attention checks alert the user's attention back to the system or require an action from the user to continue the use of the automated features [17]. For example, Tesla has implemented a feature where the vehicle will ask humans to place hands on the steering wheel, while in "Autopilot", if it has not detected their hands for fixed amount of time. If this request is ignored, the request becomes more prevalent, and eventually, the vehicle will pull to the side of the road and needs to be restarted to continue using the self-driving feature. While these attention checks may help to improve SA and performance, there has not been much research regarding the consequences of these monitoring checks. It is important to understand how these monitoring checks impact the user's trust in the system because monitoring checks could draw the user's attention to the system's imperfection. Previous research has shown that repeated false alarms result in the cry-wolf effect which leads to under-trust in the system [18], [19]. It is unclear whether these monitoring checks could create similarly decreased trust as is shown in the cry-wolf effect. It is possible that the implementation of monitoring checks could increase SA and appropriately calibrate trust in the system. In contrast, these attention checks could create under-trust and overall disuse of the system by repeatedly reminding the user that the system is imperfect. Importantly, in the case mentioned above, the prompt does not ask drivers to "pay attention" but rather prompts them to complete a simple task which amounts to simply 'jiggling' the steering wheel. The goal of the current study is to evaluate how these behavioral attention checks could potentially impact trust in the system. This is important to consider for design to promote performance, safety, and satisfaction [20].

In the current study, users were tasked with monitoring automation completing a primary task and correcting automation errors if/when they occurred. In order to mimic the real world where a user engages in other tasks, a secondary task was introduced where the user had to complete the entire task (i.e., without automated help) while monitoring the primary task. In the primary task, the reliability (i.e., the number of errors) differed across individuals. A behavioral monitoring check was implemented at different intervals across individuals to redirect the user's attention to the primary task.

Based on previous literature and the variables manipulated in this study, there were two main hypotheses. First, it was hypothesized that individuals in the high-reliability condition (i.e., 90%) would report the highest levels of trust while those in the intermediate-reliability condition (i.e., 70%) would report lower levels of trust than the high-reliability condition and individuals in the low reliability condition (i.e., 50%) is hypothesized to report the lowest level of trust. Second, it was hypothesized that those in the greater frequency of monitoring checks condition (i.e., 30 seconds between) would report lower levels of trust while those in the intermediate frequency of monitoring checks condition (i.e., 60 seconds between) would report higher levels of trust than the 30 second condition and those in the low frequency of monitoring checks condition (i.e., 90 seconds between) would report the highest level of trust.

II. Methods

Participants. Participants (N = 92) were recruited from an available participant pool at The University of Alabama in Huntsville.

Materials. The simulation the users interacted with was created in Pygame and was presented on a Dell computer running Windows. The simulation environment included four dials with rotating arrows as the primary task, a keypad entry box as the secondary task, and a visual monitoring check displayed above the dials. An image of the simulation as it appeared to the participants is shown in Fig. 1.

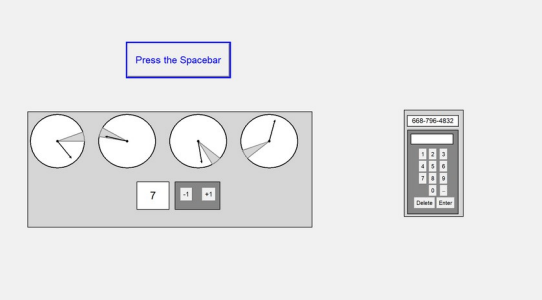


FIGURE 1. Simulation Environment

Additionally, the monitoring checks contained an auditory component, so the participants were provided with headphones. A pre-test survey containing the Automation Induced Complacency Potential Rating (AICP-R) scale assessed the participants' attitudes toward automation on two subscales: alleviating workload and monitoring [21]. This is a 10-question survey graded on a 5-point Likert scale (1 "strongly disagree" to 5 "strongly agree"). Similarly, a post-test survey was given that consisted of the Trust of Automated Systems Test (TOAST). The TOAST has 9 questions that assess the participants' overall trust in the system [22]. The TOAST question is graded on a 7-point Likert scale (1 "strongly disagree" to 7 "strongly agree") and demonstrates high criterion validity between system understanding and performance [22]. These questionnaires and several demographic questions were compiled in Qualtrics.

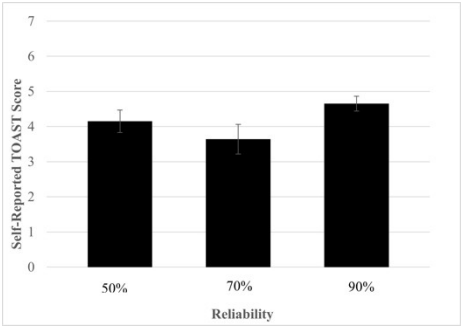
Procedures. To begin the participants were asked to present their government-issued ID to confirm their age and signed a voluntary consent form. Participants filled out the AICP-R questionnaire that was distributed through Qualtrics. This questionnaire assessed the participants' overall attitude toward automation. Next participants were shown the simulation and were told to ensure the count accuracy for the primary task by monitoring the dials as the automation counted. The total for all four dials, contained in one value to the bottom left of all the dials, increased by one every time an arrow landed in the specified range (i.e. the gray wedge in the circle) on any of the dials. For the primary task, the speed and direction of the dials changed every 500ms and the speed was no greater than 24 degrees per second. The total number of times an arrow entered the gray region varied by participant because the speed and direction of the arrow varied. As such, the number of errors differed across participants and reliability conditions. On average, an error occurred every 150 seconds for the 90% reliability conditions, every 50 seconds for the 70% reliability condition, and every 30 seconds for the 50% reliability condition. Errors were categorized as either errors of omission, where the automation failed to count an arrow in the gray region, or errors of commission, where the automation counted when it should not have. These errors occurred randomly and were to be corrected by the participant using the plus button to correct an error of omission and the minus button to correct an error of commission which were displayed next to the counter.

Participants were also asked to enter phone numbers into a keypad as a secondary task. This was a random phone number in which participants copied using the number pad on the left and hit enter to submit. A new phone number appeared every 10 seconds after hitting enter. The monitoring check appeared above the dials and stated, "Press the Spacebar" accompanied by an initial "beep". Initially these checks appeared in solid black but if participants did not dismiss the check by pressing the spacebar, the border would begin to flash in blue. If the check was still not dismissed the alert would "beep" and remain flashing. If the participants did not dismiss the alert before the next alert was queued, the alert disappeared, and the next check was displayed. Monitoring checks occurred at the selected rate based on the condition (i.e., no checks, 30, 60, or 90 seconds). The total simulation ran for 10 minutes. Upon finishing, participants were asked to complete the TOAST questionnaire , which assessed the participant's trust and understanding of the performance of the automated system, and a series of demographics questions. Once the post-test questionnaire was completed, participants were debriefed and dismissed. Participants were in one of twelve conditions based on the reliability (i.e., 50%, 70%, 90%) and the frequency of monitoring checks (i.e., no checks, every 30, 60, or 90 seconds). This study is a subset of a larger study that evaluates a secondary simulation block where all participant's automation is set to a 70% reliability condition with the same varying monitoring checks. Data collected from AICP-R, block 2, and demographics questions were not analyzed for this study. Additionally, performance metrics on the primary and secondary task were not analyzed here.

III. Results

A two-way analysis of variance (ANOVA) was used to determine whether the system reliability, the frequency of monitoring checks, or the interaction between the two impacted the level of reported trust in the automated system. Results indicated that system reliability significantly impacted the level of self-reported trust $F(2, 80) = 4.30$, $p = .02$, $\eta_p^2 = 0.08$ (see Fig. 2).

A Tukey's post hoc test was used to further analyze the differences among the reliability conditions (see Table 1). Individuals in the 70% reliability condition reported significantly less trust in the system than individuals in the 90% reliability condition ($M_{diff} = -0.68$, $t = -2.71$, $p = 0.02$). There was not a significant difference in the level of reported trust between the 50% reliability condition and the other two conditions.



Note: The error bars represent the 95% CI.

FIGURE 2. Average Overall TOAST Scores Across Reliability Conditions

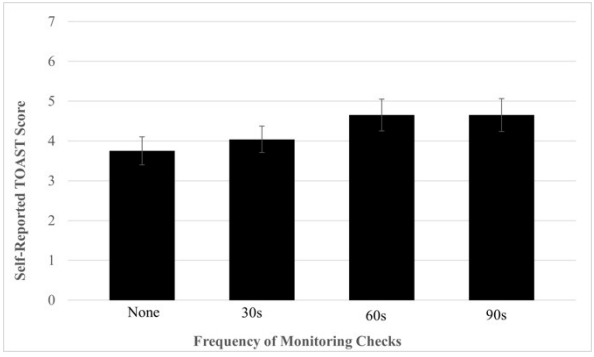
TABLE I. PAIRWISE COMPARISONS FOR TRSUT ACROSS RELIABILITY CONDITIONS

Pairwise Comparison	M_{diff}	*t*	*p*
50% - 70%	0.245	1.01	.58
50% - 90%	-0.44	-2.19	.08
70% - 90%	-0.68	-2.71	.02

The frequency of monitoring checks significantly impacted the level of self-reported trust $F(3, 80) = 5.24$, $p < .01$, $\eta_p^2 = 0.14$ (see Fig. 3).

A Tukey's post hoc test was used to further analyze the differences among the frequency of monitoring checks conditions (see Table 2). Individuals in the no check condition reported significantly less trust in the automated system than those in the 60s ($M_{diff} = -0.86$, $t = -3.25$, $p < 0.01$) and 90s ($M_{diff} = -0.71$, $t = -2.76$, $p = 0.04$) conditions. Individuals in the 30s condition reported significantly less trust than those in the 60s condition ($M_{diff} = -0.77$, $t = -2.72$, $p = 0.04$). All other comparisons were non-significant.

The interaction between reliability and frequency of monitoring checks on reported trust was not statistically significant $F(6, 80) = 1.82$, $p = .10$, $\eta_p^2 = 0.10$.



Note: The error bars represent the 95% CI.

FIGURE 3. AVERAGE OVERALL TOAST SCORES ACROSS MONITORING CHECK CONDITIONS

TABLE II. PAIRWISE COMPARISONS FOR TRUST ACROSS MONITORING CHECK CONDITIONS

Pairwise Comparison	M_{diff}	*t*	*p*
None - 30s	-0.10	-0.41	.98
None - 60s	-0.86	-3.25	< .01
None - 90s	-0.71	-2.76	.04
30s - 60s	-0.77	-2.72	.04
30s - 90s	-0.61	-2.23	.12
60s - 90s	0.16	0.56	.95

IV. DISCUSSION

A. Reliability

We found a significant difference in the level of self-reported trust such that individuals in the 70% reliability condition trusted the system significantly less than those in the 90% condition. There was no difference between the trust reported in the 50% reliability and 70% reliability condition. These results contradict our hypotheses as we believed the 50% reliability condition would create less trust than the 70% reliability condition; however, previous research has shown that the threshold for reliability to create increased confidence in a system is about 80% [23]. It is not entirely clear why the trust reported for the 50% and 90% reliability conditions were not significantly different as we would have expected. However, the level of reported trust varied between the 70 and 90% conditions as would be expected by a dynamically learned trust [4]. The users were unaware of the reliability of the system at the beginning of the simulation; however, as errors occurred the user could gauge the accuracy based on their knowledge and level of attention allotted to the primary task.

B. Frequency of Monitoring Checks

We found a significant difference in the level of reported trust across the frequency of monitoring checks. Individuals in the no-check condition reported significantly less trust than those in the 60s and 90s conditions. It is possible that there was low transparency of the system for individuals who did not receive an attention check because they were not receiving appropriate feedback that the system was correctly executing its task. In oversight of systems with low

transparency, the individual monitoring the system relies on their own capabilities to ensure the accuracy of the primary which further demonstrates imperfections in the system and decreases the level of reported trust.

Furthermore, there were differing levels of trust associated with the different frequencies of attention check conditions. All 3 conditions of monitoring checks increased the level of trust from the no-check condition; however, the 30s condition reported significantly less trust than the 60s condition, while the others did not differ from one another. These results suggest there is a complex relationship between the frequency of monitoring checks and trust. Specifically, considering the 30s to 90s comparison as we would expect a significant difference in these conditions if the cry-wolf effect for false alarms caried over in this situation. The cry-wolf effect has been shown to decrease trust and lead to decreased use of the system and to some extent, this is replicated in the difference between the 30s and 60s conditions; however, if it was fully supported we would expect to see the difference in the 30s to 90s comparison.

When evaluating the implementation of monitoring checks to an automated system, most argue that doing so will promote human attentiveness and increase overall SA [16]. However, it is not known whether there is potential for these monitoring checks to create more harm than good. For example, when we evaluate the 30s monitoring check condition, where there was significantly less trust in the system than the other monitoring check conditions, the higher frequency of checks could have distracted from the overall tasks. By implementing additional distractors, we could have negatively impacted SA and encouraged further task switching. In order to analyze this, a measure of the user's performance in each condition would be useful. In this study, the no-check condition acted as a control in which the user would be monitoring an autonomous system without any sort of monitoring check. As such, these results suggest a need for further research regarding the use of monitoring checks to aid in calibrating trust when a human operator monitors a system. In conclusion, this study demonstrated the fluctuations in user-reported trust across the frequency of monitoring checks and expanded the cry-wolf effect into a monitoring system setting. Additionally, this study is among the first of its kind when evaluating the consequences of monitoring checks on user trust.

C. Limitations and Future Research

In future studies, there are several things to consider when evaluating the consequences of the frequency of monitoring checks. In this study, trust was evaluated using a self-report measure which is inherently subjective and may be impacted by biases from the user and the experimental studies. Further research should evaluate how the user performs as a function of correctly adjusting the errors from the system, which vary across the reliability conditions, as well as the number of phone numbers entered compared to the self-reported trust measure. These measures would demonstrate where users are allotting their attention, objectively, during the study and thus demonstrate whether the user actually trusts the automation. Secondly, the exploration of different types of monitoring checks may be beneficial in further examining the consequences of trust in automation. The current attention checks were displayed above the dials and asked the user to press the spacebar, questions remain as to if these are the most effective monitoring checks for this type of system. Lastly, this study was initially designed to understand monitoring checks in the application to monitoring automated driving so expanding this experimental design into a more realistic driving setting could have different results than the current study.

REFERENCES

[1] T. B. Sheridan and R. Parasuraman, "Human versus automation in responding to failures: An expected-value analysis," *Hum. Factors*, vol. 42, no. 3, pp. 403–407, 2000, doi: 10.1518/001872000779698123.

[2] Natasha Merat and John D. Lee, "Preface to the Special Section on Human Factors and Automation in Vehicles: Designing Highly Automated Vehicles With the Driver in Mind," *Hum. Factors*, vol. 54, no. 5, pp. 681–686, 2012.

[3] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.

[4] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

[5] "de Vries, Midden, Bouwhuis - 2003 - The effects of errors on system trust, self-confidence, and the allocation of control in route plann.pdf."

[6] N. R. Bailey and M. W. Scerbo, "Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust," *Theor. Issues Ergon. Sci.*, vol. 8, no. 4, pp. 321–348, 2007, doi: 10.1080/14639220500535301.

[7] R. Molloy and R. Parasuraman, "Monitoring an automated system for a single failure: Vigilance and task complexity effects," *Hum. Factors*, vol. 38, no. 2, pp. 311–322, Jun. 1996, doi: 10.1518/001872096779048093.

[8] E. de Visser *et al.*, "Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams," *Int. J. Soc. Robot.*, vol. 12, no. 2, pp. 459–478, May 2020, doi: 10.1007/s12369-019-00596-x.

[9] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," *Hum. Factors*, vol. 57, no. 4, pp. 545–556, 2015, doi: 10.1177/0018720814564422.

[10] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997, doi: 10.1518/001872097778543886.

[11] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Hum. Factors*, vol. 37, no. 1, pp. 32–64, 1995, doi: 10.1518/001872095779049543.

[12] S. M. Casner and J. W. Schooler, "Thoughts in flight: Automation use and pilots' task-related and task-unrelated thought," *Hum. Factors*, vol. 56, no. 3, pp. 433–442, 2014, doi: 10.1177/0018720813501550.

[13] B. Kidwell, G. L. Calhoun, H. A. Ruff, and R. Parasuraman, "Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles," *Proc. Hum. Factors Ergon. Soc.*, no. 2007, pp. 428–432, 2012, doi: 10.1177/1071181312561096.

[14] K. Kilingaru, J. W. Tweedale, S. Thatcher, and L. C. Jain, "Monitoring pilot 'Situation Awareness,'" *J. Intell. Fuzzy Syst.*, vol. 24, no. 3, pp. 457–466, Jan. 2013, doi: 10.3233/IFS-2012-0566.

[15] L. Onnasch, C. D. Wickens, H. Li, and D. H. Manzey, "Human performance consequences of stages and levels of automation: An integrated meta-analysis," *Hum. Factors*, vol. 56, no. 3, pp. 476–488, 2014, doi: 10.1177/0018720813501549.

[16] F. Naujoks, C. Purucker, and A. Neukum, “Secondary task engagement and vehicle automation - Comparing the effects of different automation levels in an on-road experiment,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 38, pp. 67–82, 2016, doi: 10.1016/j.trf.2016.01.011.

[17] S. Said, S. AlKork, T. Beyrouthy, M. Hassan, O. E. Abdellatif, and M. Fayek Abdraboo, “Real time eye tracking and detection- A driving assistance system,” *Adv. Sci. Technol. Eng. Syst.*, vol. 3, no. 6, pp. 446–454, 2018, doi: 10.25046/aj030653.

[18] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, “Automation failures on tasks easily performed by operators undermine trust in automated aids,” *Hum. Factors*, vol. 48, no. 2, pp. 241–256, Jun. 2006, doi: 10.1518/001872006777724408.

[19] C. D. Wickens, B. L. Hooey, B. F. Gore, A. Sebok, and C. S. Koenicke, “Identifying black swans in nextgen: Predicting human performance in off-nominal conditions,” *Hum. Factors*, vol. 51, no. 5, pp. 638–651, Nov. 2009, doi: 10.1177/0018720809349709.

[20] J. D. Lee, C. D. Wickens, Y. Liu, and L. N. Boyle, *Designing for People*. Charleston, SC, SC, 2017.

[21] S. M. Merritt *et al.*, “Automation-induced complacency potential: Development and validation of a new scale,” *Front. Psychol.*, vol. 10, no. FEB, 2019, doi: 10.3389/fpsyg.2019.00225.

[22] H. M. Wojton, S. Lane, D. Porter, S. T. Lane, C. Bieber, and P. Madhavan, “Initial validation of the trust of automated systems test (TOAST),” *J. Soc. Psychol.*, vol. 160, no. 6, pp. 735–750, 2020, doi: 10.1080/00224545.2020.1749020.

[23] D. Huegli, S. Merks, and Schwaniger, Adrian, “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,” 2020.