Near-Perfect Automation: Investigating Performance, Trust, and Visual Attention Allocation

Cyrus K. Foroughi, U.S. Naval Research Laboratory, Washington, DC, USA, **Shannon Devlin**, U.S. Naval Research Laboratory, Washington, DC, USA, and University of Virginia, Charlottesville, USA, **Richard Pak**, Clemson University, South Carolina, USA, **Noelle L. Brown**, **Ciara Sibley**, and **Joseph T. Coyne**, U.S. Naval Research Laboratory, Washington, DC, USA

Objective: Assess performance, trust, and visual attention during the monitoring of a near-perfect automated system.

Background: Research rarely attempts to assess performance, trust, and visual attention in near-perfect automated systems even though they will be relied on in high-stakes environments.

Methods: Seventy-three participants completed a 40-min supervisory control task where they monitored three search feeds. All search feeds were 100% reliable with the exception of two automation failures: one miss and one false alarm. Eye-tracking and subjective trust data were collected.

Results: Thirty-four percent of participants correctly identified the automation miss, and 67% correctly identified the automation false alarm. Subjective trust increased when participants did not detect the automation failures and decreased when they did. Participants who detected the false alarm had a more complex scan pattern in the 2 min centered around the automation failure compared with those who did not. Additionally, those who detected the failures had longer dwell times in and transitioned to the center sensor feed significantly more often.

Conclusion: Not only does this work highlight the limitations of the human when monitoring near-perfect automated systems, it begins to quantify the subjective experience and attentional cost of the human. It further emphasizes the need to (1) reevaluate the role of the operator in future high-stakes environments and (2) understand the human on an individual level and actively design for the given individual when working with near-perfect automated systems.

Application: Multiple operator-level measures should be collected in real-time in order to monitor an operator's state and leverage real-time, individualized assistance.

Keywords: automation, performance, trust, visual attention

Address correspondence to Cyrus K. Foroughi, U.S. Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375, USA; e-mail: cyrus.foroughi@nrl.navy.mil

HUMAN FACTORS
Vol. 00, No. 0, Month XXXX, pp. 1-16
DOI:10.1177/00187208211032889
Article reuse guidelines: sagepub.com/journals-permissions

Over the past decade, the number of deployed automated technologies has sharply increased. We are quickly moving to a new phase of human–automation interaction where humans may be monitoring near-perfect automated systems. In many cases, these systems will be 99.9% reliable or higher (e.g., United States Department of Defense, 2017). However, humans are still likely to be tasked to intervene when it does fail, and those failures are projected to be costlier than before (Onnasch et al., 2014). Although there is an immediate, real-world need to understand how humans interact with these near-perfect automated systems, there is a dearth of research. Here, our goal was to holistically assess performance, trust, and visual attention during the monitoring of a near-perfect automated system that fails .1% of the time.

HUMAN PERFORMANCE AND AUTOMATION

Introducing automation to offset a human's limitations in attention and reduce manpower hours seems primarily advantageous. However, research has clearly shown that trade-offs exist when introducing automation such as the loss of situation awareness (Endsley & Kiris, 1995), manual skill (Bainbridge, 1983), and overall system trust (Hoff & Bashir, 2015). Researchers have used many terms to describe these trade-offs: for example, "automation conundrum" (Endsley, 2017) and "irony of automation" (Bainbridge, 1983). Endsley (2017) describes the problem well: "The more automation is added to a system, and the more reliable and robust that automation is, the less likely that

human operators overseeing the automation will be aware of critical information and able to take over manual control when needed."

Automated systems of the future will likely have near-perfect reliability (e.g., 99.9%), and it is unlikely that humans will be able to reliably detect these rare-event failures and, even less likely, be able to then step in to correct said failures. Very little research has evaluated how well humans can detect rare-event automation failures in these near-perfect automated systems. A bulk of the previous research has evaluated how well humans detect failures with automation ranging from 60% to 90% reliability (e.g., Chancey et al., 2017; Dixon & Wickens, 2006; Dixon et al., 2007; Foroughi et al., 2019; Rovira et al., 2007). Some researchers have found human performance increases as automation reliability increases (e.g., Chancey et al., 2017), while others have found that human performance improves when interacting with a varied reliability automated system as opposed to a consistently reliable system (Parasuraman et al., 1993). Recently, our group showed that although the combined human–automation accuracy increased as automation reliability increased, the contribution from the human's detecting of automation failures (specifically, when it missed a target) remained relatively stable as the automation reliability increased. That is, the contribution from the human remained mostly consistent as the reliability of the automated system increased (Foroughi et al., 2019).

We do not expect humans to perform well in a task where automation failures are extremely rare (e.g., vigilance task; see Parasuraman, 1986; Warm et al., 2008). However, establishing a specific point estimate that can be considered "poor" performance at the onset of the experiment is challenging. In realistic terms, unless humans are able to detect these rare-event failures at a high rate, their role as an automation monitor may not be worthwhile. With that being said, including additional measures such as subjective trust ratings and attention allocation as a function of whether someone detected the automation failures helps in holistically understanding the human's role when monitoring these systems. Including these analyses could inform how to achieve a more effective human–automation interaction and subsequently improve technology design or training practices. For example, studying a human's trust calibration process of automation has led to an increased understanding of human–automation interactions as they happen in real time.

TRUST AND AUTOMATION

Trust is a human's attitude that another entity (e.g., human, machine, system) will help achieve one's goals in the face of uncertainty and vulnerability (Lee & See, 2004). A human's behavior with a system can be dramatically affected by their level of trust (Muir, 1994; Muir & Moray, 1996). For example, very high trust in an automated system can lead to a person over-relying (not enough monitoring) or over-complying (blind acceptance), even if it is unreliable, a state known as complacency (Parasuraman & Riley, 1997). On the other hand, under-trust leads to humans shunning automation and suffering the negative effects of manual performance in intensive situations (e.g., experiencing mental overload or catastrophic performance outcomes). This relationship between reliability and trust is sometimes referred to as "trust calibration" (Lewandowsky et al., 2000; Parasuraman & Riley, 1997). Calibration is a continuous process as it updates and evolves with the present situation.

Because of the important role that trust plays in human–automation performance (Lee & Moray, 1994), a great deal of research has sought to examine what affects trust and how it affects performance (Hoff & Bashir, 2015; Lee & See, 2004). One of the most well studied factors is the reliability, or perceived reliability of the system. In an early investigation of how trust is influenced by system characteristics, Lee and Moray (1994) found that human trust in a system could simply be predicted by, among other things, the level of reliability of the system itself. However, research has found that trust is lost faster than it is regained (Wiegmann et al., 2001). Additionally, humans have been found to narrow their attentional resources on the area of automation where it did fail, leading to decreased surveillance of the rest of the system (Dixon & Wickens, 2006; Thomas & Wickens, 2004). While the coupling of human

trust to system reliability has been confirmed in subsequent investigations (e.g., Hancock et al., 2011), the nature of the error has an important differential effect on trust (Meyer & Ballas, 1997). For example, when an automated system fails to detect a signal, it is often called a *miss*. Alternatively, when it indicates it has detected a signal, but in reality it has not, it is often called a *false alarm*. Although both are failures needing correction, a miss is not as salient as a false alarm, and misses been found to have less severe effects on trust (Davenport & Bustamante, 2010).

However, with regard to near-perfect automation, there are two related remaining questions regarding performance and the dynamics of trust. First, consistent with prior research, we expect humans to have high levels of trust with exposure to more reliable systems. Does this high level of trust result in complacent behaviors, such as marked, quantitative decrease in attentional narrowing, and contribute to reduced performance in detecting extremely rare failures? The second question is in regard to the dynamics of trust: how is trust affected by extremely rare failures? It could be argued that extremely rare failures are more memorable and could result in more extreme trust dynamics than with moderate reliability automation. However, trust may be able to be adequately rebuilt due to the system's overwhelming reliability the majority of the time, which has been a challenge with moderately reliable systems. These questions can be assessed via measuring trust levels over several different time points, and can be especially informative when they come after a rare-event automation failure. However, another potential way to quantify and further understand the impact that these extremely rare failures have on the operator is by studying their real-time attention allocation.

ATTENTION ALLOCATION AND AUTOMATION

Attentional resources have been defined as both the fuel and bottleneck of the human information processing system (Wickens et al., 2012). Selective attention is the process of directing attentional resources to entities in an environment for information extraction. Eye tracking has been successful in capturing both overt and covert selective attention in real-time (Liechty et al., 2003). Nevertheless, the former has been more applicable in human factors research as it contributes to understanding several constructs, such as workload (Coral, 2016), expertise (Jarodzka et al., 2010), and learning (Rehder & Hoffman, 2003). It has also been able to directly quantify how the human monitors automation across many contexts. For the purposes of the present work, eye tracking is used to quantify overt selective attention, which we term as *visual attention allocation* henceforth.

Previous research studying human–automation interaction has relied on a variety of different eye-tracking metrics (Bagheri & Jamieson, 2004; Dehais et al., 2015; Sarter et al., 2007; Thomas & Wickens, 2004). Recently, research has specifically investigated how eye tracking can be used as an objective measure of operator's trust of an automated system (Glaholt, 2014; Hergeth et al., 2016; Parasuraman & Manzey, 2010; Victor et al., 2018). Research generally supports monitoring frequency to be inversely related to human trust—meaning the more the human trusts the automation, the less frequently it will be monitored (Bagheri & Jamieson, 2004; Brown & Noy, 2004; Hergeth et al., 2016; Moray & Inagaki, 1999). Hergeth et al. (2016) found this to be evident, but investigated if it was primarily due to a decrease in monitoring in general. They compared the total amount of time monitoring the automated tasks to the total amount of time monitoring all the other nonautomated tasks. This ratio measure was positively correlated, meaning changes in monitoring the automation was not solely due to changes in monitoring in general. Hergeth et al. (2016) suggest future research should continue to study monitoring ratios when humans have more "decisional freedom," for example, when they are not explicitly instructed on how to attend to tasks and to investigate if other eye-tracking metrics can be sensitive and reliable measures of trust. When research expands to additional eye-tracking metrics, it usually captures the static and aggregate patterns of visual attention, and sometimes only in reference to a specific area of interest (AOI). For example, Victor et al. (2018) found that in a simulated autonomous vehicle environment, the percentage of glances on the road was not able to predict a human's ability to intervene in a timely

and appropriate manner when the autonomous vehicle failed. However, the human–automation relationship may be further informed by measuring visual attention allocation with metrics that are composites of basic metrics. For example, when applied to gaze data, entropy quantifies the predictability of the observed pattern of fixation transitions (Shiferaw et al., 2019). Entropy measures have been found to be very informative on the measurement of several human factors constructs such as information processing, task difficulty, individual differences, and sleep deprivation (Kruizinga et al., 2006; Raptis et al., 2017; Shic et al., 2008; Shiferaw et al., 2018).

One of the goals for this research was to understand how visual attention allocation differed between those who did and did not detect rare-event automation failures as a means to better understand the human's behavior toward near-perfect automation. In this work, a subset of eye-tracking metrics that were informative of global (i.e., general) and local (i.e., specific) visual attention allocation patterns was studied. Both the global and local metrics were calculated for the entire experimental session and for a specified time window during each automation failure. Given that the overarching goal of the present work was to observe any differences in visual attention allocation patterns between participants who did and did not detect each automation failure, both types of metrics and analyses were included. Finally, eye-tracking metrics that are found to be consistently different between performance groups will be analyzed in the same method as the trust ratings (i.e., by detection rate and over time) in order to make comparisons between the two results.

CURRENT STUDY MOTIVATION AND GOALS

The current study was motivated by the immediate need to understand how humans interact with near-perfect automated systems by assessing three information streams essential to successfully monitoring and detecting rare-event automation failures: performance, trust, and visual attention allocation. To do this, we deployed the Supervisory Control Operations User Testbed (SCOUT), an automated supervisory control environment (Figure 1) that was designed to simulate the current and future demands of unmanned aerial vehicle (UAV)

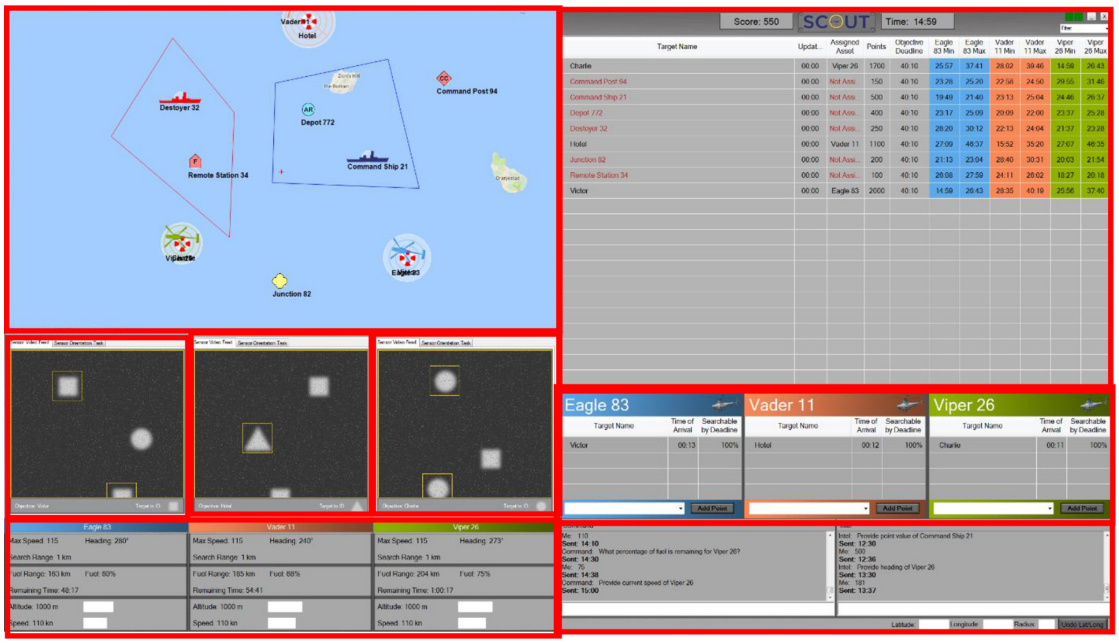

Figure 1. The supervisory control user testbed (SCOUT).

pilots (Sibley et al., 2016). The reliability of this environment was 99.9%, meaning participants encountered only two automation failures while completing the experiment. Subjective trust questions were asked throughout the experiment, and eye-tracking data were collected as a real-time index of visual attention allocation. Our goals were to (1) determine how well participants could detect the rare-event automation failures, (2) determine how subjective trust changes as a function of detection rates, and (3) determine the relation between visual attention allocation and detection type.

METHOD

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Boards at both the U.S. Naval Research Laboratory and George Mason University. Informed consent was obtained from each participant.

Participants

Seventy-three students with normal or corrected-to-normal vision (*M* age = 20.5 years, SD age = 4.2 years, 51 females) from George Mason University participated in this research for course credit.

Tasks

The Supervisory Control Operations User Testbed (SCOUT) is a simulated supervisory control environment (Figure 1) designed by scientists at the U.S. Naval Research Laboratory (Sibley et al., 2016) to simulate the current and future demands of UAV pilots. This testbed requires individuals to plan a search mission using three UAVs, then monitor those UAVs while completing secondary tasks. Some of these tasks include responding to chat updates from command (e.g., confirming flight status or relaying intelligence) and updating UAV information (e.g., updating flight speed or altitude). SCOUT includes many self-report probes including trust, fatigue, and workload.

Importantly, when a UAV reaches its target, the sensor search feed for that UAV becomes active, and the user must monitor the search feed to identify possible targets. The search feed is automated such that the system will help the user identify targets by highlighting possible targets with a gold box (Figure 2). This automation is immediately displayed with no delay. Each sensor search feed had a different target shape—either a triangle, circle, or square (see Target ID in Figure 2), meaning all other shapes for that feed were defined as distractors. All objects would enter at the top of the feed and then vertically scroll down it for 14 s. In that time, the automation was tasked to highlight each target with a gold box. For example, if a sensor search feed's target was a triangle, the participant would need to ensure that all of the triangles (i.e., potential targets) that scrolled across the screen were highlighted, and none of the circles or squares (i.e., distractor targets) were not highlighted. The state of any object (i.e., highlighted or not highlighted) could be changed by clicking on that object. Each search feed had a different target resulting in participants searching for triangles in one feed, circles in another feed, and squares in the third feed. See https://youtu.be/HehmW_Ha-9M for a demonstration of SCOUT.

Equipment

A 24-inch Dell P2415Q monitor set at 2560 × 1440 resolution was used for this experiment. The participants used a standard mouse and QWERTY keyboard to complete the task. For the eye-tracking data collection, a Gazepoint GP3 eye tracker with a sampling rate of 60 Hz and 0.5–1 degree of visual accuracy was used and placed right below the monitor. Participants sat approximately 65 cm (25.6 in) from the monitor. The eye tracker was calibrated for each participant using a 9-point calibration program built by Gazepoint. The GP3 provides left and right eye point of gaze in pixels and assigns a binary quality measure to each point to indicate whether the system believes the data is valid or not. Based on these three values, each data point was marked as either valid or invalid for analysis. A valid data point was one where its quality measure was maximized and both the left and right point of gaze was a positive coordinate value. Only valid data points were used for eye-tracking metric calculations.

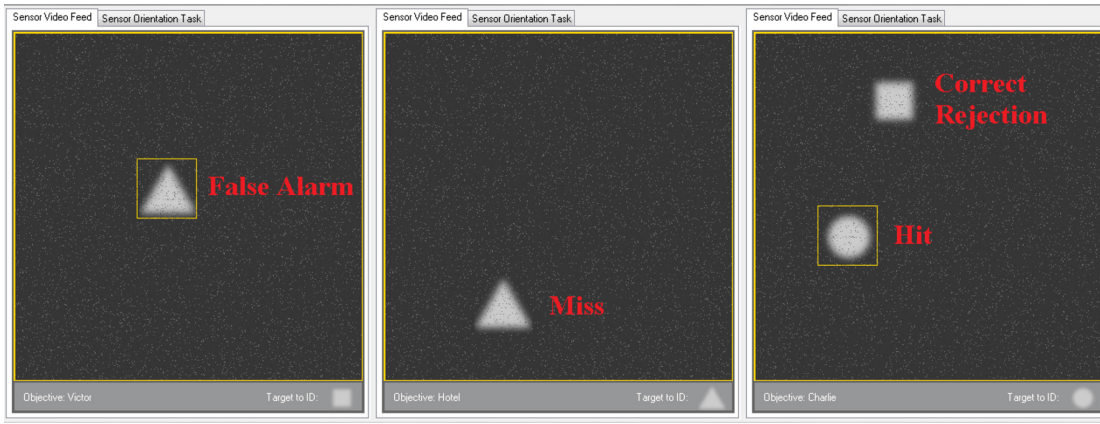

Figure 2. This is an example of the sensor search feeds from SCOUT. There is an icon below each sensor search feed indicating the target of interest: square, triangle, and circle from left to right respectively, as noted by the red arrows. The automated system automatically highlights targets by placing a gold box around them. Participants were tasked to ensure that the automated system accurately identifies the correct targets. If the automated system misses a correct target (miss) or incorrectly highlights the wrong target (false alarm), participants must click on the object to fix the error. In this specific example, we have shown all four possible outcomes of what the automated system could do. The red labels are added for this figure and are not in the experiment. To correct the automation failure (i.e., miss or false alarm), participants would need to click on the shape to either select or deselect it appropriately.

Procedure

After signing an informed consent form, participants were instructed to be comfortably seated in the desk chair where the experiment would take place. First, the participant calibrated to the eye tracker using the Gazepoint GP3 software. Next, participants completed a fixation test as an additional calibration tool. Participants then completed a luminance change task and the shortened automated operation span. These tasks were not analyzed for this manuscript, as both are part of a larger individual differences project that is not yet complete (see Rovira et al., 2017 for more information regarding individual differences and automation).

Participants then completed a SCOUT training session to learn how to properly complete the task. During training, participants were informed that the automation may not be perfect and that they would need to ensure that all targets were correctly identified. After completion, participants were given a short comprehension test about SCOUT to ensure that they understood all of the features of the task. Participants were shown a static screenshot of SCOUT and asked to answer questions about features within the task (e.g., Can you tell me the current speed of Vader 11? How many targets are in Vader 11's sensor feed?"). Participants were required to answer every question correctly to continue. All participants answered all the comprehension questions correctly on their first attempt.

Participants then completed a 40-min experimental scenario within SCOUT. For this experiment, all three UAVs had preset targets and no participants deviated the UAVs from their targets. All three search feeds activated within 1 s of each other ensuring near equal display time. Participants had 14 s to decide if any object was incorrectly highlighted or not highlighted, and to correct the object accordingly. Objects appeared at a rate of 1 every 5s, on average for each sensor search feed. Objects could be on multiple search feeds at once. Chat queries (e.g., What percentage of fuel is remaining for Eagle 83?) occurred every 60 s on the lower right side of SCOUT. These events did not coincide with the manually

injected automation errors (mentioned below) to avoid split attention.

For this experiment, with the exception of the two manually injected automation failures, the automation reliability was set to 100%. These manually injected automation failures (namely, one automation miss and one automation false alarm) occurred at approximately 19:05 and 39:05 in the center sensor feed (for concerns about the impact of center-bias, see Supplemental Material). Participants were not given specifics on which sensor feed an automation failure could occur. The types of failures (i.e., miss and false alarm) were counterbalanced. These two automation failures made the overall automation reliability of the system 99.9% across the entire experiment. Additionally, participants were prompted with a trust question at four time points: approximately 10:15, 19:25, 30:15, and 39:25. They were specifically asked "To what extent do you trust (i.e., believe in the accuracy of) the automation aid in this scenario?" and were able to respond using a sliding scale from "Not at all" to "Completely." After completing the SCOUT scenario, participants completed a short demographics survey.

Metrics

Performance. Each participant encountered two automation failures: one miss and one false alarm. Individually, each participant could have detected no failures, one failure, or both failures. We calculated the percentage of automation failures detected by each type across all participants.

Subjective trust. Each participant responded to the trust question four times using a sliding scale that ranged from "Not at all" to "Completely." We mapped these responses to a 0–100 scale for analyses.

Eye tracking. Details of the eye-tracking metrics are in the supplemental materials. Table 1 highlights the chosen metrics.

RESULTS

All analyses were screened for outliers and violations in normality. Outliers were considered to be anything beyond 1.5x of the interquartile range (IQR). If outliers were detected, they were removed from the dataset and if normality was not met, corresponding nonparametric tests (e.g., Mann–Whitney) were used. The selected significance level was $\alpha = .05$. For omnibus tests, partial eta squared (η_p^2) is reported for effect size, where the values of .01, .06, .14 are interpreted as small, medium, and large effect size, respectively (Cohen, 1988). For tests of means, effect size is reported by using Cohen's d and values of 0.2, 0.5, 0.8, which indicate a small, medium, and large effect size, respectively (Cohen, 1988).

Performance

Overall, 34% (25 of 73) of the participants correctly identified the automation miss and 67% (49 of 73) correctly identified the automation false alarms. As for the distribution of participants detecting failures in general, 18 did not detect any failure, 36 detected one failure (i.e., the first *or* second failure), and 19 detected both failures. To summarize how automation failure type and timing impacted performance, when the first failure was an automation miss (37 of 73 participants), 15 participants detected no failure across the entire experiment, nine participants detected both failures, four participants detected only the first failure (miss), and nine participants detected only the second failure (false alarm). When the first failure was an automation false alarm (36 of 73 participants), three participants detected no failure across the entire experiment, ten participants detected both failures, 21 participants detected only the first failure (false alarm), and two participants detected only the second failure (miss). To summarize, the data show that the main driver of performance was the type of automation failure (i.e., miss or false alarm) as opposed to the timing of the failure.

Trust

Figure 3 presents the changes of subjective trust ratings as a function of time and detection performance. Using the lme4 package within R (Bates et al., 2014), we ran a mixed-effects model with time and detection rate as predictors (fixed effects), and subjective trust rating as the outcome variable. Time is a within-subject

TABLE 1: Candidate Eye-Tracking Metrics

Eye-Tracking Metric	Formula/Definition	Explanation/Application	Source
Capturing/measuring global visual attention metrics (i.e., how visual attention allocation is in general)			
Mean time between fixations (MTBF)	The average time elapsed between two fixations.	Measures how quickly a person is scanning and processing stimuli.	Bagheri and Jamieson (2004)
Normalized gaze transition entropy (GTE)	$H(x) = -\sum_{i=1}^{n} p_i \sum_{j=1}^{n} p(i,j) \log_2 p(i,j)$ p_i is the overall probability of a fixation being in state *i*, p_{ij} is the probability of transitioning from state *i* to state *j*, *S* is the set of states in the system, and *s* is the number of states in the system	The higher the value, the more random the fixation transitions.	Ellis and Stark (1986); Shiferaw et al. (2019)
Normalized stationary gaze entropy (SGE)	$H(x) = -\sum_{i=1}^{n} (p_i) \log_2 (p_i)$ where p_i is the total number of fixations in state *i*, *S* is the set of states in the system, and *s* is the number of states in the system	The higher the value, the more distributed the fixations were across the AOIs of the display.	Shiferaw et al. (2019)
Local visual attention allocation metrics (i.e., how it relates to a specific AOI)			
Total dwell ratio	Compare the total amount of time spent in a specific AOI in comparison to all other AOIs. In this work, the specific AOI is the one where the automation failures occurred (i.e., the center sensor feed).	Used to understand how time spent in an AOI changed specifically versus generally. Previously used in automation failure research as a proxy measure for trust.	Hergeth et al. (2016)
Number of transitions	The number of fixations transitioning to a specific AOI. In this work, the specific AOI is the one where the automation failures occurred (i.e., the center sensor feed).	Previously used to evaluate design, expertise, and importance of an area. In HAI research, used to measure automation trust, with higher values indicating lower levels of trust.	Hergeth et al. (2016); Holmqvist et al. (2011); Parasuraman and Manzey (2010)

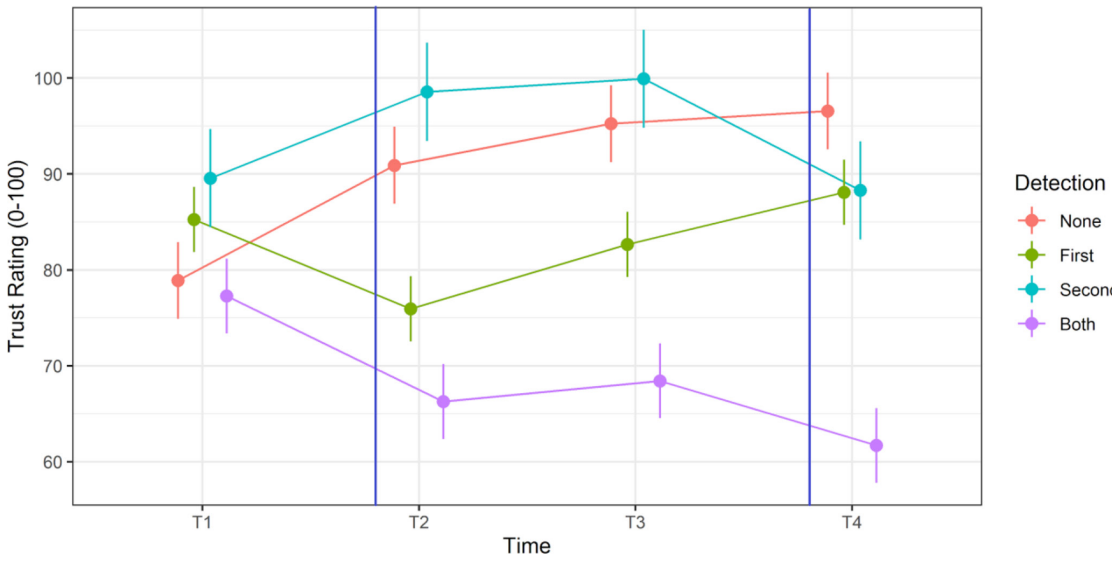

Figure 3. Mean subjective trust ratings (±SE) as a function of time and detection. The vertical blue lines provide an approximate visual representation of when the automation failures occurred. Time is a within-subject factor with four levels: T1 (before a failure occurred), T2 (after one failure occurred), T3 (10 min after the first failure occurred), and T4 (after the second failure occurred). Detection is a between-subject factor with four levels: None (did not detect either failure), First (detected the first failure only), Second (detected the second failure only), and Both (detected both failures).

factor with four levels: Time 1 (before a failure occurred), Time 2 (after one failure occurred), Time 3 (10 min after the first failure occurred), and Time 4 (after the second failure occurred). Detection rate is a between-subject factor with four levels: None (did not detect either failure), First (detected the first failure only), Second (detected the second failure only), and Both (detected both failures). We found a significant main effect of time ($F(3,69) = 8.5$, $p < .001$), a significant main effect of detection ($F(3,207) = 4.1$, $p = .007$), and a significant interaction ($F(9,207) = 18.6$, $p < .001$). We were interested in determining whether subjective trust ratings changed following the detection of a failure. This resulted in running different analyses by detection group. Consistent with the first failure effect (Wickens & Xu, 2002), for the group that detected the first failure, a paired contrast between T1 and T2 revealed a significant decrease between subjective trust scores ($M_{DIFF} = -6.72$, SE = 2.01, $p = .001$, $d = .50$). For the group that detected the second failure, a paired contrast between T3 and T4 revealed a significant decrease between subjective trust scores ($M_{DIFF} = -11.64$, SE = 3.04, $p < .001$, d = 2.19). To further explore the simple main effect of time, we ran a one-way repeated-measures ANOVA for the group that did not detect any failure and one for the group that detected both failures. For the group that detected no failures, a one-way repeated measures ANOVA revealed an effect of Time ($F(3,51) = 12.41$, $p < .001$, $\eta_p^2 = .422$) such that subjective trust rating increased over time. For the group that detected both failures, a one-way repeated-measures ANOVA revealed an effect of Time ($F(3,54) = 14.46$, $p < .001$, $\eta_p^2 = .446$) such that subjective trust rating decreased over time.

Eye Tracking

Overall eye-tracking analysis. Mean time between fixations (MTBF) was not significantly different for those who did or did not detect the miss or for those who did and did not detect the

false alarm (all $p > .05$). This result suggests the speed of visual attention allocation was not significantly different between performance groups. Normalized gaze transition (GTE) and stationary gaze entropy (SGE) were not significantly different between those who did and did not detect the miss or between those who did and did not detect the false alarm (all $p > .05$).

The local metrics, that is, the ones focused on the center sensor feed as that was the specific AOI associated with the automation failure, were then calculated. For total dwell ratio of the center sensor feed, those who detected the miss had significantly higher total dwell ratio (M = .31, SD = .09) than those who missed the miss (M = .19, SD = .08; ($t(33.027) = -4.3532$, $p < .001$, $d = 1.442$). Those who detected the false alarm (M = .263, SD = .09) had a significantly higher total dwell ratio than those who missed the false alarm (M = .183, SD = .09; $t(15.371) = -2.3698$, $p = .031$, $d = .8965$). This suggests that those who detected the miss and false alarm had a significantly higher proportion of time in the center sensor feed than those who did not detect the miss and false alarm. For the number of transitions to the center sensor feed, those who detected the miss had significantly more transitions, to this feed (M = 2185, SD = 722.3) than those who did not detect the miss (M = 1413.8, SD = 629.1; $t(35.719) = -3.548$, $p = .001$, $d = 1.141$). Similarly, those who detected the false alarm had significantly more transitions to the center sensor feed, (M = 1938.8, SD = 738.2) than those who did not detect the false alarm (M = 1356.4, SD = 739.83; $t(15.647) = -2.148$, $p = .047$, $d = .7881$). This suggests that those who detected the miss and false alarm transitioned to the center sensor feed more frequently than those who did not detect the miss and false alarm.

Two-minute window centered around each automation failure. For MTBF, there was no significant difference between those who did and did not detect the miss or false alarm ($p > .05$). There was no significant difference in normalized GTE for those who did and did not detect the miss ($p > .05$). However, there was a significant difference between those who did (M = .41, SD = .064) and did not detect the false alarm (M = .35, SD = .049; $t(13.634) = -2.366$, $p = .0334$, $d = .9815$), suggesting the 2-min scan sequence of those who detected the false alarm was more complex than those who did not detect the false alarm. There was no significant difference in normalized SGE for those who did and did not detect the miss or false alarm ($p > .05$).

As for metrics focused specifically on the center sensor feed, that is, the one experiencing the automation failure, those who detected the miss had significantly higher total dwell ratio (M = .37, SD = .09) than those who missed the miss (M = .21, SD = .07; $U = 10$, $p < .001$, $d = 1.906$). Those who detected the false alarm (M = .38, SD = .15) also had a significantly higher total dwell ratio than those who did not detect the false alarm (M = .14, SD = .07, $t(18.247) = -5.2144$, $p < .001$, $d = 2.096$). So for the minute before and after the automation failure, those who detected the miss and false alarm spent a significantly higher proportion of time in the center senor feed (relative to all other AOIs), than those who did not detect the miss and false alarm. For the number of transitions to the center sensor feed it was found, again, that those who detected the miss had significantly more transitions, to the center sensor feed (M = 114.9, SD = 31.1) than those who did not detect the miss (M = 67.33, SD = 23.7, $U = 17$, $p < .001$, $d = 1.733$). Similarly, those who detected the false alarm had significantly more transitions to the center sensor feed, (M = 107.9, SD = 42.1) than those who did not detect the false alarm (M = 43.8, SD = 4.65, $t(20.67) = -6.647$, $p < .001$, $d = 2.741$). This suggests that for the minute before and after the automation failure, those who detected the miss and false alarm transitioned to the center sensor feed significantly more frequently than those who missed the miss and false alarm.

Over time analysis for the local eye-tracking metrics as a function of detection. In an attempt to robustly address our third research goal (i.e., examine the relation between visual attention allocation and detection type), we also analyzed the local eye-tracking metrics in the same format as the trust ratings (as a function of detection rate and over time). We limited this analysis to the local eye-tracking metrics only given the consistent significant differences

found with these metrics between those who do and do not detect each type of automation failure. A two-way mixed ANOVA where the between-subject effect was the four performance groups (no detection, only first failure detected, only second failure detected, detected both failures) and within-subject effect was time period (i.e., the durations of T1–T4) was used for both local eye-tracking metrics. For dwell ratio, there was a main effect of performance group ($F(3,35) = 6.87$, $p < .001$) but not time ($F(3,105) = 2.355$, $p = .076$) nor their interaction ($F(9,105) = 1.889$, $p = .061$). For number of transitions, there was a main effect of performance group ($F(3,33) = 4.708$, $p = .08$) and time ($F(3,99) = 225.904$, $p < .001$) but no significant interaction ($F(9,99) = 1.810$, $p = .076$). Therefore, individuals may not update their visual attention strategies even when they detect errors in near-perfect automation which is in stark contrast to the trends found with the trust ratings. Table 2 summarizes the findings from the eye-tracking analyses.

DISCUSSION

The goal for this research was to improve our understanding of how humans interact with near-perfect automated systems by assessing three important human-automation interaction features: performance, trust, and attention

TABLE 2: Summary of the Eye-Tracking Analysis

Eye-Tracking Metric	Analysis Results	Interpretation
Mean time between fixations (MTBF)	No significant differences between performance groups	Scanning speed does not impact the ability to detect certain types of failures happening with near-perfect automation
Normalized gaze transition entropy (GTE)	For the 2 min centered around the false alarm, those who detected the false alarm had higher GTE than those who did not detect the false alarm	**Instances of *more*** random/complex visual attention allocation transitions may be better able to detect automation false alarms.
Normalized stationary gaze entropy (SGE)	No significant differences between performance groups	The spread of fixations across the AOIs does not impact the ability to detect failures in near-perfect automation
Total dwell ratio	For both the entirety of the mission and for the 2 min centered on each automation failure, those who detected the miss and false alarm had a higher total dwell ratio than those who did not detect the miss and false alarm.	Consistently allocating higher proportions of consecutive visual attention to the area where automation failed is key to detecting rare-event automation failures
Number of transitions	For both the entirety of the mission and for the 2 min centered on each automation failure, those who detected the miss and false alarm had more transitions to the center search feed than those who did not detect the miss and false alarm.	Consistently transitioning visual attention *t* to the area where automation failed (sometimes more than **twice** as much) is key to detecting rare-event automation failures.

Note. AOI = area of interest.

allocation. Overall, 34% of the participants correctly identified the automation miss, and 67% correctly identified the automation false alarm. Consistent with prior research (e.g., Bliss, 2003; Chancey et al., 2015), participants detected significantly more false alarms than misses. Unfortunately, misses are often costlier than false alarms (e.g., bomb detection), and although false alarms are often considered annoying and can lead to "cry wolf" syndrome (Parasuraman & Riley, 1997), some evidence suggests that domain experts are more accepting of false alarms than misses (Masalonis & Parasuraman, 1999). Regardless, in general, the results found that the number of participants who detected both automation failures and the number who detected neither was practically equal, whereas the number of participants that detected one of the failures was approximately twice as many as either group. In summary, the performance results from this research show that that humans are only marginally reliable (34% and 67%) at intervening to correct rare-event automation failures. One could argue that any intervening detection from a human could be worthwhile, even if the improvement is marginal. The key assumption to this argument is that the additional cost of that improvement is minimal. It is possible that training or expertise could improve these detection trends, but previous research in supervisory control suggests it is unlikely training alone will lead to acceptable performance levels (e.g., Victor et al., 2018) and training would come at some cost (e.g., money, time, etc.). In summary, if near-perfect automation systems are going to include human monitoring as a layer of overall system reliability, research needs to study the human's monitoring process in these environments and design for them accordingly. Part of this process is the person's trust calibration process.

The trust ratings from participants trended as expected: trust decreased when participants detected the automation failure(s) and increased when they did not. However, the rate at which trust was lost and rebuilt was unexpected. One interesting finding from this analysis is that those who detected the first automation failure, but not the second, reported that their trust levels recovered to a level that was similar to the first trust reading, (i.e., the first 10 min of the simulation where no automation failures occurred) and similar to those who detected no automation failures. However, trust decreased rapidly from start to end for those who detected both automation failures. For context, the automation was 99.9% reliable, even with the two failures, but trust dropped to ~60% for those who detected both failures. This further supports that the trust calibration process is not directly proportional to automation reliability and is highly variable as the human detects failures. Future studies should precisely examine the relationship between the number of automation failures and the dynamics of trust recovery. The eye-tracking analysis helps to clarify the discrepancy between automation reliability and trust.

There were no significant differences between the two performance groups (those who did and did not detect the automation failures) when comparing global visual attention patterns over the entire experimental session, which is inconsistent with some previous work (Bagheri & Jamieson, 2004). This may be due to the length of the scenario being 40 min and possibly "washing out" any general visual attention allocation trends. Given that operators in the field may be tasked to this role for much longer amounts of time, this emphasizes the need for eye-tracking analyses to be analyzed on a more "real-time" basis in order to accurately capture the current state of the operator. This is somewhat supported in the present work, given that gaze transition entropy was significantly different between those who did and did not detect the false alarm for the 2-min analysis only, meaning the more complex scan patterns for those who detected the false alarm was only evident during the 2 min centered around the automation failure. This finding suggests global eye-tracking metrics should be analyzed on a more granular basis if they are to be informative of performance differences with near-perfect automation. Future research should further corroborate these findings with the exploration of different types of metrics and time intervals.

Alternatively, the local eye-tracking metrics (i.e., the ones associated with the specific sensor search feed where the automation failures

happened) were significantly different between performance groups (i.e., those who did and did not detect each type of automation failure) and for all analyses (i.e., the entire experimental session, the 2 min centered around the automation failure, and overall detection rates). Overall, participants who detected the automation failures spent the most time monitoring the center sensor feed for an average of 21%–24% of all monitoring time. They also visited this feed 1.2–1.8 times more than any other search feed (i.e., the left and right sensor feeds) and 1.3–9.2 times more than any other AOI. These results directly quantify how participants narrow their attention when near-perfect automation failed, which is consistent to previous work (Dixon & Wickens, 2006; Thomas & Wickens, 2004). These results also begin to make direct comparisons on how visual attention patterns differ between the trust levels of those who detected none and both automation failures. Interestingly, the trust ratings changed dramatically over time depending on detection rates, but the local eye-tracking metrics did not. There are two potential explanations for this: the first being a characteristic of the system and the second being a characteristic of the human. The first potential explanation of these diverging trends is due to a positive feedback loop (Smith & Smith, 1987): if you detect automation failures, you believe you are sufficiently monitoring the automation, so you do not change your monitoring approach. If you do not detect errors, you are unaware that automation needs to be monitored at all, so you do not change your monitoring approach. Second is the monitoring rates of automation are trait and not state based, that is, monitoring rates are more dependent on the characteristics of the person than the characteristics of the environment. Future studies could directly address these competing theories, but regardless, both will need to eventually inform how to provide active, real-time assistance to the operator. This is clearly warranted because regardless of some operators monitoring the automation "sufficiently" (whatever is defined as sufficient for the environment/automation at hand) and some not, the current evidence suggests those monitoring rates are relatively stagnant over time even as failures are detected, suggesting that failure detection is not sufficient feedback to impact changes in visual attention allocation patterns. Furthermore, the analysis of the local eye-tracking metrics highlights that the level of sufficiency may come at a high and unrealistic visual attention cost to the operator (e.g., spending ~21% of time monitoring one sensor feed of the whole display). Even more concerning is this cost may not lead to a reciprocal benefit of substantially improved system reliability as participants were not overwhelmingly reliable in correcting automation failures. As a sanity check, all eye-tracking data were screened to ensure participants were not attending to a secondary task when the automation failed. Given that no participants were attending to a secondary task, this suggests that participants either (1) missed the failure even when their point of gaze was in the center sensor feed that is, inattentional blindness or (2) they were allocating their attention elsewhere without being prompted to do so. To investigate if there were instances of inattentional blindness, the eye-tracking data were used to determine if at least one fixation was present in the center sensor feed at some point during an automation failure (i.e., the 14 s it was in the center sensor feed) yet it was still not detected. Of the 59 instances where an automation failure was not detected (across both automation failure types), 42 had at least one fixation in the center sensor feed during the automation failure (i.e., 71.2% of all instances). This alarming percentage of inattentional blindness seems to further indicate that the *current* cost of humans monitoring near-perfect systems outweighs any benefit. In total, the eye-tracking analysis shows the need to study eye-tracking metrics in more granular units of time to detect potential performance decrements, to include local eye-tracking metrics (i.e., ones that contextually relate to the task's goals) and to determine optimal visual attention allocation patterns. Future work should thoroughly validate all of these aspects before delivering final design guidance for near-perfect automated systems.

This work is not without limitations. The generalizability of the work needs to be limited as this was a lab-based experiment. This experiment was only 40-min long, and it is likely that real-world operators will be interacting with

near-perfect automated systems for much longer periods of time. However, this could suggest that our performance findings are understating the negative effects (e.g., vigilance decrement). Relatedly, the participants may not be representative of the person who would be tasked to this kind of monitoring, making it even more important to tease out state- and trait-based effects in monitoring. Finally, the monitoring task itself was (purposefully) simple, in order to have participants reach task proficiency in a relatively short amount of time. Realistically, monitoring tasks will be more contextually relevant to a specific aspect in a given field and will most likely be done by an expert, which may make the task more engaging and better prioritized. Future research should incorporate these elements, as well as the suggestions made above, when investigating human performance, trust, and visual attention allocation in near-perfect automated environments.

CONCLUSION

Taken together, the performance, trust, and eye-tracking data show that humans are not well suited for monitoring near-perfect automated systems. Performance is inadequate and calls into question whether humans should ever be in these roles. Additionally, inadequate performance is problematic as it dictates the trust calibration process, coming at a large cost to the human's attentional resources. From a human factors standpoint, improving the human–computer interface design of the system may be an appropriate first step. For example, uncertainty communication has been found to increase automation transparency and assist in correcting operator's mental models of the automation, which informs the trust calibration process (Beller et al., 2013; Endsley, 2017; Victor et al., 2018). Incorporating eye-tracking to aid the operator's overt visual attention allocation may improve performance, but as evidenced by data from this experiment, it is not a certainty (i.e., 71.2% of inattentional blindness instances). More generally, trying to understand the traits and current state of the operator may be more informative on their ability to successfully complete these tasks. Eye tracking may be able to aid in that understanding (e.g., apply the method used in Mracek et al., 2014 to parse out the trait and state levels of eye-tracking metrics that have found to differ on both of these levels, for example, de Haas et al., 2019; Tsukahara et al., 2016), but more work in this domain is needed. In conclusion, this research shows that humans are not well suited in the monitoring of near-perfect automated systems. Should humans be pushed into these roles, far more research is needed to understand how to best design for them.

ACKNOWLEDGMENTS

We would like to thank the Command Decision Making program, within the Office of Naval Research, for funding support.

KEY POINTS

- This research assessed performance, trust, and visual attention during the monitoring of a near-perfect automated system.
- Participants correctly identified 67% of the automation false alarms and 34% of the automation misses.
- Subjective trust increased when participants did not detect the automation failures and decreased when they did.
- Participants who detected the false alarm had a more complex scan pattern compared to those who did not in the 2 min centered around the automation failure. Additionally, those who detected the failures had longer dwell times in and transitioned to the center search feed significantly more often.
- Directing covert visual attention may not be sufficient to improve human detection performance. Using eye tracking in this context may require more than traditional human–computer interface design.

ORCID iD

Shannon Devlin https://orcid.org/0000-0003-2652-012X

SUPPLEMENTAL MATERIAL

The online supplemental material is available with the manuscript on the *HF* website.

REFERENCES

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, *1*, 212–217.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*, 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction: An approach using automation uncertainty. *Human Factors*, *55*, 1130–1141. https://doi.org/10.1177/0018720813482327

Bliss, J. P. (2003). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology*, *13*, 249–268. https://doi.org/10.1207/S15327108IJAP1303_04

Brown, C. M., & Noy, Y. I. (2004). Behavioural adaptation to in-vehicle safety measures: Past ideas and future directions. In T. Rothengatter & R. D. Huguenin (Eds.), *Traffic and transport psychology: Theory and application* (pp. 25–46).

Chancey, E. T., Bliss, J. P., Liechty, M., & Proaps, A. B. (2015). False alarms vs. misses: Subjective trust as a mediator between reliability and alarm reaction measures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 647–651. https://doi.org/10.1177/1541931215591141

Chancey, E. T., Yamani, Y., Brill, J. C., & Bliss, J. P. (2017). Effects of alarm system error bias and reliability on performance measures in a multitasking environment: Are false alarms really worse than misses? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*, 1621–1625. https://doi.org/10.1177/1541931213601890

Cohen, J. (1988). The t test for means. In *Statistical power analysis for the behavioral sciences* (2nd ed.,pp. 19–74). Lawrence Erlbaum Associates, Inc., Publishers. http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.

Coral, M. P. (2016). *Analyzing cognitive workload through eye-related measurements: A meta-analysis*. Wright State University.

Davenport, R. B., & Bustamante, E. A. (2010). Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*, 1513–1517. https://doi.org/10.1177/154193121005401933

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 11687–11692. https://doi.org/10.1073/pnas.1820553116

Dehais, F., Peysakhovich, V., Scannella, S., & Fongue, J. (2015). *"Automation surprise" in aviation: Real-time solutions* [Conference session]. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems CHI 2015, Republic of Korea.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*, 474–486. https://doi.org/10.1518/001872006778606822

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572. https://doi.org/10.1518/001872007X215656

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors*, *28*, 421–438. https://doi.org/10.1177/001872088602800405

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*, 5–27.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*, 381–394. https://doi.org/10.1518/001872095779064555

Foroughi, C. K., Sibley, C., Brown, N. L., Rovira, E., Pak, R., & Coyne, J. T. (2019). Detecting automation failures in a simulated supervisory control environment. *Ergonomics*, *62*, 1150–1161. https://doi.org/10.1080/00140139.2019.1629639

Glaholt, M. G. (2014). Eye tracking in the cockpit: A review of the relationships between eye movements and the aviator's cognitive state (Report No. DRDC-RDDC-2014-R153): Defence Research and Development Canada.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*, 517–527. https://doi.org/10.1177/0018720811417254

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*, 509–519. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*, 407–434. https://doi.org/10.1177/0018720814547570

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods, paradigms, and measures* (2nd ed.). CreateSpace Independent Publishing Platform.

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*, 146–154. https://doi.org/10.1016/j.learninstruc.2009.02.019

Kruizinga, A., Mulder, B., & De Waard, D. (2006). *Eye scan patterns in a simulated ambulance dispatcher's task*. Shaker Publishing.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*, 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lewandowsky, S., Mundy, M., & Tan, G. P. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*, 104. https://doi.org/10.1037//1076-898x.6.2.104

Liechty, J., Pieters, R., & Wedel, M. (2003). Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika*, *68*, 519–541. https://doi.org/10.1007/BF02295608

Masalonis, A. J., & Parasuraman, R. (1999). Trust as a construct for evaluation of automated aids: Past and future theory and research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*, 184–187. https://doi.org/10.1177/154193129904300312

Meyer, J., & Ballas, E. (1997). A two-detector signal detection analysis of learning to use alarms. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *41*, 186–189. https://doi.org/10.1177/107118139704100143

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*, 203–211. https://doi.org/10.1177/014233129902100408

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy, J. H., & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human Factors*, *56*, 1401–1413. https://doi.org/10.1177/0018720814533964

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, 1905–1922. https://doi.org/10.1080/00140139408964957

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, 429–460. https://doi.org/10.1080/00140139608964474

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*, 476–488. https://doi.org/10.1177/0018720813501549

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Volume 2: Cognitive processes and performance* (pp. 43.1–43.39). Wiley.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*, 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, *3*, 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253. https://doi.org/10.1518/001872097778543886

Raptis, G., Katsini, C., Belk, M., Fidas, C., Samaras, G., & Avouris, N. (2017). *Using eye gaze data and visual activities to infer human cognitive styles: Method and feasibility studies* [Conference session]. Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 164–173. https://doi.org/10.1145/3079628.3079690

Rehder, B., & Hoffman, A. B. (2003). Eyetracking and selective attention in category learning. *Proceedings of the Annual Meeting of the Cognitive Science Behaviour*, 276–281. https://doi.org/ISBN%20978-0-9768318-8-4

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*, 76–87. https://doi.org/10.1518/001872007779598082

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science*, *18*, 573–591. https://doi.org/10.1080/1463922X.2016.1252806

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors*, *49*, 347–357. https://doi.org/10.1518/001872007X196685

Shic, F., Chawarska, K., Bradshaw, J., & Scassellati, B. (2008). Autism, eye-tracking, entropy. *7th IEEE International Conference on Development and Learning, ICDL*, 73–78. https://doi.org/10.1109/DEVLRN.2008.4640808

Shiferaw, B., Downey, L., & Crewther, D. (2019). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, *96*, 353–366. https://doi.org/10.1016/j.neubiorev.2018.12.007

Shiferaw, B. A., Downey, L. A., Westlake, J., Stevens, B., Rajaratnam, S. M. W., Berlowitz, D. J., Swann, P., & Howard, M. E. (2018). Stationary gaze entropy predicts Lane departure events in sleep-deprived drivers. *Scientific Reports*, *8*, 2220. https://doi.org/10.1038/s41598-018-20588-7

Sibley, C., Coyne, J., & Thomas, J. (2016). Demonstrating the supervisory control operations user testbed (SCOUT). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 1324–1328. https://doi.org/10.1177/1541931213601306

Smith, T. J., & Smith, K. U. (1987). Feedback control mechanisms of human behavior. In G. Salvendy (Ed.), *Handbook of human factors* (pp. 251–293). Wiley.

Thomas, L. C., & Wickens, C. D. (2004). Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*, 223–227. https://doi.org/10.1177/154193120404800148

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*, 109–123. https://doi.org/10.1016/j.cogpsych.2016.10.001

United States Department of Defense. (2017). *Unmanned systems integrated roadmap FY2017-2042*.

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human Factors*, *60*, 1095–1116. https://doi.org/10.1177/0018720818788164

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*, 433–441. https://doi.org/10.1518/001872008X312152

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2012). Attention in perception and display space. In *Engineering Psychology and Human Performance* (4th ed.). East Sussex. https://books.google.co.mz/books?id=_rFmCgAAQBAJ&pg=PR1&hl=pt-PT&source=gbs_selected_pages&cad=2#v=onepage&q&f=false.

Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention HMI 02-03*.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*, 352–367. https://doi.org/10.1080/14639220110110306

Winnefeld, J. A., & Kendall, F. (2013). Unmanned systems integrated roadmap FY 2013-2038. *Approved for Open Publication Reference*, 0553.

Cyrus K. Foroughi is a research scientist at the Naval Research Laboratory. He earned his PhD in human factors and applied cognition from George Mason University in 2016.

Shannon Devlin is a PhD candidate in Systems Engineering at the University of Virginia. She earned her MS in industrial engineering from Clemson University in 2018.

Richard Pak is a professor in the Department of Psychology at Clemson University. He received his PhD in psychology in 2005 from the Georgia Institute of Technology.

Noelle L. Brown is a research scientist at the Naval Research Laboratory. She earned her PhD in cognitive psychology from Louisiana State University in 2011.

Ciara Sibley is a research scientist at the Naval Research Laboratory. She earned an MA in human factors and applied cognition from George Mason University in 2009 and is currently completing a PhD in computational social science.

Joseph T. Coyne is a research scientist at the Naval Research Laboratory. He earned his PhD in human factors from Old Dominion University in 2004.

Date received: January 4, 2021
Date accepted: June 14, 2021