Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Detecting automation failures in a simulated supervisory control environment

Cyrus K. Foroughi, Ciara Sibley, Noelle L. Brown, Ericka Rovira, Richard Pak & Joseph T. Coyne

To cite this article: Cyrus K. Foroughi, Ciara Sibley, Noelle L. Brown, Ericka Rovira, Richard Pak & Joseph T. Coyne (2019): Detecting automation failures in a simulated supervisory control environment, Ergonomics, DOI: 10.1080/00140139.2019.1629639

To link to this article: https://doi.org/10.1080/00140139.2019.1629639

Accepted author version posted online: 10 Jun 2019.
Published online: 26 Jun 2019.

ERGONOMICS
https://doi.org/10.1080/00140139.2019.1629639

Taylor & Francis
Taylor & Francis Group

ARTICLE

Detecting automation failures in a simulated supervisory control environment

Cyrus K. Foroughi[a,b], Ciara Sibley[a], Noelle L. Brown[a], Ericka Rovira[c], Richard Pak[d] and Joseph T. Coyne[a]

[a]U.S. Naval Research Laboratory, Washington, DC, USA; [b]Department of Psychology, George Mason University, Fairfax, VA, USA; [c]Department of Behavioral Sciences and Leadership, U.S. Military Academy, West Point, NY, USA; [d]Department of Psychology, Clemson University, Clemson, SC, USA

ABSTRACT

The goal of this research was to determine how individuals perform and allocate their visual attention when monitoring multiple automated displays that differ in automation reliability. Ninety-six participants completed a simulated supervisory control task where each automated display had a different level of reliability (namely 70%, 85% and 95%). In addition, participants completed a high and low workload condition. The performance data revealed that (1) participants' failed to detect automation misses approximately 2.5 times more than automation false alarms, (2) participants' had worse automation failure detection in the high workload condition and (3) participant automation failure detection remained mostly static across reliability. The eye tracking data revealed that participants spread their attention relatively equally across all three of the automated displays for the duration of the experiment. Together, these data support a system-wide trust approach as the default position of an individual monitoring multiple automated displays.

Practitioner Summary: Given the rapid growth of automation throughout the workforce, there is an immediate need to better understand how humans monitor multiple automated displays concurrently. The data in this experiment support a system-wide trust approach as the default position of an individual monitoring multiple automated displays.

Abbreviations: DoD: Department of Defense; UA: unmanned aircraft; SCOUT: Supervisory Control Operations User Testbed; UAV: unmanned aerial vehicle; AOI: areas of interest

ARTICLE HISTORY
Received 11 September 2018
Accepted 2 June 2019

KEYWORDS
Automation; automation failures; human-automation interaction; supervisory control; attention allocation; system-wide trust; eye-tracking

1. Introduction

The push to automate technologies by both the private and government sector has consequently changed the role of a human operator from an active, functional participant to a passive monitor. For example, Uber is currently testing self-driving cars that have humans sitting in the driver's seat monitoring the vehicle. The United States Department of Defence (DoD) is on record pushing for more autonomy while moving humans to a monitoring role:

> The UA [Unmanned Aircraft] must improve to higher levels of autonomy and the human to higher levels of management. This would migrate operational responsibility for tasks from the ground station to the aircraft, the aircraft gaining greater autonomy and authority, the humans moving from operators to supervisors, increasing their span of control while decreasing the manpower requirements to operate the UA. (DoD Unmanned Aircraft System Roadmap 2005)

Given this rapid growth and ubiquity of automation, there is an immediate need to better understand how humans interact with automated systems (i.e. human-automation interaction). In a recent review of this domain, Endsley (2017) highlighted that human operators often become automation *monitors* when automated software is introduced to a system or task. That is, humans are monitoring the automated system and must intervene if they notice something wrong (e.g. an automation failure). Importantly, in reference to humans becoming automation monitors, Hancock (2013) eloquently noted, 'the human operator is arguably magnificently disqualified for this particular form of sustained attentive response' (see also, Hancock 1991). Hancock's remarks are exemplified by a recent tragedy. On 18 March 2018, an automated car struck

CONTACT Cyrus K. Foroughi cyrus.foroughi@nrl.navy.mil U.S. Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375, USA

and killed a pedestrian in Arizona as she was crossing the street. Importantly, a human driver was present and was supposed to monitor the automated car, to act as a line of defense against automation failures. However, the driver did not notice that the automated system failed to detect the pedestrian crossing the street. This system breakdown is known as an automation miss and is not an isolated incident. Therefore, it is critical to further investigate how to design these systems for operators in these roles.

1.1. Monitoring

Humans are limited in their ability to continuously monitor information over time. Some of the first evidence to support this statement comes from World War II. Radar operators were failing to detect enemy submarines and were even misidentifying friendly vessels and large whales as enemy submarines (Mackworth 1948, 1950; see also Head 1923). In addition, the operators' performance appeared to worsen over time. Importantly, this was not a result of lack of motivation or skill (Parasuraman 1987). This phenomenon, referred to as the vigilance decrement, is the 'deterioration in the ability to remain vigilant for critical signals with time, as indicated by a decline in the rate of the correct detection of signals' Parasuraman (1987). The majority of vigilance research shows clear performance decrements in humans over time (see Parasuraman (1987) and Hancock (2013) for more information on this topic). Importantly, Schmidtke (1966) showed that the vigilance decrement existed not only in a controlled, experimental study but in an actual operational setting (experienced radar operators during ship navigation).

Not surprisingly, with the rise of automated technologies, researchers became interested in joint human-automation monitoring (i.e. combining human and automated monitoring of a system). Thus, two separate entities would be monitoring a system, reducing the chance of missing a system failure. The results of this early work were promising, but not ideal. One common finding revealed that when humans were being aided by automation, more signals were detected overall, but false alarms rates also increased (i.e. a criterion shift within a signal detection framework; Murrell 1977). Thus, overall performance (i.e. sensitivity within a signal detection framework) did not always increase with the addition of automation. This early work and interest in automated detection systems gave rise to research on how humans interact with automated systems. This criterion shift, compared to a sensitivity change, potentially implicates changes in operator trust and overall trust strategies.

1.2. Automation reliability and trust

Ideally, automated systems would be 100% reliable and not only improve human operator performance, but also offset task interference and reduce workload (e.g. Dixon, Wickens, and Chang 2005). Unfortunately, automated aids are rarely 100% reliable, meaning understanding how humans interact with imperfect automated systems is critical. Thus far, research has shown that unreliable automation has led to human operator complacence and reliance (Dixon, Wickens, and McCarley 2007; Dixon and Wickens 2006; Metzger and Parasuraman 2005; Parasuraman, Molloy, and Singh 1993; Wickens et al. 2005), different states of trust (Parasuraman and Riley 1997) and performance reductions (Molloy and Parasuraman 1996). Despite this, unreliable automation can aid human operators in many scenarios, such as when a task is difficult (Maltz and Shinar 2003).

The level of automation reliability also impacts detection performance. Bagheri and Jamieson (2004) found that participants detected more automation failures when automation reliability was low. However, other researchers have found that higher reliability rates can lead to improved performance (e.g. Chancey et al. 2015; Chancey et al. 2017; Rovira, McGarry, and Parasuraman 2007). The impact of reliability and trust is even more complex since individuals often need to interact with multiple automated systems simultaneously (e.g. in aviation, a plane has an automated system for avoiding terrain and a separate system for detecting and avoiding collisions). One open question is how individuals will perform when interacting with multiple automated systems of varying reliability across different types of automation failures.

1.3. Automation failures (false alarms and misses)

Often times, research on human-automation interaction evaluates how automation's imperfect event detection impact operator behaviour (e.g. Dixon and Wickens 2006; Parasuraman, Molloy, and Singh 1993). This research also maps nicely onto many real-world jobs that employ automated decision aids (e.g. TSA operators, satellite imaging and detection, and x-ray detection). These automation failures come in two forms: automation misses and automation false alarms. Humans tend to be better able to detect automation

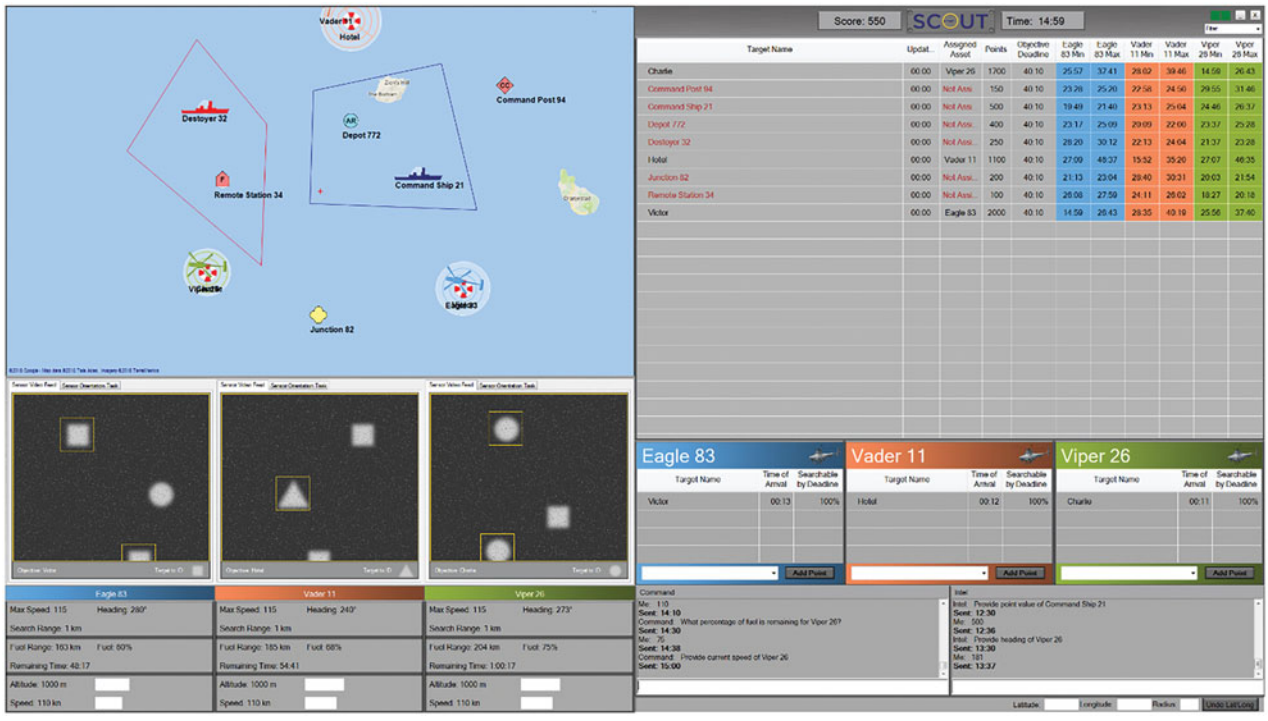

Figure 1. The Supervisory Control User Testbed (SCOUT).

false alarms compared to automation misses (e.g. Bliss 2003). This is a consequence of automation bias (Parasuraman and Manzey 2010; Wickens and McCarley 2008). That is, when automation failure results in a false alarm, that failure has cued the human to the target, drawing their attention to it for cognitive processing. However, in many cases, misses are costlier than false alarms (e.g. bomb detection). Although false alarms are often considered annoying and can lead to 'cry wolf' syndrome (Parasuraman and Riley 1997), some evidence suggests that domain experts are more accepting of false alarms than misses (Masalonis and Parasuraman 1999).

Other work in this area has evaluated detection of automation failures in regards to compliance and reliance (e.g. Dixon, Wickens, and McCarley 2007; Wickens et al. 2005). Compliance being how an operator responds when an automated alert sounds while reliance characterises what an operator does during periods of no alerts or silence. For example, a compliant operator will redirect their attention immediately when an automated alert sounds (i.e. complying with the automation) and a reliant operator will focus on other tasks assuming the automation will alert them if and when anything fails (i.e. relying on the automation). Results from this work suggest that compliance and reliance are not entirely independent of each other (Dixon, Wickens, and McCarley 2006).

1.4. Current study motivation, goals and practical importance

The current study was motivated by the need to better understand how humans monitor automated displays and detect automation failures. Given the United States' Department of Defence's active push towards automating unmanned aircrafts, we chose a simulated supervisory control task as our testbed. Specifically, we employed the Supervisory Control Operations User Testbed (SCOUT), a newly developed supervisory control environment (see Figure 1) that was designed by scientists at the U.S. Naval Research Laboratory (Sibley, Coyne, and Thomas 2016) to simulate the current and future demands of unmanned aerial vehicle (UAV) pilots. Given the likelihood that UAV pilots will be monitoring multiple automated displays in the future, we experimentally manipulated the reliability of three separate automated displays responsible for detecting targets within SCOUT. Additionally, given that workload demands are rarely static, workload was manipulated. Finally, we used eye tracking to get a real-time index of overt attention allocation.

Here our specific goals were (1) to determine how well individuals detected automation misses and false alarms (i.e. failures) when monitoring multiple automated displays that have varied automation reliability and (2) to identify where individuals directed their

overt attention when monitoring multiple automated displays that have varied automation reliability.

This work has practical importance because the role of a human in an automated system is shifting to becoming a monitor of information and detector of automation failures. Thus, there is a need to determine how well humans can monitor automated systems and detect automation failures. Understanding how humans direct their attention while monitoring automated systems and detecting automation failures will also further inform the overt attention allocation strategies used in these environments.

1.5. Hypotheses related to goal 1

1.5.1. Automation detection (failure type)

Because automation false alarms cue the human operator to the target, we expect that individuals will be able to detect more automation false alarms than misses:

H_1: Humans will detect FA more than Misses

1.5.2. Automation detection (workload)

Because the attentional demands placed on a human increase as workload increases, we expect individuals to rely on the automation more in the high workload condition resulting in worse automation failure detection performance:

H_2: Humans will perform worse in the high workload condition

1.5.3. Automation detection (reliability)

Previous research has not agreed on how automation failure detection performance varies as automation reliability changes. Some work shows that failure detection increases (e.g. Bagheri and Jamieson) when reliability is lower while other work shows that detection increases as reliability increases (e.g. Chancey et al. 2015; Chancey et al. 2017). This results in two competing hypotheses:

$H_{3\text{-}1}$: Humans will detect more automation failures when automation reliability is lower

$H_{3\text{-}2}$: Humans will detect more automation failures when automation reliability is higher

1.6. Hypotheses related to goal 2

There are two theories that can be used to make hypotheses related to identifying where individuals will direct their attention when monitoring multiple automated displays that have varied automation reliability: the system-wide trust theory and the component-specific trust theory (see Keller and Rice 2009 for more information). The system-wide trust theory predicts that humans will trust all of the automated visual monitoring tasks equally regardless of reliability, thus treating them as one 'system'. Conversely, the component-specific trust theory predicts that humans will have different levels of trust in each automated visual monitoring tasks, thus treating each as an individual 'component.' Two competing attentional-based hypotheses present as a result of these two theories:

$H_{4\text{-}1}$: Individuals will spread their attention equally (e.g. total time spent viewing each task) across all of the automated tasks supporting a prediction made by the system-wide trust theory.

$H_{4\text{-}2}$: Individuals will not spread their attention equally (e.g. more time spent viewing task with lowest automation reliability) across all of the automated tasks supporting a prediction made by the system-wide trust theory.

2. Method

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board's at both the U.S. Naval Research Laboratory and George Mason University (#1037839). Informed consent was obtained from each participant.

2.1. Participants

Ninety-six students (M age = 20.4 years, SD age = 4.9 years, 70 females) from George Mason University participated in this research for course credit.

2.2. Primary task

The Supervisory Control Operations User Testbed (SCOUT) is a simulated supervisory control environment (see Figure 1) designed by scientists at the U.S. Naval Research Laboratory (Sibley, Coyne, and Thomas 2016) to simulate the current and future demands of UAV pilots. This task 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). Importantly, when a UAV reaches its target, the sensor search feed for that

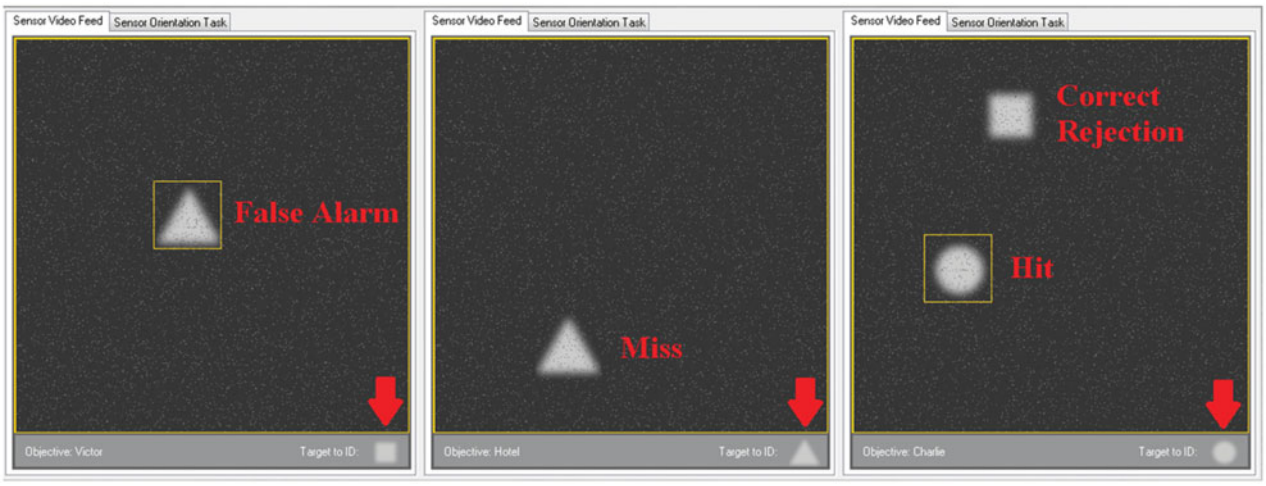

Figure 2. This is an example of the sensor feeds from SCOUT. There is an icon below each sensor feed indicating a unique target to identify: square, triangle and circle from left to right in this figure, as noted by the red arrows. The automated system automatically highlights targets by placing a gold box around them. Participants are to ensure that the automated system accurately identifies the correct targets. If the automated system misses a correct target 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. Participants would need to click on the false alarm and miss to correct the automation errors.

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. The automation reliability can be set from 0% to 100% reliable for both misses on possible targets and false alarms on distractor targets.

2.3. Equipment

A 24-inch Dell P2415Q monitor set at 2560 × 1440 resolution was used for this experiment. Participants' heads were approximately 65 cm away from the display. Eye tracking data were collected at 60 Hz using a GazePoint GP3 Desktop Eye Tracker. The eye trackers were calibrated for each participant using a 9-point calibration programme built by GazePoint for use with their system. The GP3 provides left and right point of gaze in pixels, as well as left and right pupil diameter in pixels and in millimetres. Each data point in the GP3 data stream also has a binary quality measure associated with it, which indicates whether the system believes the data is valid. Luminance levels were standardised for all participants as everyone completed the experiment in the same windowless room with consistent lighting conditions.

2.4. Procedure

After signing an informed consent form, participants were instructed to get comfortably seated. First, the eye tracker was set up and calibrated using the GazePoint GP3 software. Next, participants completed a fixation test as an additional calibration tool. Participants then completed the colour change task and the shortened automated operation span. The colour change task and operation span were not analysed for this manuscript. They are both being combined with more data from other projects as part of a larger individual differences project that is not yet complete (see Rovira, Pak, and McLaughlin 2017 for more information regarding individual differences and automation).

Participants then completed a SCOUT training session to teach them how to properly complete the task. After completion, participants were given a short comprehension test about SCOUT to ensure they understood all of the features of the task. Participants were required to answer every question correctly to continue. All participants answered all of the comprehension questions correctly on their first attempt.

Participants then completed the experimental scenario of SCOUT. Participants controlled three separate UAVs that were each searching different targets concurrently. 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. In each search feed, the participant was informed to look for one of three possible targets (see Figure 2). Triangles, circles and squares were the objects used as the targets and distractors for this

experiment. All objects would enter at the top of the sensor and then vertically scroll down the feed. The sensor feed task was automated such that some objects were automatically highlighted if it matched the target for that specific sensor feed. The participants' responsibility was to ensure that targets were highlighted and distractors were not highlighted for each sensor feed. For example, if the participant was searching for the triangle in a specific search feed, the participant would need to ensure that all of the triangles (i.e. targets) that scrolled across the screen were highlighted, and any of the circles or squares (i.e. distractors) were not highlighted. Participants could change the status of any object 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. The target within each search feed remained constant across the experiment.

Participants completed two conditions in a counterbalanced order: high and low workload. In the low workload condition, the distractors appeared on the search feeds every 15 ± 3 s and the targets appeared on the search feeds every 18 ± 3 s. In the high workload condition, this rate tripled resulting in distractors appearing every 5 ± 1 s and targets appearing every 6 ± 1 s. This resulted in targets appearing 45.5% of the time and distractors appearing 55.5% of the time. Across both conditions, the scroll time in every search feed was 14 s; that is, participants had 14 s to alter whether any object was highlighted or not highlighted. In both conditions, each search feed had a different automation reliability rate: (1) 70% accurate, (2) 85% accurate, and (3) 95% accurate. Automation failures were evenly distributed with half being misses and half being false alarms. Additionally, chat queries appeared in the chat box every 15 s. For example, 'Provide the fuel range of Eagle 83'. Participants were instructed to complete all of the tasks in SCOUT as quickly and accurately as possible. At the end of each condition, participants completed the crew status survey, a subjective fatigue and workload scale (Ames and George 1993; Samn and Perelli 1982). See Figure 2 for a visualisation of how the automation operates.

2.5. Measures

2.5.1. SCOUT performance

We present the automated display performance data in two forms. First, we present the data as a function of the human only. That is, assessing how well the participants corrected the automation failures without adding in the base automation accuracy (e.g. 85%). Following that, we will add in the automation base accuracy rate and present the data as a function of both the human and automation's performance. Additionally, chat prompt accuracy is reported.

2.5.2. Subjective measures

The crew status survey (Ames and George 1993; Samn and Perelli 1982) was employed to subjectively assess an individual's maximum and average workload as well as fatigue during both the high and low workload conditions.

2.5.3. Eye-Related measures

Pupil size was measured across the experiment as a proxy for workload (Kahneman and Beatty 1966; Hess and Polt 1964). Total time viewing each automated display was recorded as a measure of attention allocation.

3. Results

3.1. Eye data reduction

Participants were required to have 70% valid data or higher in both eyes to be considered for eye tracking analyses. This resulted in removing 22 participants' eye data from consideration. The 74 remaining participants had an average of 85% valid data in both eyes. Pupil data were cleaned before analysis. Data marked invalid by GazePoint were removed, then data were within-subject standardised and windsorized to $z = \pm 3$. The entire dataset was used for all other analyses.

3.2. SCOUT performance

3.2.1. Automated displays

3.2.1.1. Humans only. Figure 3 presents the percentage of automation failures that participants failed to detect for both types of failures (namely miss and false alarm), both levels of workload (namely high and low), and all three reliability levels (namely 70%, 85% and 95%). For example, if a participant was searching for triangles and 100 triangles scrolled during the experiment, but only 85 were properly highlighted by automation (i.e. 85% automation reliability), the participant would have had to manually highlight the other 15. Of these 15, if the participant highlighted 10 and failed to highlight the other 5, the percentage of automation failures that the participant missed would be 33.3% (5 out of 15).

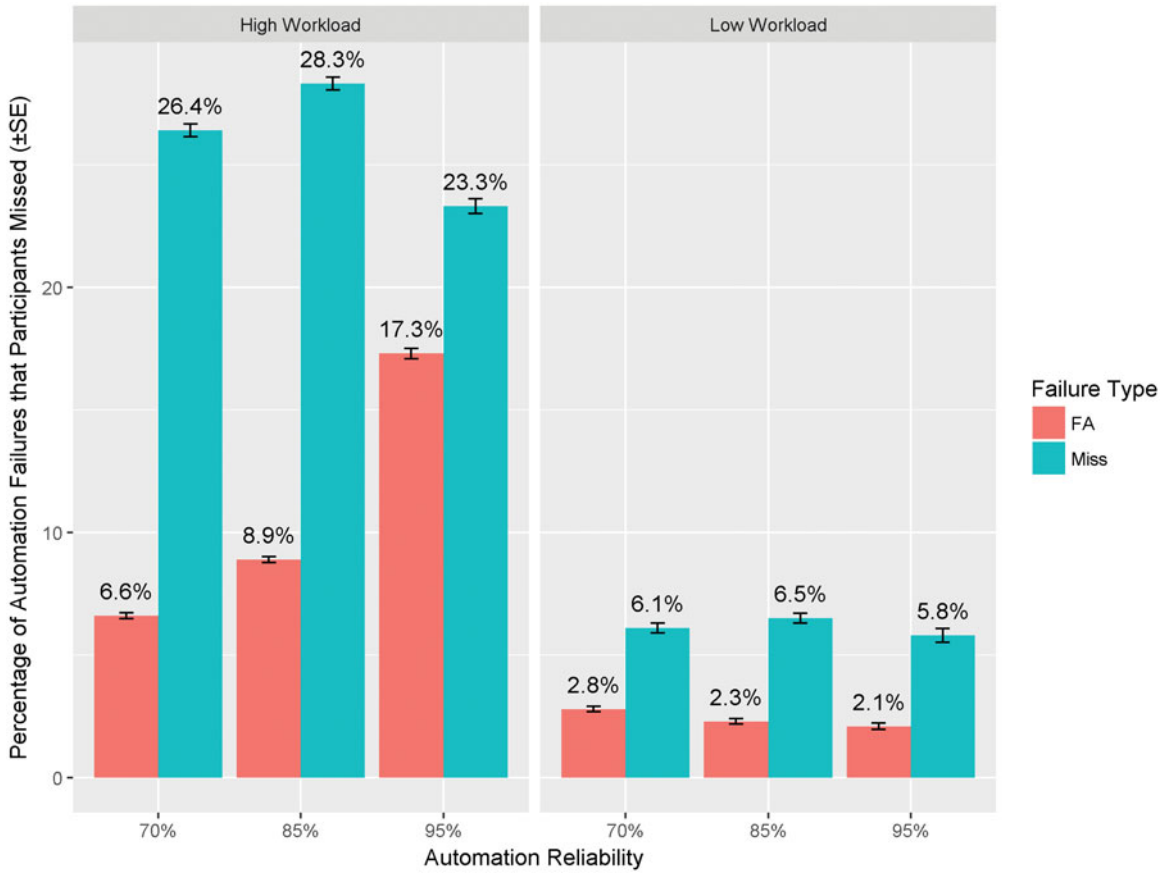

Figure 3. The percentage of automation failures (±SE) that participants failed to detect across both conditions and all reliability levels. Higher percentages are worse performance.

For misses, there was a main effect of condition such that the detection of automation misses was significantly worse in the high workload condition compared to the low workload condition (F (1,569) = 96.2, $p < .001$). No effect of sensor reliability (F (2,569) = .661, p = .52) or interaction were detected (F (2,569) = .39, p = .68). For false alarms, there was an interaction between condition and reliability (F (2,570) = 9.6, $p < .001$), a main effect of condition (F (1,570) = 58.5, $p < .001$), and a main effect of reliability (F (2,570) = 7.6, $p < .001$). Contrasts using mixed-effects modelling with a Bonferroni correction revealed that within the high workload condition, detection of the false alarms in the 95% reliable sensor was significantly worse than both the 85% reliable sensor (t = −6.6, $p < .001$, d = .52) and the 70% reliable sensor (t = −8.4, $p < .001$, d = .65). No other contrasts were significantly different.

3.2.1.2. Combined human-automation. Figure 4 presents the combined human-automation accuracy for both types of failures (namely miss and false alarm), both levels of workload (namely high and low) and all three reliability levels (namely 95%, 85% and 70%). For example, if a participant was searching for circles and 100 circles scrolled in X time at an automated accuracy of 70% (i.e. 70 circles were highlighted properly and 30 were not), and the participant manually corrected 22 of the 30 targets, the overall accuracy would be 92% ((70 + 22)/100 = 92%).

For misses, there was a main effect of condition, F (1,570) = 73.5, $p < .001$, a main effect of reliability, F (2,570) = 33.8, $p < .001$, and an interaction, F (2,570) = 10.4, $p < .001$. Contrasts using mixed-effects modelling with a Bonferroni correction revealed that within the high workload condition, detection of the miss within the 95% reliable sensor was significantly greater than both the 85% sensor (t = −7.27, $p < .001$, d = 1.07) and the 70% reliable sensor (t = −12.4, $p < .001$, d = 1.20). Additionally, detection of the miss within the 85% reliable sensor was significantly greater than the 70% sensor (t = −12.4, $p < .001$, d = .53). Within the low workload condition, detection of the miss within the 95% reliable sensor was significantly greater than the 70% reliable sensor (t = −3.5, p = .007, d = .38). No other contrasts were significantly different for miss detection. For false alarms, there was a main effect of condition such that participants

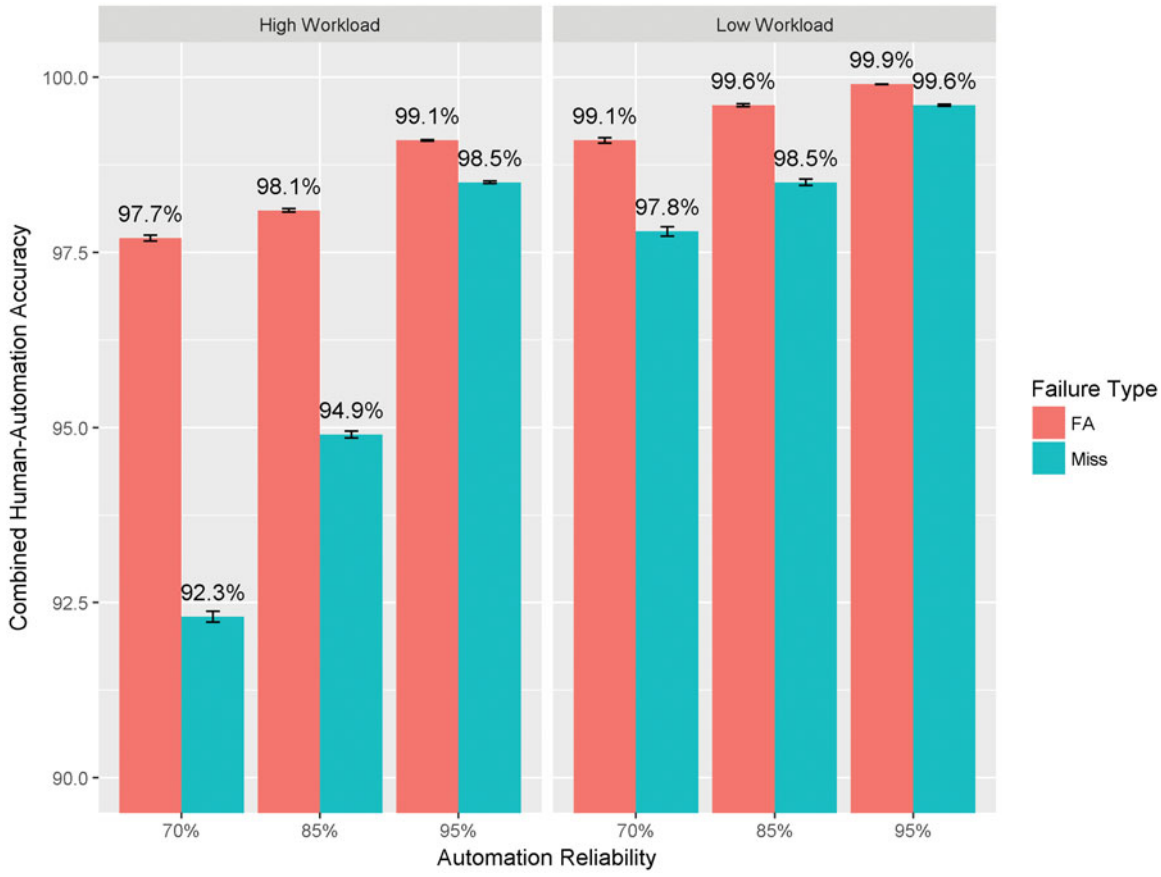

Figure 4. The combined human-automation accuracy (±SE) across both conditions and all reliability levels. Higher percentages are better performance.

detected more automation false alarms in the low workload condition compared to the high workload condition, F (1,570) = 32.9, $p < .001$. A main effect of reliability such that participants detected more false alarms as automation reliability increased, F (2,570) = 8.3, $p < .001$. There was no interaction effect F (2,570) = 1.03, $p = .36$.

3.2.2. Chat prompts

As expected, participants successfully answered more chats in the low workload condition ($M = 81\%$ $SD = 10.5$) compared to the high workload condition ($M = 72.8\%$, $SD = 15$), $t = -6.65$, $p < .001$, $d = .63$.

3.2.3. Workload and fatigue

3.2.3.1. Subjective workload. Participants self-reported higher levels of maximum workload in the high workload condition ($M = 4.58$, $SD = 1.17$) compared to the low workload condition ($M = 4.1$, $SD = 1.2$), $t = 5.47$, $p < .001$, $d = .41$. Participants also self-reported higher levels of average workload in the high workload condition ($M = 3.76$, $SD = 1.09$) compared to the low workload condition ($M = 3.34$, $SD = 1.13$), $t = 4.31$, $p < .001$, $d = .38$.

3.2.3.2. Subjective fatigue. Participants did not self-report differences in fatigue between the high ($M = 3.07$, $SD = 1.34$) and low ($M = 3.14$, $SD = 1.24$) workload conditions, $t = -.73$, $p = .47$, $d = .05$.

3.2.3.3. Pupil size. Participants had significantly larger pupils in the high workload condition compared to the low workload condition, $M_{Diff} = .15$, $SD_{Pooled} = .25$, $p = .01$, $d = .61$.

3.3. Attention allocation

We were interested in determining where the participants were directing their attention while monitoring the three automated sensor feeds over time. We, therefore, created three areas of interest (AOI) within SCOUT for each sensor search feed and calculated the percentage of time that each participant viewed each sensor search feed (when viewing any of the sensor search feeds). Thus, if the data were randomly distributed, we would expect participants to view each screen 33.33% of the time.

For the low workload condition, we did not detect a main effect of AOI (F (2,1704) = 1.3, $p = .27$) or time

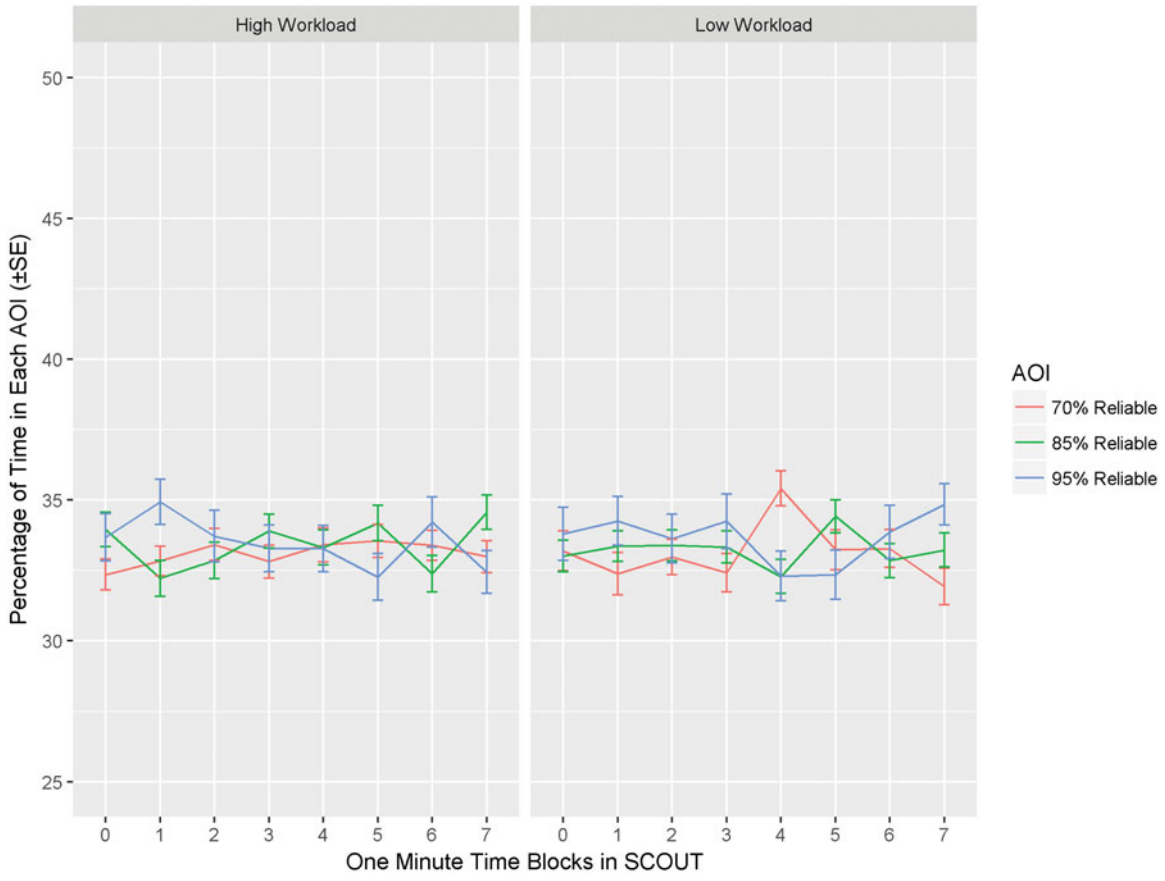

Figure 5. The percentage of time viewing each AOI (i.e. individual sensor feed) when participants were only viewing the sensor feeds (±SE).

(F (7,1704) = 0.01, p = .999) but did detect an interaction of AOI and time (F (14,1704) = 2.18, p = .007). For the high workload condition, we did not detect a main effect of AOI (F (2,1704) =.72, p = .49) or time (F (7,1704) = 0, p = 1) but did detect an interaction of AOI and time (F (14,1704) = 1.8, p = .03). Post-hoc contrasts were not significantly different when using a Bonferroni correction (Figure 5). On average, participants viewed approximately each sensor search feed equally over time.

4. Discussion

The goal of this research was to determine how individuals perform and where they direct their attention when monitoring multiple automated displays of different reliability. When just looking at operator performance (Figure 3), (1) participants' failed to detect automation misses approximately 2.5 times more than false alarms in both workload conditions, supporting H_1, (2) participants' performance was worse in the high workload condition compared to the low, supporting H_2, and (3) participant detection of automation failures remained mostly stable across reliability in both workload conditions, not supporting either $H_{3\text{-}1}$ or $H_{3\text{-}2}$. The eye data revealed that participants spread their attention relatively equally across all three sensor feeds for the duration of the experiment in both the high and low workload conditions, supporting $H_{4\text{-}1}$ (system-wide trust).

The finding that participants' detection of automation failures remained mostly stable across reliability is interesting and is not entirely consistent with previous work (Bagheri and Jamieson, 2004; Chancey et al. 2015; Chancey et al. 2017). The lone exception to this finding was the reduced detection rate in the high workload condition for false alarms, which most likely represents an outlier given the rest of the data.

One possible explanation for the relatively stable human detection performance is the system-wide trust strategy that the participants appeared to employ. That is, if participants were equally spreading their attention across the automated displays, it is not unreasonable to assume that their detection performance would be approximately the same, regardless of automation reliability. Other research in this area often

compares detection performance as a function of reliability where participants are interacting with one system at a time and reliability is manipulated within- or between-subject. Here, participants interacted with all three reliability rates/levels concurrently.

Another possible explanation for this overall finding is the calculation method used to determine human-only and human-automation system performance. Here, we presented both methods. This allowed us to determine the participant detection performance without adding in the baseline automated system reliability. Thus, it makes sense that when adding in the automation base detection rate, Figure 4 shows a steady increase in human-automation detection performance with an increase in reliability rate. So, if detection of automation failures remains stable regardless of the automation reliability, then more reliable automation will lead to better overall combined performance (see Figures 3 and 4). To our knowledge, researchers do not commonly report both methods, and most seem to report the combined human-automation performance. However, this analysis shows increases in detection performance are a function of increases automation reliability and not any improvement of the human operator.

The eye tracking/gaze data revealed that across both workload conditions, participants spent a similar amount of time viewing each sensor feed. This result supports a system-wide trust approach (e.g. Keller and Rice 2009; Walliser, de Visser, and Shaw 2016). That is, the participants appeared to equally trust each sensor feed even though they varied in automation reliability. This suggests operator's default approach to view multiple automated displays is to spread their attention equally, which falls in the line with the aforementioned performance results. The Gestalt design principle of similarity, which suggests that people naturally group like items together, may explain why the three automated displays were monitored with equal frequency (see Rock and Palmer 1990). This has important design implications as system designers need to realise that operators may not accurately account for individual components within a system to vary in reliability, and therefore systems should be designed to assist operators in accurately calibrating their trust to each system. Future research could investigate how well specific design features or training/instruction can overcome this issue.

These data support the position that we should aim for automated systems that are highly reliable because human performance may remain relatively static across reliability. Although this may seem like an obvious point, some research suggests that variable reliability may be more ideal (e.g. Parasuraman, Molloy, and Singh 1993) and may reduce automation-induced complacency. One possible caveat to this design approach is the challenge in human detection of very rare and unexpected automation failures, known as 'black swans' (Molloy and Parasuraman 1996; Sebok and Wickens 2017; Wickens et al. 2009). Often times, these types of failures are completely missed, and Craig (1984) estimated that detecting this type of failure when it is critical may occur as rarely as once in two weeks in the field. If failures occurred at such a low rate, it is possible that unless a critical signal is present, the human operator would miss it regularly.

This work helped reveal how individuals complete and attend to multiple automated displays with varying reliability. Overall, the data support a system-wide trust approach as the default position of an individual monitoring multiple automated displays. More work in this area is needed to better understand how individuals attend to and process information when interacting with automated systems. Like all laboratory research, this work has limitations. The results could have varied if different populations were tested and replication of the findings would be wise. Our immediate future work hopes to expand these results by testing different levels of automation reliability, including extremely reliable systems (e.g. 99.9%).

Funding

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

ORCID

Cyrus K. Foroughi http://orcid.org/0000-0002-9699-6812
Ericka Rovira http://orcid.org/0000-0002-4820-5828
Richard Pak http://orcid.org/0000-0001-9145-6991

References

Ames, L. L., and E. J. George. 1993. Revision and verification of a seven-point workload estimate scale (Air Force Flight Test Center Report number AFFTC-TIM-93-01) Edwards Air Force Base, CA.

Bagheri, N., and G. A. Jamieson. 2004. "Considering Subjective Trust and Monitoring Behavior in Assessing Automation-Induced "Complacency"." Proceedings of the Human Performance, Situation Awareness and Automation Conference, 54–59. Marietta, GA: SA Technologies.

Bliss, J. P. 2003. "Investigation of Alarm-related Accidents and Incidents in Aviation." *The International Journal of Aviation Psychology 13* (3):249–268.

Chancey, E. T., J. P. Bliss, A. B. Proaps, and P. Madhavan. 2015. "The Role of Trust as a Mediator between System Characteristics and Response Behaviors." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57(6): 947–958. doi:10.1177/0018720815582261.

Chancey, E. T., Y. Yamani, J. C. Brill, and J. P. Bliss. 2017. "Effects of Alarm System Error Bias and Reliability on Performance Measures in a Multitasking Environment: Are False Alarms Really Worse than Misses? In." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61(1):1621–1625.. doi:10.1177/1541931213601890.

Craig, A. 1984. "Human Engineering: The Control of Vigilance." In *Sustained Attention in Human Performance*, edited by J.S. Warm, 247–291. Chichester, England: Wiley.

Dixon, S. R., and C. D. Wickens. 2006. "Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48(3):474–486. doi:10.1518/001872006778606822.

Dixon, S. R., C. D. Wickens, and D. Chang. 2005. "Mission Control of Multiple Unmanned Aerial Vehicles: A Workload Analysis." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 47(3):479–487. doi:10.1518/001872005774860005.

Dixon, S. R., C. D. Wickens, and J. S. McCarley. 2007. "On the Independence of Compliance and Reliance: Are Automation False Alarms Worse than Misses?" *Human Factors* 49(4):564–572. doi:10.1518/001872007X215656.

Endsley, M. R. 2017. "From Here to Autonomy: Lessons Learned from Human–Automation Research." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59(1):5–27. doi:10.1177/0018720816681350.

Hancock, P. A. 1991. "On Operator Strategic Behavior." Proceedings of the Sixth International Symposium on Aviation Psychology, 999–1007. Columbus: Ohio State University.

Hancock, P. A. 2013. "In Search of Vigilance: The Problem of Iatrogenically Created Psychological Phenomena." *American Psychologist* 68(2):97. doi:10.1037/a0030214.

Head, H. 1923. "The Conception of Nervous and Mental Energy 1 (ii) 'Vigilance." *British Journal of Psychology. General Section* 14(2):126–147. doi:10.1111/j.2044-8295.1923.tb00122.x.

Hess, E. H., and J. M. Polt. 1964. "Pupil Size in Relation to Mental Activity during Simple Problem-Solving." *Science* (*New York*, N.Y.) 143(3611):1190–1192. doi:10.1126/science.143.3611.1190.

Kahneman, D., and J. Beatty. 1966. "Pupil Diameter and Load on Memory." *Science* (*New York*, N.Y.) 154(3756): 1583–1585.

Keller, D., and S. Rice. 2009. "System-Wide versus Component-Specific Trust Using Multiple Aids." *The Journal of General Psychology: Experimental, Psychological, and Comparative Psychology* 137(1):114–128. doi:10.1080/00221300903266713.

Mackworth, N. H. 1948. "The Breakdown of Vigilance during Prolonged Visual Search." *Quarterly Journal of Experimental Psychology* 1(1):6–21. doi:10.1080/17470214808416738.

Mackworth, N. H. 1950. *Researches on the Measurement of Human Performance*, 268. London: His Majesty's Stationery Office.

Maltz, M., and D. Shinar. 2003. "New Alternative Methods of Analyzing Human Behavior in Cued Target Acquisition." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 45(2):281–295. doi:10.1518/hfes.45.2.281.27239.

Masalonis, A. J., and R. Parasuraman. 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(3): 184–187. doi:10.1177/154193129904300312.

Metzger, U., and R. Parasuraman. 2005. "Automation in Future Air Traffic Management: Effects of Decision Aid Reliability on Controller Performance and Mental Workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 47(1):35–49. doi:10.1518/0018720053653802.

Molloy, R., and R. Parasuraman. 1996. "Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 38(2):311–322. doi:10.1177/001872089606380211.

Murrell, G. A. 1977. "Combination of Evidence in a Probabilistic Visual Search and Detection Task." *Organization Behavior and Human Performance* 18(1):3–18. doi:10.1016/0030-5073(77)90015-0.

Parasuraman, R. 1987. "Human-Computer Monitoring." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 29(6):695–706. doi:10.1177/001872088702900609.

Parasuraman, R., and D. H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human factors 52* (3):381–410.

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

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

Rock, I., and S. Palmer. 1990. "The Legacy of Gestalt Psychology." *Scientific American* 263(6):84–91.

Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision-Making in a Simulated Command and Control Task." *Human Factors* 49(1):76–87.

Rovira, E., R. Pak, and A. McLaughlin. 2017. "Effects of Individual Differences in Working Memory on Performance and Trust with Various Degrees of Automation." *Theoretical Issues in Ergonomics Science* 18(6):573–591. doi:10.1080/1463922X.2016.1252806.

Samn, S. W., and L. P. Perelli. 1982. *Estimating Aircrew Fatigue: A Technique with Application to Airlift Operations (ADA125319)*. Brooks Air Force Base, TX: USAF School of Aerospace Medicine.

Schmidtke, H. 1966. "Untersuchungsziele." In *Leistungsbeeinflussende Faktoren im Radar-Beobachtungsdienst*, edited by H Schmidtke, 7–9. Wiesbaden: VS Verlag für Sozialwissenschaften.

Sebok, A., and C. D. Wickens. 2017. "Implementing Lumberjacks and Black Swans into Model-Based Tools to Support Human–Automation Interaction." *Human* Factors*:* The Journal of the Human Factors and Ergonomics Society 59(2):189–203. doi:10.1177/0018720816665201.

Sibley, C., J. Coyne, and J. Thomas. 2016. "Demonstrating the Supervisory Control Operations User Testbed (SCOUT)." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60(1):1324–1328. doi:10.1177/1541931213601306.

Unmanned Aircraft Systems Roadmap 2005–2030. 2005. Office of the Secretary of Defense, United States of America.

Walliser, J. C., E. J. de Visser, and T. H. Shaw. 2016. "Application of a System-Wide Trust Strategy when Supervising Multiple Autonomous Agents." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60(1):133–137. doi:10.1177/1541931213601031.

Wickens, C., S. Dixon, J. Goh, and B. Hammer. 2005. "Pilot Dependence on Imperfect Diagnostic Automation in Simulated UAV Flights: An Attentional Visual Scanning Analysis." Presented at the 13th International Symposium on Aviation Psychology.

Wickens, C. D., B. L. Hooey, B. F. Gore, A. Sebok, and C. S. Koenicke. 2009. "Identifying Black Swans in NextGen: Predicting Human Performance in off-Nominal Conditions. *Human*." *Human* Factors*:* The Journal of the Human Factors and Ergonomics Society 51(5):638–651. doi:10.1177/0018720809349709.

Wickens, C. D., and J. S. McCarley. 2008. *Applied Attention Theory*. Boca Raton: CRC Press.