Using System-Wide Trust Theory to Reveal the Contagion Effects of Automation False Alarms and Misses on Compliance and Reliance in a Simulated Aviation Task

Kasha Geels-Blair a, Stephen Rice a & Jeremy Schwark a

a Department of Psychology, New Mexico State University, Las Cruces, New Mexico

Published online: 28 Jun 2013.

To cite this article: Kasha Geels-Blair, Stephen Rice & Jeremy Schwark (2013) Using System-Wide Trust Theory to Reveal the Contagion Effects of Automation False Alarms and Misses on Compliance and Reliance in a Simulated Aviation Task, The International Journal of Aviation Psychology, 23:3, 245-266, DOI: 10.1080/10508414.2013.799355

To link to this article: http://dx.doi.org/10.1080/10508414.2013.799355
Using System-Wide Trust Theory to Reveal the Contagion Effects of Automation False Alarms and Misses on Compliance and Reliance in a Simulated Aviation Task

Kasha Geels-Blair, Stephen Rice, and Jeremy Schwark

Department of Psychology, New Mexico State University,
Las Cruces, New Mexico

System-wide trust theory (Keller & Rice, 2010) suggests that automated aids associated with multiple independent gauges tend to be treated as one system. Meyer’s (2001, 2004) compliance–reliance model indicates that false-alarm-prone and miss-prone automated aids affect operator behavior differently. This study integrates system-wide trust theory with Meyer’s compliance–reliance model. Participants monitored 8 gauges, each augmented by an automated aid. Aid 1 was either 100% or 70% reliable (either false alarm- or miss-prone), whereas the other aids were perfectly reliable. Participants generally employed a system-wide trust strategy, but this effect was stronger for false alarms compared to misses.

Aviation psychologists tend to make predictions about pilot performance using existing theories from cognitive psychology. For example, diagnostic automated aids associated with a task typically assist the operator by reducing cognitive load (Wickens & Dixon, 2007). Unfortunately, this added benefit can bring about potential negative consequences. A diagnostic automated aid in an unmanned aerial vehicle (UAV) setting can be helpful to pilots by providing recommendations, but pilots might choose not to use the aid if their level of trust in the aid is reduced. The reduction of trust in a single automated aid has been widely studied, but the impact of reduced trust from one aid to multiple aids has only been recently explored.

We have chosen system-wide trust (SWT) theory (Keller & Rice, 2010) and Meyer’s (2001, 2004) model of compliance and reliance as the foundation for
making predictions about contagion effects in a diagnostic automation task that required participants to monitor eight gauges for system failures. Each gauge was supported by an independent diagnostic aid. The aid corresponding to the leftmost gauge (Gauge 1) was either 100% or 70% reliable and failed by producing either false alarms (FAs) or misses. Gauges 2 through 8 were always augmented by 100% reliable aids. The gauges were intended to simulate what might be encountered when interacting with UAVs to understand the effects that an unreliable automated aid could have on other nearby automated aids in an aviation setting.

In the following sections, we discuss the previous literature regarding trust, SWT theory, and the compliance–reliance model. Knowledge of the previous literature from these three separate areas is essential to understand the foundation for the current methodology. We then introduce the merging of SWT theory with the compliance–reliance model and, finally, present our subsequent hypotheses for this study to predict the effects of an unreliable automated aid on nearby automated aids in a simulated aviation task.

**TRUST IN AUTOMATION**

Trust, as a construct in social psychology, refers to the predictability of another person (Deutsch, 1958; Eckel & Wilson, 2004; Ergeneli, Saglam, & Metin, 2007). For example, having a high level of trust in someone means that you are fairly confident you can predict that another person will do what you expect (and, generally, what you expect that person to do will be positive). Research demonstrates that humans interact with machines similarly to how they interact with other humans (Reeves & Nass, 1996), which allows the construct of trust to be applied to human–machine interaction and, more specifically, interactions with automation.

Numerous studies have been conducted to examine how trust impacts automation use in various fields, including aviation (Dixon & Wickens, 2006; Dixon, Wickens, & Chang, 2005; Dixon, Wickens, & McCarley, 2007; Geels, Rice, Schwark, & Sandry, 2011; Lee & Moray, 1994; Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2000; Rice, 2009; Rice, Trafimow, Clayton, & Hunt, 2008; Wiegmann, Rich, & Zhang, 2001). The general consensus is that an operator’s trust in an automated aid and his or her use of the automated aid is positively correlated. An operator is more likely to rely on an automated aid if he or she has a high level of trust in that aid.

Response times (RTs), agreement rates, and self-reports are the most common measures of operator trust (Dixon & Wickens, 2006; Dixon et al., 2005; Dixon et al., 2007; Geels et al., 2011; Keller & Rice, 2010; Rice, 2009; Rice & Geels, 2010; Rice et al., 2008; Wickens & Colcombe, 2007). An operator will make a faster response in a task when the automated aid makes few mistakes. The faster response most likely stems from the idea that operators do not need to spend time performing the task themselves because trust levels are high and the aid can be
used to assist with the operator’s decision. Agreement rates with the automation are used to measure trust due to the previously mentioned large, positive correlation between trust (measured by various other constructs) and automation use. Self-reports might also serve as a measure of trust, but can be more subjective. For this study, we use objective measures of RTs and agreement rates to determine the effects on trust. Arguments against the use of these two measures exist (Dekker & Hollnagel, 2004; Dekker & Woods, 2002), but we use these measures in this study because they have been previously accepted as measures of trust (Lee & Moray, 1994; Parasuraman, Sheridan, & Wickens, 2008; Rice, 2009; Rice & Geels, 2010).

The factors that influence operator trust have been of primary interest to automation researchers. Generally, mistakes made by an automated aid cause a decrease in the operator’s trust in that aid (Meyer, 2001, 2004; Rice, 2009). Meyer (2001, 2004) proposed that the two types of mistakes made by automated aids (false alarms and misses) differentially affect operators’ trust in an aid (explained further in the compliance and reliance section of this article). However, more recent studies have shown that both types of mistakes have an overall negative impact on operator trust (Rice, 2009).

**SYSTEM-WIDE TRUST THEORY**

Prior to the introduction of SWT theory, previous research in trust toward automation had focused primarily on the single-aid paradigm (e.g., Rice, 2009), whereby participants were exposed to a single automated aid (e.g., Maltz & Shinar, 2003; Meyer, 2001, 2004), even if the paradigm as a whole involved multiple tasks (e.g., Dixon & Wickens, 2006; Dixon et al., 2007; Lee & Moray, 1994; Parasuraman et al., 1993). Very little was known about how operators would respond to multiple independent tasks augmented by multiple independent aids. Furthermore, a theory that explains what would happen to operator trust and human–automation performance when an imperfect aid was paired with perfectly reliable aids had not yet been produced.

Keller and Rice (2010) proposed a continuum of possibilities for the effects of an imperfect aid on a perfectly reliable aid. At one extreme end of the continuum, component-specific trust predicted that operators would differentially place their trust in the various aids, due to the difference in reliabilities. For example, when exposed to two aids of varying reliabilities, the operator would be able to discern the difference in reliabilities and act accordingly. Performance for each aid would show a rational attempt to calibrate to that aid’s reliability regardless of the other aid’s reliability. On the other extreme end of the continuum, SWT predicted that operators would treat all the aids as part of a “system” and calibrate their trust to the system rather than to each individual aid. If one aid was unreliable, it could...
“pull down” trust in a perfectly reliable aid (or vice versa), which causes contagion from one aid to another (i.e., a system-wide trust effect).

To test their predictions, Keller and Rice (2010) asked participants to perform a simulated unmanned aerial systems (UAS) task, whereby participants manually flew a UAS while concurrently monitoring for system failures in two gauges below the flight display. These gauges were augmented by 100%, 85%, or 70% reliable aids, with the unreliable aid producing only FAs. Participants were free to either agree with the aid or ignore it; that is, the final decision rested with the participant. Performance was measured by accuracy and RTs. The data were clear and definitive. SWT not only best predicted the data, but the contagion effect was quite strong, causing trust in the perfectly reliable aid to drop to levels of the unreliable aid.

Two later studies (Geels et al., 2011; Rice & Geels, 2010) followed up the first study with a more thorough test of SWT by requiring participants to perform a similar task. Participants in these studies had to interact with four gauges rather than two, thus increasing the ratio of perfect to imperfect aids from 1:1 to 3:1. Instead of using the dual-task paradigm adopted by Keller and Rice (2010), the follow-up studies used a single-task paradigm to more efficiently control for any potential dual-task confounds. The amount of information given to the participants about the reliability of the aids was also manipulated to see whether knowledge of the aids’ reliabilities and their own performance would affect participants’ trust levels. Misses were used instead of FAs to determine whether a system-wide effect occurs with both types of errors. Also, agreement rates and RTs to the automated aids—rather than accuracy—were measured to show effects on operator trust. Data supported findings from the previous study. SWT best predicted the data in almost all cases. Even when participants were provided with detailed information about the reliability of each automated aid, participants still succumbed to the effects of SWT throughout all four aids.

COMPLIANCE AND RELIANCE

In the past decade, the differential effects of automation FAs and misses on operator trust and performance have been thoroughly examined (e.g., Bainbridge, 1983; Lee & Moray, 1994; Lee & See, 2004; Meyer, 2001, 2004; Parasuraman & Riley, 1997; Parasuraman et al., 2000). Meyer’s (2001, 2004) model of compliance and reliance garnered much deserved attention for proposing the distinction between how the two types of errors differentially affect operator behaviors. According to the model, automation FAs tend to cause a reduction in compliance, where compliance is defined as behaviors toward automation alerts. Conversely, automation misses tend to cause a reduction in reliance, where reliance is defined as behaviors toward automation nonalerts, or quiet periods when the automation is silent. A strict form of this model would assume that FAs only affect compliance and
misses only affect reliance, with no crossover between the two types of errors or behaviors. Meyer pointed out that other contributing factors (e.g., experience, alert sensitivity, and criterion setting) along with error type could prevent a strict model from always occurring.

Subsequent data show that a more liberal model of compliance—reliance might be more appropriate. Dixon and Wickens (2006) had participants fly a simulated UAV mission while monitoring system failures and found that FAs not only affected compliance, but they also affected reliance. Admittedly, the data were sparse due to the inherent real-world nature of the task, so they followed this up with a more controlled laboratory test where they found very strong and convincing data that FAs reduced both compliance and reliance (Dixon et al., 2007). Interestingly, the authors noted that in both studies, automation misses only appeared to negatively affect reliance and not compliance.

Dixon et al. (2007) noted that these two studies used a dynamic multiple-task paradigm that resulted in FAs being more perceptually and cognitively salient; that is, participants were likely to have noticed the FAs more easily than the misses. For example, during the 30-sec window that the system gauge was fluctuating, if an FA sounded, then the operator would immediately notice it. However, because nonalerts were actually periods of silence from the automation, misses would go unnoticed until the end of the trial when the participants were given feedback, assuming they even were able to process the feedback appropriately because they were still performing the primary task of flying the UAV. A later study using detection of weapons in a luggage screening task (Rice & McCarley, 2011) also found an asymmetry for the two errors on compliance and reliance, even when alerts and nonalerts were matched on their salience. It was concluded in both studies that perceptual framing—meaning the perceptual differences between FAs and misses—could have contributed to the differential treatment of alerts and nonalerts by the operator.

Rice and McCarley (2011) proposed that the differential effects between the two errors might also be due to the difference in cognitive salience between the errors when operators come across them. Generally, FAs are more salient than misses, and if it is important to keep safe and failure readings equivalent in saliency, designers must take this observation into account. Even when FAs and misses were equated on their levels of salience, FAs still tend to have a more negative impact on trust. This might be because FAs and misses generally differ in cognitive saliency for the operator.

To equalize the negative perceptual and cognitive effects of FAs and misses, Rice (2009) used a single-task paradigm whereby participants searched aerial photographs of Baghdad for enemy weapons. Prior to each image, a diagnostic aid also scanned the image and gave a recommendation as to whether or not the aid detected a weapon. The aid was either FA-prone or miss-prone, and varied in reliability from 100% to 55%, in 5% increments. Participants were then able to
view the image as long as they needed to make a determination. The final decision rested with the participant. The data revealed that both types of automation errors negatively impacted compliance and reliance. Automation FAs caused a large reduction in compliance, with a smaller, but still significant effect on reliance, whereas misses produced the reverse effect. This was the first study of its kind to show that automation misses also affected compliance.

The data from Rice (2009) have been supported using a different philosophy. The confluence theory (Trafimow, 2009), taken from social psychology and recently applied to human–automation interaction (Rice, Trafimow, Keller, & Bean, 2012), would suggest that these seemingly orthogonal processes (the effect of FAs and misses on both compliance and reliance) are actually related through a general trust of the automation itself. As general trust fluctuates, so does trust in both alerts and nonalerts. FAs have an impact on general trust, which leads to an effect not only on compliance, but on reliance as well. Misses also have an effect on general trust, so that leads to their effect on compliance.

MERGING SYSTEM-WIDE TRUST THEORY AND THE COMPLIANCE–RELIANCE MODEL

The studies just described have not yet examined the effects of automation FAs and misses on operator trust and performance across multiple aids. One of the primary goals of this study was to do exactly that by manipulating the bias of the automation so that it produced either FAs or misses in one aid to closely examine the contagion effect on the other aids. This study is the first to combine SWT (Geels et al., 2011; Keller & Rice, 2010; Rice & Geels, 2010) with the compliance–reliance model (Meyer, 2001, 2004; Rice, 2009). The two are combined by measuring the contagion effect of an unreliable automated aid as it spreads to seven nearby 100% reliable aids, which are monitored simultaneously. Not only are we able to see the effects of imperfect automation on multiple 100% reliable aids, but we can also see whether contagion effect differences exist between an automated aid producing FAs compared to an aid that produces misses.

Although Rice and Geels (2010) briefly discussed the model of compliance and reliance (Meyer, 2001, 2004), the effects of FAs and misses on multiple automated aids have not yet been examined while also applying SWT theory. SWT predicts that there will be a contagion effect from the unreliable aid to each of the 100% reliable aids. As described previously, the strength of the effect and the type of behavior (compliance or reliance) that is affected might differ depending on whether the aid produces FAs or misses. Although Meyer (2001, 2004) stated that FAs tend to strictly affect compliance while having no effect on reliance, Rice (2009) found a strong effect on compliance and a small effect on reliance. The reverse was also found for misses.
We hope to find stronger support for these data while applying the findings toward multiple automated aids using a single-task paradigm; that is, a paradigm in which participants are given a single task that must be performed on each of the eight gauges. Rice (2009) used a single-task paradigm to compare the effects of FAs and misses, but did not use a multiple automated aid paradigm, and therefore could not determine whether a contagion effect would exist with multiple automated aids. The previous SWT studies (Geels et al., 2011; Keller & Rice, 2010; Rice & Geels, 2010) used only one error type, but this study aims to examine both types of errors within the same task.

Keller and Rice (2010) measured accuracy and RTs to determine whether an SWT strategy existed, whereas the other two studies (Geels et al., 2011; Rice & Geels, 2010) used agreement rates to measure overall dependence and RTs to determine participants’ level of trust. In this study, trust and contagion effects will be measured by agreement rates to both alerts and nonalerts (corresponding to compliance and reliance, respectively) and RTs. If SWT occurs, a contagion effect of Gauge 1 (the unreliable automated aid) exists.

**THIS STUDY**

The purpose of this study was to merge the theoretical contributions of SWT theory and Meyer’s compliance–reliance model. To do this, we had participants perform a simulated gauge-monitoring task, where they were responsible for monitoring eight gauges for system failures. Each of the gauges was augmented by an independent diagnostic aid that provided a safe or unsafe notification during each trial for each gauge. The aids operated independently of one another. Gauge 1 was either 100% or 70% reliable, and in the latter case, was either FA-prone or miss-prone. Gauges 2 through 8 were always 100% reliable. The following hypotheses were generated:

Hypothesis 1: For the gauge augmented by imperfect automation (Gauge 1), we predicted that automation FAs and misses would affect both compliance and reliance behaviors in the same manner found in Rice (2009).

Hypothesis 2: When tasked with monitoring eight gauges, we predicted that participants would still employ a system-wide trust strategy, despite the increase in the ratio of perfect to imperfect aids. In the original Keller and Rice (2010) study, the percentage of imperfect aids was 50%. This was later reduced to 25% (Geels et al., 2011; Rice & Geels, 2010) and to 12.5% in this study. If contagion from one aid could spread to seven other perfectly reliable aids, then this would provide the strongest evidence to date that SWT is a pervasive force when dealing with multiple aids.
Hypothesis 3: We predicted that the contagion seen in Gauges 2 through 8 would be differentially affected by FAs and misses.

Hypothesis 3a: Automation FAs in Gauge 1 would have a strong effect on compliance in Gauges 2 through 8, with a weaker effect on reliance.

Hypothesis 3b: Automation misses in Gauge 1 would have a strong effect on reliance in Gauges 2 through 8, with a weaker effect on compliance.

METHOD

Participants

One hundred thirty-three undergraduates (84 female, 49 male, $M_{age} = 20.27$, $SD = 4.75$) from a large Southwestern university participated in the experiment for partial course credit. All participants were tested for normal or corrected-to-normal vision and colorblindness.

Materials and Stimuli

The experimental display was presented on a Dell computer with a 3.3 GHz processor and a 22-in. monitor, using $1024 \times 768$ resolution.

The experimental display simulated a task that would be encountered when interacting with UAVs. The gauge display consisted of eight gauges in a row from left to right, each of which presented eight randomly assigned values (see Figure 1). A reading on the gauges was determined by the position of three “hands.” The short wide hand represented units of 1,000, the medium hand represented units of 100, and the longest thin hand represented units of 10. Note that these displays were not meant to be designed for efficiency, but instead were meant to represent the types of displays found in UAV control settings that often require heavy cognitive resources. By making the displays difficult, we were able to more effectively test our theoretical points (Dixon et al., 2007).

Above each gauge was a range indicator that provided information about an ideal value and range. These values were also randomly assigned per trial, using four- and three-digit values. For example, the indicator might read 1,330 (550). This meant that the ideal value was 1,330 and the safe range was 1,330 ± 550 (780 to 1,880).

Below each gauge was an indicator that provided the recommendation of the diagnostic aids for each gauge. This indicator was either green for “Safe” or red for “Failure,” depending on the recommendation. The leftmost aid varied in reliability from either 100% to 70% reliable. Depending on the condition, the errors were either FAs (incorrectly recommending “Failure” when no failure had occurred) or misses (incorrectly recommending “Safe” when a failure had occurred). Participants were told prior to beginning the experiment how reliable
FIGURE 1  A sample display from an experimental trial. See text for a description of the display components (color figure available online).
the aid was and whether it was prone to FAs or misses. They were also told that all the aids were independent of each other and that the other seven aids were always 100% reliable. The position of the unreliable aid remained constant to test whether (a) participants always responded to the unreliable aid first, and (b) the contagion effect would be affected by distance of the reliable aids from the imperfect aid. A baseline condition was employed whereby no automation recommendations were given for any of the aids.

Thus, there were four conditions. In Condition 100, all the aids were 100% reliable. In Condition 70FA, the leftmost aid was 70% reliable with FAs (all other aids were 100% reliable). In Condition 70M, the leftmost aid was 70% reliable with misses (all other aids were 100% reliable). The baseline condition did not have an automated aid, so it was not included in our analyses, as discussed later. Participants were randomly placed into each condition.

Procedure

Participants first signed consent forms and were then seated comfortably in a chair facing the experimental display. Viewing distance was controlled for by the use of a chin rest, centered approximately 21 in. from the display. Scripted verbal instructions were provided by the experimenter and participants were informed that they could ask questions before beginning the experiment. Five practice trials were presented with unlimited time for each trial to allow participants ample time to become familiar with the task.

At the beginning of each trial, the task display presented the gauge information for 20 sec, during which time participants were asked to determine if the value in each gauge fell within the safe range. Responses were made by mouse-clicking on the appropriate key for Safe or Failure. Participants could agree with the automation if desired, but were told the final choice rested with them. Once the participant had completed all eight choices, he or she pressed a Continue button at the bottom right of the screen.

Feedback was provided in the form of red or green displays that outlined each of the safe–failure choice boxes, such that participants could see their accuracy for each gauge. A red outline meant that the participant made an incorrect choice, whereas a green outline meant that the participant made a correct choice. This feedback remained present for 5 sec. Participants completed 50 trials in each condition. The experiment lasted approximately 40 min. On completion of the experiment, participants were debriefed and dismissed.

Design

A two-way mixed factorial design was employed with Aid/Gauge (Aid for the agreement rates and Gauge for the RTs) and Condition as factors. Aid/Gauge
consisted of eight levels that represented each of the eight automated aids or gauges. As mentioned earlier, only one gauge was accompanied by an unreliable aid, and the other seven gauges were accompanied by perfectly reliable aids. Condition consisted of three levels that represented the three different automation groups in which participants were randomly assigned (as discussed previously in the stimuli section).

RESULTS

This section is divided into two parts. First, the data from Aid 1 are presented to test the first hypothesis. Second, a discussion of the agreement rate data from all automated aids as a function of whether or not the automation provided an alert to test the second and third hypotheses are presented. Aid refers to each automated aid that corresponds to each gauge, and conditions refers to each of the three conditions in which participants interacted with an automated aid. Baseline data could not be included in the analyses because of the inability to gather agreement rate data from a condition that had no automation.

Gauge 1 Performance

The analyses for Gauge 1 are presented to test Hypothesis 1. To reiterate, Hypothesis 1 states that FAs and misses will affect both compliance and reliance in Gauge 1. Previously, Rice (2009) had shown that FAs can have a strong negative impact on compliance, with a weaker impact on reliance, and that misses will have a mirrored effect on reliance and compliance respectively for an unreliable automated aid.

Agreement rates. Agreement rates were measured by how often the participant agreed with the automated aid. These were broken down between alerts and nonalerts for Aid 1 to show the effects on compliance and reliance and are presented in Figure 2. A two-way analysis of variance (ANOVA) using condition and alert or nonalert as factors revealed a main effect of Condition, \( F(2, 111) = 72.97, p < .001, \eta^2 = .57 \), indicating that performance in the 100% reliable condition was superior to both the imperfectly reliable conditions, \( t(111) = 11.99, p < .001 \). There was also a significant Condition × Alert/Nonalert interaction, \( F(2, 111) = 10.10, p < .001, \eta^2 = .15 \).

Of particular interest was the significant interaction between the 70F and 70M conditions (after removing the 100% condition data) and the alert–nonalert factor, \( F(1, 74) = 13.51, p < .001, \eta^2 = .15 \). This revealed that when the aid provided alerts, compliance suffered the most for the 70F condition compared to the 100 condition, \( p < .001 \), but also still suffered significantly for the 70M condition, \( p < .001 \). In contrast, when the aid provided nonalerts, reliance suffered the most
for the 70M condition compared to the 100 condition, \( p < .001 \), but still suffered significantly for the 70F condition, \( p < .001 \). This crossover effect, whereby FAs hurt compliance but still affect reliance, and vice versa, is consistent with the data from the Rice (2009) study and Hypothesis 1.

**Response times.** RTs were measured for the responses to each gauge separately, as participants evaluated and responded to each gauge one at a time. RTs were collected starting from the time the participant began the trial to when the participant responded to any gauge. After the first response the clock was reset for each gauge afterward, so that RTs only reflected the time between responding to each gauge.

A two-way ANOVA using condition and alert–nonalert as factors revealed a main effect of condition, \( F(2, 111) = 14.11, p < .001, \eta^2 = .20 \), indicating that responses in the 100% reliable condition were superior to both the imperfectly reliable conditions, \( t(111) = 5.28, p < .001 \). There was also a main effect of alert–nonalert, \( F(1, 111) = 6.18, p < .05, \eta^2 = .05 \), indicating faster responses to nonalerts than alerts. A significant Condition \( \times \) Alert–Nonalert interaction was also found, \( F(2, 111) = 5.77, p < .01, \eta^2 = .09 \); however, \( t \) tests conducted between the 70F and 70M conditions found no differences between either of the conditions for both alerts and nonalerts (both \( ps > .10 \)).

Because agreement rates differed significantly between the 70F and 70M conditions, but RTs did not, any concern about a speed–accuracy trade-off was alleviated.

**Contagion Effects**

According to Hypothesis 2, the unreliable automated aid should reduce trust throughout the other seven automated aids, as seen in previous SWT studies.
(Geels et al., 2011; Keller & Rice, 2010; Rice & Geels, 2010). Although the ratio of unreliable to reliable aids was higher in previous studies, it is predicted that SWT will still take place even with an increased amount of automated aids. Hypothesis 3 predicts that SWT will impact compliance and reliance, which has only been previously observed for a single automated aid. FAs should have a strong negative effect on compliance and a weaker effect on reliance throughout each of the automated aids. Misses, on the other hand, should have a strong negative effect on reliance, with a weaker effect on compliance for all of the automated aids.

**Agreement rates to alerts (compliance).** Compliance was measured by how often the participant agreed with the automated aid when it provided an alert (Dixon & Wickens, 2006; Dixon et al., 2007; Keller & Rice, 2010; Rice, 2009). Figure 3a presents these data. A two-way ANOVA using condition and aid as factors revealed a main effect of condition, $F(2, 111) = 10.53, p < .001, \eta^2 = .16$, a main effect of aid, $F(7, 777) = 59.17, p < .001, \eta^2 = .35$, and a significant Condition $\times$ Aid interaction, $F(14, 777) = 17.12, p < .001, \eta^2 = .24$, indicating

![Agreement Rates - Compliance](image)

![Response Times - Compliance](image)

**FIGURE 3** Agreement rates and response times during automation alerts (compliance). Standard error bars included.
that the drop in agreement rates from the 100% reliable condition to the other imperfect conditions was not equal across all aids.

To further test Hypotheses 2 and 3, further analysis was done on the reliable aids only to determine whether the conditions with the imperfect Gauge 1 aid caused contagion in the perfectly reliable aids. A two-way ANOVA using condition and aids 2 through 8 as factors revealed a main effect of condition, $F(2, 111) = 3.27, p < .05, \eta^2 = .06$, indicating that there was a drop in agreement rates for those aids in the 70F condition, $F(1, 74) = 4.29, p < .05, \eta^2 = .06$, and in the 70M condition, $F(1, 74) = 4.06, p < .05, \eta^2 = .05$, compared to the 100 condition. There was also a main effect of aid, $F(6, 666) = 9.56, p < .001, \eta^2 = .08$. The Condition × Aid interaction was marginally significant, $F(12, 666) = 1.76, p = .051, \eta^2 = .03$. A post-hoc comparison using the Fisher least significant difference (LSD) test revealed that the 100 condition had higher agreement rates than both the 70F and 70M conditions and the 70F condition had lower agreement rates than the 70M condition.

**Response times to alerts (compliance).** Figure 3b presents these data. A two-way ANOVA using condition and gauge as factors revealed a main effect of condition, $F(2, 111) = 4.75, p = .01, \eta^2 = .08$; a main effect of gauge, $F(7, 777) = 23.54, p < .001, \eta^2 = .18$; and a significant Condition × Gauge interaction, $F(14, 777) = 5.73, p < .001, \eta^2 = .09$. These data mirror those of the agreement rates.

Again, further analysis was done on the reliable aids only to test further Hypotheses 2 and 3. A two-way ANOVA using condition and Gauges 2 through 8 as factors revealed a significant main effect of condition, $F(2, 111) = 3.21, p < .05, \eta^2 = .06$, indicating that there was a drop in performance for those gauges in the unreliable conditions compared to the 100 condition. There was a significant main effect of gauge, $F(6, 666) = 9.10, p < .001, \eta^2 = .08$, and a marginally significant Condition × Gauge interaction, $F(12, 666) = 1.75, p = .053, \eta^2 = .03$. A post-hoc comparison using the Fisher LSD test revealed that the 70F condition had longer response times than both the 70M and 100 conditions.

**Agreement rates to nonalerts (reliance).** Reliance was measured by how often the participant agreed with the automated aid when it provided a nonalert (Dixon & Wickens, 2006; Dixon et al., 2007; Geels et al., 2011; Rice, 2009; Rice & Geels, 2010). Figure 4a presents these data. A two-way ANOVA using condition and aid as factors revealed a main effect of condition, $F(2, 111) = 8.02, p = .001, \eta^2 = .13$; a main effect of aid, $F(7, 777) = 53.68, p < .001, \eta^2 = .33$; and a significant Condition × Aid interaction, $F(14, 777) = 14.55, p < .001, \eta^2 = .21$, indicating that the drop in agreement rates from the 100% reliable condition to the other imperfect conditions was not equal across all aids.

Again, further analysis was done on the reliable aids only to test Hypotheses 2 and 3. A two-way ANOVA using condition and aids 2 through 8 as factors
revealed a marginally significant main effect of condition, $F(2, 111) = 2.65, p = .075, \eta^2 = .05$, indicating that there was a drop in agreement rates for those aids in the 70F condition (marginally significant), $F(1, 74) = 3.82, p = .054, \eta^2 = .05$; and in the 70M condition, $F(1, 74) = 4.47, p < .05, \eta^2 = .06$, compared to the 100 condition. There was a significant main effect of aid, $F(6, 666) = 6.71, p < .001, \eta^2 = .06$. The Condition $\times$ Aid interaction was not significant, $F(12, 666) = 1.47, p > .10, \eta^2 = .03$.

**Response times.** Figure 4b presents these data. A two-way ANOVA using condition and gauge as factors revealed a marginally significant main effect of condition, $F(2, 111) = 2.47, p = .08, \eta^2 = .04$; a significant main effect of gauge, $F(7, 777) = 22.25, p < .001, \eta^2 = .17$; and a significant Condition $\times$ Gauge interaction, $F(14, 777) = 5.02, p < .001, \eta^2 = .08$. These data mirror those of the agreement rates.
Again, further analysis was done on the reliable aids only to test Hypotheses 2 and 3. A two-way ANOVA using condition and Gauge 2 through 8 as factors revealed a significant main effect of gauge, $F(6, 666) = 9.57, p = .001, \eta^2 = .08$; but no main effect of condition, $F(2, 111) = 1.75, p > .10, \eta^2 = .03$; or Condition $\times$ Gauge interaction, $F(12, 666) = 1.32, p > .10, \eta^2 = .02$.

DISCUSSION

The purpose of this study was threefold. First, the findings of Rice (2009) could possibly be extended to a multiple automation paradigm; that is, the Rice study showed that automation FAs strongly affected compliance with weaker, but significant effects on reliance, and vice versa for automation misses. Whereas the Rice study was performed using only a single automated aid, this study used multiple automated aids to examine whether a contagion effect exists. Second, it was predicted that participants would adopt a SWT strategy despite the increase in the number of automated aids from two (Keller & Rice, 2010) to four (Geels et al., 2011; Rice & Geels, 2010) and now eight, resulting in a contagion effect on the perfectly reliable aids. If a SWT strategy still takes place when the unreliable aid comprises 12.5% of the system, this study would be the strongest test of SWT to date. Third, and most important, the ultimate goal was to reveal the contagion effects of automation FAs and misses on adjacent aids by experimentally manipulating whether the imperfect aid was FA-prone or miss-prone. This could only be done by combining SWT and the compliance–reliance model because the previous studies had either used a single aid when comparing the effects of FAs and misses or examined the effect of only one error type on multiple automated aids.

The first hypothesis was confirmed by the data. In the Rice (2009) study, the data revealed that FAs strongly affected compliance with a weaker effect on reliance, whereas misses strongly affected reliance, with weaker effects on compliance. That was the first study to date to find that both types of automation errors affect both types of compliance–reliance behaviors. This crossover effect appeared to be quite strong in the current data as well. When analyzing the results from Aid 1 only (the imperfect aid), FAs clearly have a devastating effect on compliance (agreement rates to automation alerts), as well as a surprisingly strong negative effect on reliance. The data for automation misses mirror the data for FAs; that is, misses have a devastating effect on reliance with a surprisingly strong negative impact on compliance. This replication of the Rice study using a new paradigm with multiple gauges provides compelling evidence that the previously proposed multiple process model may be generalized to other paradigms.

With regard to the second hypothesis, these data indicate that the 70F condition caused contagion for both compliance and reliance measures. For compliance measures, the effect was noticeable in both agreement rates and RT. For the
reliance measures, it was noticeable in the agreement rates. For the 70M condition, the data also indicate contagion for both compliance and reliance (agreement rates). Participants clearly failed to use a strict component-specific trust strategy. On the other hand, they did not appear to use an extreme SWT strategy either. There was some contagion, and thus the interpretation of the data is that participant strategies fell somewhere along the continuum between the two extremes.

This was despite the increase in the number of gauges from prior studies (Geels et al., 2011; Keller & Rice, 2010; Rice & Geels, 2010). In the Keller and Rice (2010) study, participants were exposed to two aids, so in effect, 50% of the “system” was imperfect. The contagion effect of SWT was extremely powerful; participants did just as poorly with the 100% aid as they did with the less reliable aids. In the Rice and Geels (2010) and Geels et al. (2011) studies, one of the four gauges was imperfect, so in effect, 25% of the system was imperfect. The same strong contagion effect was seen in agreement rates. In this study, despite only one out of eight aids (12.5%) being imperfect, there was still a noticeable contagion effect, which extended to many of the aids, despite their distance from the imperfect aid. In fact, in some cases, the furthest aids from the imperfect aid suffered the most, discounting the notion that perhaps increasing the number of reliable aids might alleviate the contagion effect. Thus, the data reported here provide the strongest evidence to date that SWT is a powerful force when interacting with multiple gauges and aids.

The third hypothesis consisted of two parts. Hypothesis 3a predicted that if FAs negatively affected compliance (strongly) and reliance (weakly) in Gauge 1, they would also do so for Gauges 2 through 8, and vice versa for misses. However, what is surprising is that the crossover effect seen in the Gauge 1 data does not appear to exist in the contagion data. The FA condition was expected to cause more contagion in the compliance measures than did the miss condition—which were found to some degree—and the miss condition to cause more contagion in the reliance measures than did the FA condition—which did not occur. Apparently, although the crossover effect is strong for the unreliable aid itself, this crossover effect does not transfer well when contagion takes place, and FAs appear to have a more deleterious effect on adjacent reliable aids than do misses. It should be noted that none of the previous SWT studies directly manipulated the differential effects of FAs and misses, and thus this study is the first to experimentally test these differential effects on compliance and reliance behaviors.

THEORETICAL CONTRIBUTIONS

This study offers several theoretical contributions. First, the results from this study extend the findings of Rice (2009) regarding the crossover effects of automation
FAs and misses on compliance and reliance. Rice reported that FAs strongly influenced compliance, while having a lesser, yet significant effect on reliance and vice versa for automation misses. The data reported here follow this same trend and allow these crossover effects to be generalized to multiple paradigms.

Second, the overall data—agreement rates and RTs—provide the strongest evidence to date that SWT is pervasive across multiple aids, particularly for FAs. Despite the fact that only one of eight aids (12.5% of the system) was imperfectly reliable, participants still were unable to interact with the system via a component-specific strategy of treating each aid independently. Combined with the results reported by Keller and Rice (2010) and the two most recent studies (Geels et al., 2011; Rice & Geels, 2010), it is reasonable to assume that SWT might be a common strategy across multiple paradigms.

Third, the contagious effects of automation errors on other perfectly reliable aids can be analyzed using the compliance–reliance model described by Meyer (2001, 2004). The two types of automation errors (FAs and misses) appear to have differential contagion effects on other aids viewed as part of the system. Being able to use the compliance–reliance model in conjunction with SWT theory allows different theoretical predictions to be made as a function of different types of automation errors.

Fourth, the data reported here provide support for the use of a new theory within social psychology, termed the confluence theory, to be applied to automation research (Rice et al., 2011; Trafimow, 2009). The confluence theory is an all-encompassing theory that can be used to predict the same outcomes as the compliance–reliance model and SWT theory. According to the confluence theory, FAs should affect both compliance and reliance because the two behaviors stem from and are affected by a person’s general trust. The data reported here show that, in fact, FAs do impact both behaviors, suggesting that the prediction of the confluence theory is correct in assuming FAs impact both compliance and reliance via a general evaluation of each automated aid. The confluence theory would have also predicted that contagion between automated aids occurs due to a general system-wide evaluation of the entire set. Again, the findings support the predictions of the confluence theory. Although the confluence theory was not used to predict findings in this study, it is important to note its implications here because the theory provides yet another link from social psychology research to automation research.

APPLICATIONS

The practical applications from this study are numerous. First, SWT appears to be a pervasive strategy when dealing with multiple aids, particularly for an FA-prone system. UAS designers should take into consideration that multiple entities might be viewed as part of a system rather than as individual components. Even when
participants were explicitly told the reliabilities of each aid and that all aids were independent of each other, they still treated the aids as part of a system rather than using a component-specific trust strategy.

Second, the results also suggest that caution be used when implementing multiple automated aids in a system. Even automated aids that have extremely low error rates individually would result in an inflated system error rate when multiple automated aids are used concurrently, increasing the likeliness of performance decrements due to SWT. The probability that an aid in the system will make an error increases with each additional aid added to the system. Therefore, it seems advisable to limit automated aids to those necessary for the system, which decreases the chance that contamination will occur due to automation failure.

Third, it is important to understand the differential effects of the two types of automation errors in the system design. If it is important to limit negative impacts on compliance, then designers should avoid an FA-prone aid, even if all the other aids are perfectly reliable. If it is important to limit negative effects of reliance, then the reverse is true, and UAS designers should avoid automation misses in any aid.

Fourth, designers should be aware that the contagion effect of automation FAs on adjacent reliable aids appears to be stronger than that of automation misses. Thus, if an aid must have FAs—for whatever reasons—the ramifications of this decision should be considered before designing a paradigm where operators must deal with multiple aids concurrently.

LIMITATIONS AND FUTURE RESEARCH

Although there are now four studies on SWT that have corroborating data, the ability to generalize to all paradigms is still unavailable. It is still important to consider paradigms with larger numbers of gauges. Despite the fact that the contagion effect did not drop off for eight gauges, it is possible that extreme numbers of gauges might produce a contagion drop-off, or result in different types of effects.

A difficult cognitive task was chosen purposely to test SWT and the compliance–reliance model. It might be the case that an easy task might not generate the same contagion effect. Further research should focus on manipulating difficulty to see what differential effects that might have on SWT and contagion. Another problem is that a UAS is unlikely to have only one unreliable automated aid. It could be that two or more automated aids within a UAS are unreliable and can produce either the same error type or even different error types. In this case, the differential effects of the two error types can be examined using both SWT and the compliance–reliance model, as this study has done.

Also, the gauges and automated aids presented on screen were all identical. Having each gauge and automated aid look exactly the same and displaying information within the same parameters could suggest that they are not independent.
Perhaps manipulating the appearance of the gauges could have alleviating effects on SWT, although recent data suggest that at least in one paradigm, it does not (Bean, Rice, & Keller, 2011).

Furthermore, the use of undergraduate students, and not professionals, limits the ability to generalize to a wider population. Further research should include a manipulation of training or assessment of ability to provide greater generalization. Someone who is an expert (i.e., an unmanned aerial systems pilot) at a gauge monitoring task would likely perform quite differently than a nonexpert (i.e., an undergraduate student). It would be interesting to test whether experts or novices are more likely to succumb to a SWT strategy in a future study.

Finally, it is important to note that the use of RTs as a measure of trust can be troublesome. As discussed in the introduction, trust researchers disagree on the best measure of trust in automation. RTs specifically are troublesome because they can be affected by many factors aside from trust, which include workload, situational awareness, and the operator’s self-confidence (Lee & Moray, 1994). Fortunately, agreement rates were also used to measure trust and the comparison of the two measures shows that, generally, the two measures tend to mirror one another.

CONCLUSION

The findings reported here successfully integrated SWT theory (Keller & Rice, 2010) with Meyer’s (2001, 2004) compliance–reliance model as they apply to studies of multiple automated systems. A number of important findings were uncovered, one of the most notable being that SWT, as opposed to component-specific trust, was most often employed when multiple automated aids were present, creating a contagion effect. The contagion effect carried over into the perfectly reliable aids, despite the greater number of perfectly reliable aids. Additionally, findings revealed that FAs might have caused a stronger contagion effect across all perfect aids than did misses. This could only be discovered by merging SWT with the compliance–reliance model. Finally, the effects of misses and FAs on compliance and reliance were generalized to alternate paradigms than those used by Rice (2009). These findings have important implications for the future design of multiple automation systems and will serve to assist researchers in their investigations of these designs in the future.

ACKNOWLEDGMENTS

We wish to thank Joshua Sandry, Gayle Hunt, and David Keller for their helpful comments. Special thanks to Sandra Deming and Laurel Ashley for their help in collecting data and to Chandrasekhar Thotakura for programming the simulator.
REFERENCES


Manuscript first received: January 2013