2013

Applying the Compliance-Reliance Model to System-Wide Trust Theory in an Aviation Task

Casey Tunstall

Kasha Geels-Blair

Stephen Rice

APPLYING THE COMPLIANCE-RELIANCE MODEL TO SYSTEM-WIDE TRUST THEORY IN AN AVIATION TASK

Casey Tunstall
Kasha Geels-Blair
Stephen Rice
New Mexico State University
Las Cruces, New Mexico

Diagnostic automation is frequently used to assist pilots. Despite extensive research on trust in automation, multiple automation research is sparse. Two recent studies focused on effects of trust in multiple automation using system-wide trust (SWT) theory (Keller & Rice, 2010; Rice & Geels, 2010). According to SWT, operators treat multiple aids as one entity, rather than separate entities; an unreliable aid negatively affects trust in nearby aids. We combined SWT with Meyer's (2001, 2004) compliance-reliance model in order to show whether contagion effects of FAs and misses differ. Participants monitored 8 gauges for system failures. Each gauge had an automated aid that provided recommendations. The first aid was 70% reliable, and produced either FAs or misses, while the other 7 aids were 100% reliable. The data indicate that both FAs and misses cause contagion on the perfectly reliable aids, but FAs caused more contagion in both compliance and reliance measures.

It is common for operators to interact with diagnostic aids in the aviation industry. While much research has focused on operator trust in a single aid, very little has analyzed the effects of trust in multiple aids. Keller and Rice (2010) recently proposed a theory called system-wide trust (SWT), whereby operators tend to treat multiple aids as part of a "system" rather than as individual components. When one of the aids is unreliable, it affects trust in the other aids, even when the other aids are reliable and numerous (Keller & Rice, 2010; Rice & Geels, 2010). The purpose of the current study is to analyze the differential effects of automation false alarms (FAs) and misses on SWT. By combining SWT theory with Meyer's (2001, 2004) compliance-reliance model, we hope to show that the contagion effect of FAs and misses on reliable aids are qualitatively different.

System-Wide Trust Theory

Although much research has focused on a single automated aid (Rice, 2009; Maltz & Shinar, 2003; Meyer, 2001, 2004; Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Lee & Moray, 1994; Parasuraman, Molloy, & Singh, 1993), the effect that an unreliable aid would have on nearby automated aids was unknown before the introduction of system-wide trust theory (Keller & Rice, 2010; Rice & Geels, 2010). Keller and Rice (2010) proposed two possibilities when pairing an unreliable automated aid with a perfectly reliable aid. The first possibility, component specific trust (CST), states that participants would treat the automated aids differently according to their level of reliability. The second possibility states that the two automated aids would be treated as a "system" in which trust would be calibrated to both automated aids as one entity. Therefore, if one automated aid is unreliable, it may cause a negative contagion effect that spreads to nearby, perfectly reliable, aids.

Keller and Rice (2010) had participants perform a simulated unmanned aerial systems task while also monitoring for failures in two gauges. The automated aids were 70%, 85%, or 100% reliable, only erring with FAs. Participants could disagree with the automated aids if they wished, but the final decision was up to them. Performance was recorded using response times and accuracy. The data supported the second possibility. System-wide trust took place and participants' trust in the perfectly reliable aid dropped to levels of the unreliable aid.

The findings were further tested in a follow-up SWT study (Rice & Geels, 2010) using four aids and a single-task paradigm (to possibly remove any dual-task confounds). Although three of the aids were 100% reliable, the left-most aid was either 70% or 100% reliable, with the unreliable condition erring by producing only misses. In lieu of using accuracy, agreement rates were used along with the continued use of response times in order to show the effects on trust. The findings supported Keller and Rice (2010); even in conditions where participants had explicit knowledge of the reliability of each automated aid, SWT effects still occurred.

Compliance and Reliance

Much research has compared the effects of FAs and misses on operator trust (e.g. Bainbridge, 1983; Lee & Moray, 1994; Lee & See, 2004; Meyer, 2001, 2004; Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2000). For the current study, we look to Meyer's (2001, 2004) model of compliance and reliance in order to test for differential effects of the two errors on trust in multiple automation. This model predicts that FAs will have a negative effect on compliance, which is defined as behaviors toward alerts, while misses will have a negative effect on reliance, defined as behaviors toward non-alerts. The model predicts that errors will only negatively affect their respective behaviors: false alarms will only affect compliance and misses will only affect reliance.

Three studies (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Rice, 2009) were performed in order to test the compliance-reliance model. The first study (Dixon & Wickens, 2006) had participants perform a simulated unmanned aerial systems mission while comparing the effect of FAs and misses in a systems monitoring task. FAs not only reduced compliance, but also negatively affected reliance. Misses only had a negative effect on reliance. The second study (Dixon, Wickens, & McCarley, 2007) tested the findings using a more controlled laboratory setting. Dixon, Wickens, & McCarley (2007) pointed out that, in these first two studies, the presentation of FAs and misses were not equivalent. FAs were immediately noticed because an alert sounded when the automation detected a failure, while a miss was silent because non-alerts were periods of silence from the automation. This could explain the differential effects between FAs and misses in the first two studies.

In the third study, Rice (2009) alleviated this effect by making FAs and misses equally salient. The dual-task confound was also removed by having participants perform a single visual search task. Participants searched for a helicopter with the help of an automated aid that presented a recommendation before each trial. The automated aid was either FA-prone or miss-prone and varied from 55% to 100% reliable in 5% increments. Both types of errors affected compliance and reliance, with FAs strongly affecting compliance and misses strongly affecting reliance. This was the first study to show that misses can have an effect on compliance.

The Current Study: Merging System-Wide Trust Theory with the Compliance-Reliance Model

Although the previous studies examined either the contagion effects of an unreliable aid on nearby reliable aids (Keller & Rice, 2010; Rice & Geels, 2010) or the effects of FAs and misses on a single aid (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Rice, 2009), the effects of FAs and misses on multiple automated aids have not yet been compared. The current study aims to do just that by examining SWT (Keller & Rice, 2010) using the compliance-reliance model (Meyer, 2001, 2004). By combining the two, we can see the effects of FAs and misses on compliance and reliance, whether a contagion effect still takes place, and whether a difference in contagion to nearby aids exists between FAs and misses.

In this experiment, participants were responsible for monitoring 8 gauges for system failures. Each gauge was augmented by an automated aid that provided recommendations. The first aid was either 70% or 100% reliable, and failures were either FAs or misses, while the remaining 7 aids were always 100% reliable. Trust was measured using agreement rates and response times to both alerts and non-alerts. We hypothesize that 1) FAs and misses will affect both compliance and reliance in the first gauge, as described in Rice (2009), 2) that SWT would still occur, despite the increase in perfectly reliable aids from previous SWT studies (Keller & Rice, 2010; Rice & Geels, 2010), resulting in the strongest test of SWT, and 3) that contagion for the aids on gauges 2-8 would be differentially affected by FAs and misses. FAs should have a stronger effect on compliance with a weaker effect on reliance and misses should have the reverse effect.

Method

One hundred thirty-three (84 females) undergraduates participated for partial course credit; mean age was 20.27 (*SD* = 4.75). All participants were tested for normal or corrected-to-normal vision and were assigned to one of four conditions: Baseline (no automation present), 100% (all 8 automated aids were perfectly reliable), 70M (automation for Gauge 1 was 70% reliable and produced only misses, while all other aids were 100% reliable), or 70F (like 70M, except the 70% reliable aid produced only FAs). Due to the absence of automated aids, the Baseline condition was excluded from analyses. Participants were first seated at a computer and given scripted verbal instructions, followed by 5 practice trials. Once comfortable with the task, participants were told whether there would be automated aids present during the experiment, as well as the reliability of each. Viewing distance was controlled for by using a chinrest situated approximately 21” from the computer monitor.

There were 50 experimental trials which lasted 20 seconds each, followed by a feedback screen, which lasted for 5 seconds. The display consisted of eight simulated gauges with independent random values; each augmented by an automated aid (see Figure 1). Gauges were read by discerning the values given by each of three hands. The short, fat hand represented units of 1000, the long, fat hand represented units of 100, and the long, thin hand represented units of 10. Above each gauge was a range indicator which displayed random 4- and 3-digit values that changed for every trial. To determine the “safe range”, a range indicator that read “1500 (500)” would mean that the safe range is 1500 +/- 500. The automated aids gave a recommendation of whether the corresponding gauge for each was considered safe or if it had failed. The aids were located underneath the gauges and displayed a green box to state that the gauge was safe or a red box to state that it was a failure. Although the automation gave a recommendation, participants were told the final decision rested with them. They made their decision for each gauge by clicking on either the “Safe” or “Failure” options, which were located below each gauge. Once they clicked on an option for all of the gauges, a “Continue” button appeared so that participants could continue to the feedback screen. Feedback was given by outlining the choice display for each gauge. A green outline meant the participant chose correctly and a red outline meant they answered incorrectly. The experiment lasted approximately 40 minutes. Upon completion, participants were debriefed and dismissed.

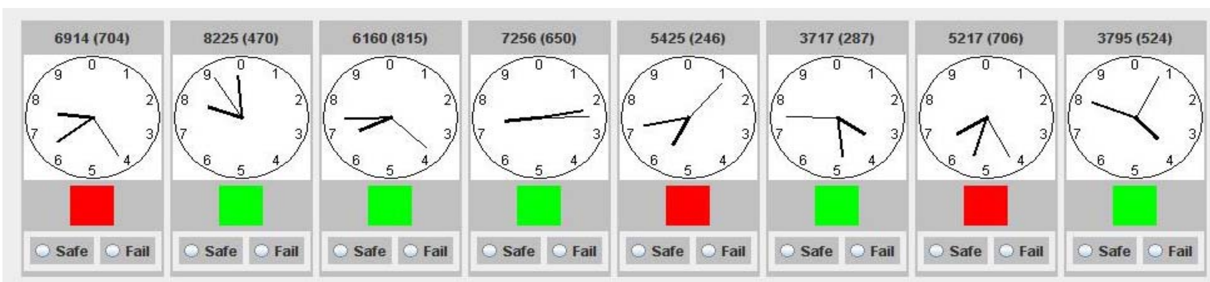

Figure 1. Gauge display. The display was originally presented in color.

Results

Gauge 1 Performance

Agreement Rates. Agreement rates were measured by the percentage in which participants agreed with the automated aid. We split up the agreement rates for alerts and non-alerts, as can be seen in Figure 2. A 2-way ANOVA using Condition and Aid showed a main effect of Condition, $F(2, 111) = 72.97$, $p < .001$, $\eta^2 = .57$, which shows that performance in the 100% condition was superior to the unreliable conditions, $t(111) = 11.99$, $p < .001$. The interaction between Condition and Aid was also significant, $F(2, 111) = 10.10$, $p < .001$, $\eta^2 = .15$. After removing the 100% condition data, we looked at the interaction between the 70F and 70M conditions and the Alert-NonAlert factor, $F(1, 74) = 13.51$, $p < .001$, $\eta^2 = .15$. This showed that for alerts, compliance suffered more in the 70F condition as compared to the 100% condition, $p < .001$, but it also suffered significantly in the 70M condition, $p < .001$. For NonAlerts, reliance suffered most in the 70M condition compared to the 100% condition, $p < .001$, but it also suffered significantly in the 70F condition, $p < .001$. This crossover effect supports Hypothesis 1.

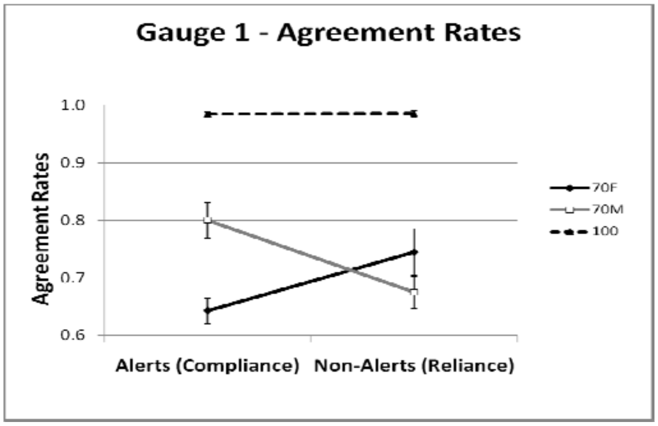

Figure 2. Agreement rates for compliance and reliance for all conditions. SE bars are included.

Response Times. Response times were measured for each gauge since participants responded one at a time. RT was collected starting from the beginning of the experiment until the participant responded to any gauge. After each response, the clock was reset so that RTs only reflected the time between responding to each gauge. A 2-way ANOVA using Condition and Alert-NonAlert as factors showed a main effect of Condition, $F(2, 111) = 14.11$, $p < .001$, $\eta^2 = .20$, which means that responses in the 100% condition was superior to the unreliable conditions, $t(111) = 5.28$, $p < .001$. A main effect of Alert-NonAlert was also present, $F(1, 111) = 6.18$, $p < .05$, $\eta^2 = .05$, which indicates faster RTs to NonAlerts than Alerts. There was also a significant interaction between the two factors, $F(2, 111) = 5.77$, $p < .01$, $\eta^2 = .09$. It should be noted that *t*-tests conducted between the 70M and 70F conditions revealed no differences between the conditions for both alerts and non-alerts (both *p*s > .10).

Contagion Effects

Agreement Rates: Alerts. Compliance was measured using the agreement rate to alerts (Figure 3). Using Condition and Aid as factors, a 2-way ANOVA indicated 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 interaction between the two, $F(14, 777) = 17.12$, $p < .001$, $\eta^2 = .24$. The drop in performance as compared to the 100% condition was not equal across all Aids. Further analysis was done on Aids 2-8 in order to further test Hypotheses 2 and 3. A 2-way ANOVA using Condition and Aids 2-8 as factors showed a main effect of Condition, $F(2, 111) = 3.27$, $p < .05$, $\eta^2 = .06$, this indicated a drop in agreement rates in the 70F condition, $F(1, 74) = 4.29$, $p < .05$, $\eta^2 = .06$, and the 70M condition, $F(1, 74) = 4.06$, $p < .05$, $\eta^2 = .05$, as compared to the 100% condition. The Interaction between Condition and Aid was marginally significant, $F(12, 666) = 1.76$, $p = .051$, $\eta^2 = .03$. Since response times mirrored agreements rates and due to lack of space, the RT analyses will not be reported here.

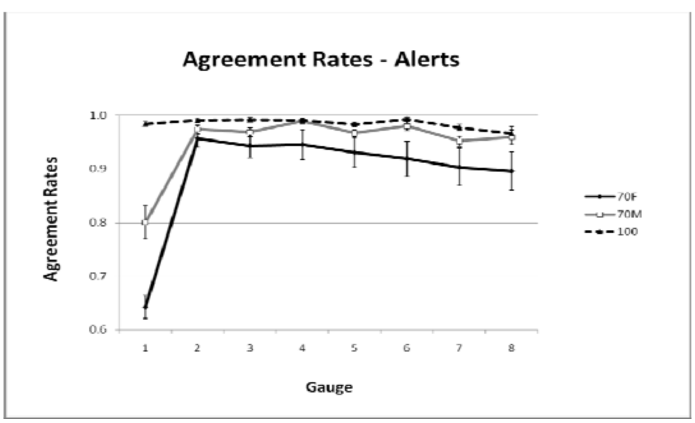

Figure 3. Compliance. SE bars are included.

Agreement Rates: Non-Alerts. Reliance was measured by how often the participant agreed with the automated aid for non-alerts (Figure 4). A 2-way ANOVA with Condition and Aid as factors showed 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 interaction between Condition and Aid, $F(14, 777) = 14.55$, $p < .001$, $\eta^2 = .21$, which means a drop in performance from 100% condition to the unreliable conditions was not equal across all Aids. Further analysis was done on the reliable Aids in order to test hypotheses 2 and 3. A 2-way ANOVA using Condition and Aids 2-8 as factors showed a marginally significant main effect of Condition, $F(2, 111) = 2.65$, $p = .075$, $\eta^2 = .05$,which indicates a drop in performance for these Aids in the 70F condition, $F(1, 74) = 3.82$, $p = .054$, $\eta^2 = .05$, and 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, $F(6, 666) = 6.71$, $p < .001$, $\eta^2 = .06$.

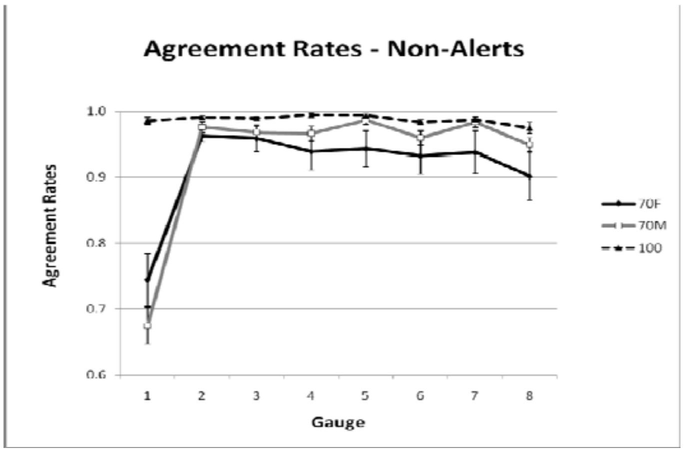

Figure 4. Reliance. SE bars are included.

Discussion

The purpose of the current study was to replicate the findings of Rice (2009) in order to test the differential effects of FAs and misses on compliance and reliance (Meyer, 2001, 2004), to find whether a contagion effect would still take place (Keller & Rice, 2010; Rice & Geels, 2010), despite the increase in reliable automated aids, and to discover the effects of FAs and misses in multiple automation. We tested three hypotheses based on findings in the previous literature. Our first prediction was that FAs would have a stronger negative effect on compliance, while also affecting reliance, and misses would have the reverse effect on the unreliable aid (Rice, 2009). Second, we predicted that SWT would still occur, despite the increase in reliable aids from previous SWT studies (Keller & Rice, 2010; Rice & Geels, 2010). Lastly, we predicted that FAs and misses in the first aid would differentially affect compliance and reliance throughout all the automated aids, with FAs having a greater negative effect on compliance and a weaker effect on reliance, while misses would have the reverse effect.

Hypotheses 1 and 2 were supported by the data, while Hypothesis 3 was only partially supported. With regards to Hypothesis 1, both error types had effects on compliance and reliance in the first aid, as seen in Rice (2009). The previous study had been the first to discover that both errors could reduce both types of trust and we have replicated the findings here. Pertaining to Hypothesis 2, SWT still took place, despite the increase in automated aids which has been shown to take place with 2, 4, and now 8 automated aids, though not as strong as we had predicted (Keller & Rice, 2010; Rice & Geels, 2010). The data indicate that both FAs and misses cause considerable contagion on the perfectly reliable aids, as predicted by SWT theory. Unfortunately, in regards to Hypothesis 3, FAs appear to cause more contagion than misses in both compliance and reliance measures, while misses did not have such a strong effect. This finding does not reflect the effect on trust in the first automated aid.

Conclusion

SWT theory (Keller & Rice, 2010) and Meyer's (2001, 2004) compliance-reliance model were combined in order to assess trust in multiple automated aids. By combining both theories, we were able to discover that FAs and misses replicate the mirrored effect on compliance and reliance (as seen in Rice, 2009) and that FAs may have caused a stronger contagion effect across all the perfect aids as compared to misses. Designers should be concerned with the possibility that operators may group multiple automated aids as an individual system rather than as independent components. Furthermore, designers should also be aware that the contagion effects may be stronger on reliable aids adjacent to an automated aid producing FAs than one producing misses.

Acknowledgments

The authors wish to thank Joshua Sandry and Gayle Hunt for their helpful comments, Sandra Deming and Laurel Ashley for their help in collecting data, and Chandrasekhar Thotakura for programming the simulator.

References

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19,* 775-779.

Dixon, S. R., & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors*. *48*(3), 474-486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*. *49*(4), 564-572.

Keller, D. & Rice, S. (2010). System-wide versus component-specific trust using multiple aids. *Journal of General Psychology*, *137*(1), 114–128.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40,* 153-184.

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

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors, 45*(2), 281-295.

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

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

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency". *International Journal of Aviation Psychology, 3*(1), 1-23.

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

Parasuraman, R., Sheridan, T., & Wickens, D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(3), 286-297.

Rice, S. (2009). Examining single and multiple-process theories of trust in automation. *Journal of General Psychology, 136*(3), 303-319.

Rice, S. & Geels, K. (2010). Using system-wide trust theory to make predictions about dependence on four diagnostic aids. *Journal of General Psychology, 137*(4), 362-375.