2011

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

Kasha Geels

Stephen Rice

Jeremy Schwark

Hayle Hunt

Joshua Sandry

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

Kasha Geels
Stephen Rice
Jeremy Schwark
Gayle Hunt
Joshua Sandry
New Mexico State University
Las Cruces, New Mexico

System-wide trust (SWT) strategy can occur when operators encounter multiple aids of differing reliabilities. Keller and Rice (2010) have shown effects of one unreliable aid influencing a perfectly reliable aid; participants had a tendency to treat both aids as one entire unit (SWT) rather than as two separate aids (component-specific trust). One limitation was that the use of only two diagnostic aids may not have been enough to generalize their results. This study seeks to further explore SWT with additional aids. Participants performed a 4-gauge monitoring task augmented by a diagnostic aid that provided recommendations of failures. The aids were either 70% or 100% reliable. Data revealed that although providing information and feedback about the aids benefited overall performance, agreement rate data showed that participants still employed a largely SWT strategy. The results from these data are applicable to the design and use of systems which contain multiple aids.

It is often useful to employ the help of an automated aid in environments with complex systems in order to increase safety and efficiency (Sheridan, 1987). This is the reason for the implementation of automated aids in places such as aircraft cockpits, surface transportation systems, and unmanned aerial systems. Parasuraman, Sheridan, and Wickens (2000) have proposed four stages of automation: synthesis, diagnosis, response selection and response execution. For the purposes of this study, we will focus on the diagnosis stage because it provides a recommendation while still leaving the decision-making to the human operator.

Diagnostic automation is used in complex systems so concurrent tasks may be performed while maintaining a specific workload. It is difficult for the human operator to perform concurrent tasks efficiently (Dixon & Wickens, 2006), including monitoring system gauges. Operators also lack the ability to perform cognitively demanding tasks while maintaining the same efficiency as an automated aid due to insufficient cognitive resources (Maltz & Shinar, 2003). Thus, the performance of the system is restricted unless automation is implemented. Even if one task is augmented by automation, multiple tasks can be performed at much higher levels of difficulty while increasing performance levels. The goal of diagnostic automation is to alert the operator of important information only when it is necessary (e.g., Wogalter & Laughery, 2006), such as warnings in an aircraft cockpit, which only alert the pilot of potential problems, so their attention can be focused on other tasks.

Diagnostic aids are not completely reliable. When a diagnostic aid errs, it produces either a false alarm or a miss (Green & Swets, 1966). A false alarm is produced when the aid detects an event that has not actually occurred, while a miss is produced when the aid does not detect an event that actually occurred. Both false alarms and misses negatively affect operator trust in the automated aid (Parasuraman & Riley, 1997; Rice, 2009).

Although previous research has shown how a single automated aid impacts trust (e.g. Dixon & Wickens, 2006; Dixon, Wickens, & Chang, 2005; Dixon, Wickens, & McCarley, 2007; Lee & Moray, 1994; Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2000; Rice, 2009; Rice, Clayton & McCarley, in press; Rice & McCarley, 2008; Rice, Hughes, McCarley & Keller, 2008; Rice, Trafimow, Clayton & Hunt, 2008; Wiegmann, Rich, & Zhang, 2001), not much is known of how trust is impacted by multiple aids. Keller and Rice (2010) have shown that operators tended to treat two separate automated aids in a task as one unit (system-wide trust) rather than as two individual units (component-specific trust), despite the fact that each aid differed in reliability. Therefore, errors produced by one aid negatively affected trust in the entire system.

Keller and Rice (2010) had participants fly simulated unmanned aerial vehicle missions while monitoring two separate gauges, each with its own diagnostic aid. Reliability of the aids varied from 70% to 100%. Trust in the 100% reliable aid was affected by the less reliable aid; thus, participants used a system-wide trust (SWT) strategy.

Six limitations to the previous study will be taken into account in the current study. One limitation to Keller and Rice (2010) is that participants only dealt with two diagnostic aids. Since one gauge makes up 50% of the system, it could be that the unreliability of one aid has a larger effect on trust in the entire system because it makes up half the entire system. The current study further employed the hypotheses to the use of four gauges with one unreliable aid; set to 70%, and three 100% reliable diagnostic aids. It is possible that one unreliable aid that only makes up 25% of the system would not have as much of an effect as one that accounts for 50% of the system. The second limitation to Keller and Rice's (2010) study is that they did not inform participants of the reliability of each aid. Perhaps a SWT strategy is only employed when participants are unaware of each aid's reliability. It is possible that making participants aware of the reliability of each aid may decrease their use of a SWT strategy. A third limitation is that participants were not given feedback on their performance after each trial. It is not known whether feedback is beneficial or not, so providing them with feedback could either help them or potentially confuse them, possibly damaging their trust in the automated aid. This is usually seen when first failures occur (Molloy & Parasuraman, 1996; Wickens & Xu, 2002). The fourth limitation is that Keller and Rice (2010) only used systems that erred with false alarms. Both false alarms and misses can affect operator trust, so the current study used misses in place of false alarms to further generalize to both types of diagnostic errors. Fifth, because the task employed by the participants in the previous study included two concurrent tasks which were unrelated, it may have confused participants as to the reliability of the diagnostic aids. The current study uses only a single-task paradigm, without the worry of a second task. The last limitation is that dependence measures were not taken into account. Dependence on the aids often positively correlates with overall accuracy; although this is not always true (Rice, 2009). So participants' accuracy may be high even if they ignore the aid. Dependence is measured by agreement rates with the aid and dependence is especially high if participants agree with the aid when it is in error.

The current study had two conditions which differed in the level of information: a) no information about reliability levels of the aids and no feedback on performance; and b) information on reliability in addition to feedback on their performance. Participants were given four gauges, with one set at 70% reliability and the remaining gauges set at 100% reliability.

Hypotheses for this study were that a) operators would employ a SWT strategy by combining all four gauges into one system, resulting in similar dependence across all four; b) the distance of each aid from the 70% reliable gauge would not make a difference in the reduction in trust; creating a systematic pull-down effect; and c) participants would have higher overall accuracy rates when given the reliability of each aid.

Method

Fifty-four (32 female, 22 male) undergraduates from a large Southwestern university participated in the experiment for partial course credit. The mean age was 19.82 (*SD* = 3.93). All participants had for normal or corrected-to-normal vision. The experiment was run on a Dell computer with a 3.3 GHz processor and a 22" monitor, with the resolution set to 1024 x 768. Viewing distance was controlled for using a chin rest centered at approximately 21" from the screen. The experimental display consisted of four gauges lined up from left to right; each displayed a randomly assigned 4-digit value. A range indicator was located on the top of each gauge, giving an ideal range for each gauge; which was also randomly assigned. For example, the 4-digit value and range might be 1836 (332). This means that the gauge's safe range is 1836 +/- 332 (between 1504 and 2168), if it went over or under the range, the gauge failed. Underneath each gauge, the aid provided a recommendation of either "Safe" or "Failure" and the left-most aid was 70% reliable. Errors were only misses; determining the gauge was "Safe" when a failure had actually occurred. The other aids were always 100% reliable. The position of the 70% reliable aid stayed constant so participants had to respond to the unreliable gauge first and to determine whether the pull-down effect would still occur even with the differences in distance of each 100% reliable aid from the unreliable aid (whether the right-most gauge would be as affected by the unreliable aid as the second gauge would).

There were two conditions in which the amount of information given to participants regarding their, and the aids', performance was manipulated. In the first condition (NI-NF: No Info, No Feedback), participants were told that the reliability levels of the aids were unknown. There was no feedback after each trial. In the second condition (I-F: Information, Feedback), participants were told the reliability of each aid at the beginning of the experiment, and were also given feedback after each trial.

Participants first signed consent forms and were then seated comfortably in a chair facing the experimental display. They read instructions on screen and were informed that they may ask questions before beginning the experiment. Each trial began with a fixation display that lasted 500 msecs. Following this, the task display presented the Gauge information for 10 seconds, after which a Choice display required participants to determine if the value in each gauge fell within the safe range. They responded by pressing the appropriate key for "Safe" or "Failure". They could agree with the automation if they desired, but were told the final choice was decided by them. Participants were required to respond to the gauges from left to right, individually. In the Feedback condition, a feedback display was provided for 1000 msecs after each trial. In the No-Feedback condition, a blank screen was presented for 1000 msecs. Participants completed 100 trials. The experiment lasted approximately 20 minutes. Upon completion of the experiment, participants were debriefed and dismissed. A mixed design was employed, whereby the Information factor was between participants, and the Reliability factor was within participants. Participants were randomly placed in each of the between-participant conditions.

Results

The between-participants Information factor referred to whether participants were given: a) no information and no feedback, or b) information and feedback. The within-participants Reliability factor referred to the reliability level of each aid (70%, 100%, 100%, and 100%). These data are presented in Figure 1.

Accuracy

Accuracy was measured by the proportion of successes divided by the total number of trials. An overall 2-way ANOVA using Information and Reliability as factors found a significant main effect of Information, $F(1, 52) = 14.29$, $p < .001$, and a main effect of Reliability, $F(3, 156) = 53.54$, $p < .001$, with a significant interaction between the two factors, $F(3, 156) = 5.40$, $p < .05$. These data indicate that providing more information to participants improved their performance across all reliability levels. Furthermore, performance in the 100% reliable aids was generally superior to performance in the 70% reliable aid.

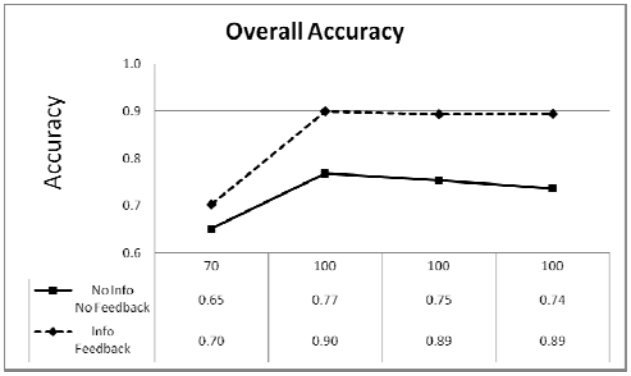

Figure 1. Overall Accuracy.

Dependence

Dependence was measured by the agreement rates between participants and the diagnostic aids (higher agreement rates equal stronger dependence) and is shown in Figure 2. When the aid correctly determined that the gauge had a "Failure", there was a significant main effect of Information, $F(1, 52) = 14.21$, $p < .001$, with no main effect of Reliability, $F(3, 156) = 1.68$, $p > .10$; however, this was qualified by a significant interaction between the two factors, $F(3, 156) = 3.75$, $p < .05$. Even though providing information and feedback benefited overall agreement rates, it did not appear to cause participants to treat the 100% reliable aids as more reliable than the 70% reliable aid.

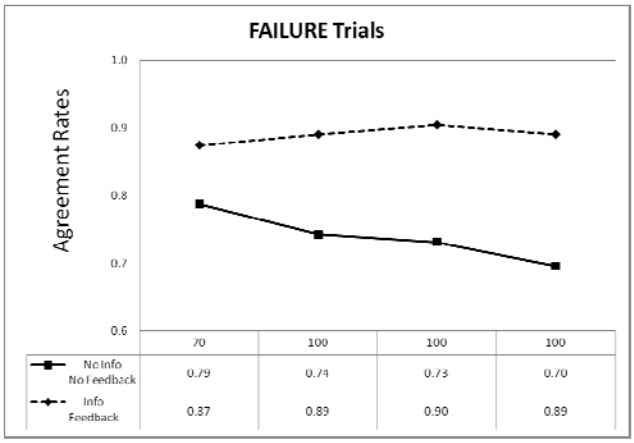

Figure 2. Failure trials.

When the aid correctly determined that the gauge was "Safe", there was a significant main effect of Information, $F(1, 52) = 6.33$, $p < .05$; however, there was no main effect of Reliability, $F(3, 156) = 2.04$, $p > .10$. This was qualified by a significant interaction between the two factors, $F(3, 156) = 6.04$, $p < .01$. As Figure 3 shows, only when information and feedback was presented were participants better able to differentiate between the reliabilities of the aids, and employ a more component-specific trust strategy (although not perfectly so).

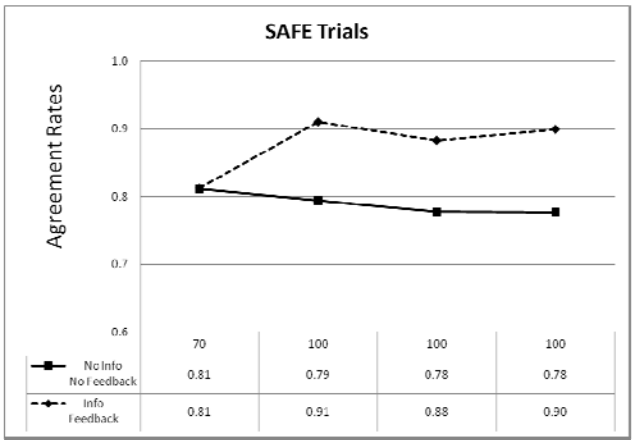

Figure 3. Safe trials.

During trials when the aid incorrectly determined that the gauge was safe (this only occurred for the unreliable aid), adding information and feedback reduced dependence on the aid for these trials, improving overall performance (Agreement Rates: NI-NF = 71%; I-F = 59%; marginally significant: $t(52) = 1.58$, $p = .06$); however, the agreement rates were still quite high given that the aid had erred, causing the drop in overall accuracy for that corresponding gauge.

Discussion

The purpose of this study was to further the findings of Keller and Rice (2010) as it applies to multiple aids while using only a single-task paradigm. Keller and Rice found that dependence on a perfectly reliable aid suffered when it was paired with an unreliable aid. Interestingly, they found that performance for both aids was almost equal. This showed that participants treated these two aids as having the same reliability and demonstrated the use of a SWT strategy.

Data in the current study further support the findings of Keller and Rice (2010). When one aid failed, the other three were treated similarly. Having additional information and/or feedback improved overall performance, but did not increase dependence on the reliable aids as compared to the unreliable aid. Low agreement rates fall in line with previous research which shows that miss-prone automation has a propensity to affect operator trust for both

alerts and non-alerts (Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007; Rice, 2009). When participants were given information and feedback, they then agreed with the 100% reliable aids more than they did with the unreliable aid. It is possible that they trusted these aids more because they realized the aids would never err (produce misses).

Although a purely system-wide trust strategy did not occur in the present experiment, component-specific trust strategy was rarely used. The trust strategy employed depended upon whether the aid determined a gauge was safe or not and if they were given information and feedback. A pull-down effect occurred in such a way that all three aids were almost equally affected by the unreliable aid. The increase in distance between the gauges did not prevent the pull-down effect from occurring all the way to the right-most aid.

As predicted, overall performance increased when participants were given information as to the reliability of each aid and when they were given feedback of their performance during each trial. This increase was seen equally on all the aids so it did not affect their trust strategy, except for the one situation in which the aid determined the gauge to be "Safe" and participants were given information and feedback.

Findings from the data provide important considerations when designing or using systems which contain multiple aids. Both designers and operators must be aware of the implications of interacting with flawed diagnostic aids and their effect on trust. If one diagnostic aid fails, an operator may lose trust in the other diagnostic aids when, in fact, the aids are performing optimally. This results in a loss of trust that is not necessary and could lead to falsely ignoring important automated aids (e.g. alarms, engine light, computer warnings, etc.). It could also lead to cognitive overload for the user because they take on the tasks normally performed by the diagnostic aid. Overall performance would also decrease because the operator's attention is taken away from the tasks they normally perform while they try to override the automated task.

Conclusion

Results from this study show the possibility of an operator treating multiple aids as one entire system, rather than as individual aids. When designing such systems, designers must be wary of the implementation of system-wide trust strategies which operators may employ. As shown in both the previous study (Keller & Rice, 2010) and the current one, all automated aids in a display can be affected by a loss of trust in only one aid. The designers of a system should try to avoid this decrease in overall performance by reducing the dependence on only the failed diagnostic aid while maintaining trust in the reliable diagnostic aids.

Acknowledgements

The authors wish to thank Eric Johnson, Audrey Rosenblatt, David Flemming, Sandra Deming, Eduardo Rubio, Crystal Sandy, Rachael Currier, Crystal Garcia, and Natasha Devries for their help in collecting data.

References

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., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors, 47*(3), 479-487.

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

Green, D.M., Swets J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

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.

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

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38,* 311-322.

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., Clayton, K., & McCarley, J. (in press). The effects of automation bias on operator compliance and reliance. *Human Factors Issues in Combat Identification*.

Rice, S., Hughes, J., McCarley, J. & Keller, D. (2008). Automation dependency and performance gains under time pressure. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.

Rice, S. & McCarley, J. (2008). The effects of automation bias and saliency on operator trust. *Proceedings of the XXIX International Congress of Psychology*.

Rice, S., Trafimow, D., Clayton, K., & Hunt, G. (2008). Impact of the contrast effect on trust ratings and behavior with automated systems. *Cognitive Technology Journal, 13*(2), 30-41.

Sheridan, T.B. (1987). *Handbook of Human Factors*. New York: Wiley.

Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention*. (AHFD-02-14/MAAD-02-2). Savoy, IL: University of Illinois, Aviation Research Lab.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*(4), 352-367.

Wogalter, M. S., & Laughery, K. R. (2006). Warnings and hazard communications. In G. Salvendy (Eds), Handbook of Human Factors and Ergonomics (pp. 889-911). New York: Wiley.