2009

Providing Evidence of a Multiple-Process Model of Trust in Automation

Stephen Rice

Gayle Hunt

PROVIDING EVIDENCE OF A MULTIPLE-PROCESS MODEL OF TRUST IN AUTOMATION

Stephen Rice
Gayle Hunt
New Mexico State University
Las Cruces, New Mexico

This study focuses on the effects of human responses to computer automation aids. Previous research has shown that different types of automation errors (false alarms and misses) affect human trust in different ways. False alarms tend to negatively affect operator compliance, whereas misses tend to negatively affect operator reliance. Participants were asked to determine whether an enemy target was present or absent in a series of images, a task similar to what a UAV operator might be asked to perform. A diagnostic aid provided recommendations before participants viewed each image. Reliability and type of automation error were manipulated in order to provide data to determine which of four theoretical models is most accurate. Analyses provided conclusive evidence that a multiple-process theory of operator trust is the only model which accurately explains behavior outcomes in this type of situation. A discussion of theoretical and practical implications of this finding is included.

The use of automation has accelerated so rapidly that it has outpaced the formation of a comprehensive theoretical understanding of human interaction with automated systems. Specifically, it is alarming to take notice of the lack of theory explaining how errors in automation affect human trust in this technology, which has an impact on the level of human dependence on automated systems. Although it is possible to find theoretical literature on this topic (e.g. Meyer, 2001; 2004; Parasuraman & Riley, 1997), it is nonetheless limited and must be studied more extensively in order to achieve a full understanding of the human cognitive process during human-automation interaction. Both trust (a cognitive state) and dependence (a behavior) are key factors in these studies, although they are not always correlated.

Many researchers have studied the cognitive and perceptual benefits of the use of automation on multi-tasking. In particular, it is presumed that when automation can be held responsible for completing a task, this should free up cognitive resources for the operator to focus on a different task at hand. The number of simultaneous tasks performed should equal the number of automated tasks plus the task being performed by the operator (e.g. Dixon, Wickens, & Chang, 2005).

Recently, there has been some attention on research focused on transferring cognitive resources away from an automated system and toward another task. An example of this research might be visual search and supervisory control. Many times, these two tasks are performed together (e.g. Dixon & Wickens, 2006). A real world example of this would be when a UAV operator must monitor all controls while simultaneously searching images for enemy targets (e.g. Maltz & Shinar, 2003).

While presently there are multiple people employed to operate a single UAV, there is an eventual goal to assign only one operator to each UAV, making it essential to understand the cognitive process of performing multiple tasks as well as working with, trusting, and depending on automation to lighten the load. It is also crucial to study automated systems that are not perfectly reliable and discover their implications on operator trust and performance of human-automation teams.

It is thought that automation takes place over four stages that are somewhat related to stages of human information processing (Parasuraman, Sheridan & Wickens, 2000). The first stage is information synthesis, which occurs by directing focus to particularly important environmental factors. The second stage is diagnosis, when automation provides an assumption about the information that has been taken in. Some examples may include warning alarms, which serve to focus the operator's attention on important events by utilizing auditory and/or visual warnings. The third stage is selection of response, and the fourth stage is execution of that selection. This paper will focus primarily on automation diagnosis, the second stage.

When stage 2 diagnosis is being utilized, it is possible that the operator may not have access to the information that is being processed. Instead the operator may only have access to the information provided by the automation. In the situation where an operator has raw data to confirm or disconfirm automation warnings, the performance outcome may be very different than when an operator has no raw data and must choose whether or not to follow the automation blindly (Sorkin & Woods, 1985). Further, human dependence on automation systems may be much more

fickle when there is no raw data to allow the operator to confirm the diagnosis, which may cause a grave failure. This study will focus only on diagnostic automation with raw data readily available to the operator.

Although reliable automation warnings can greatly increase positive performance (e.g. Dixon, Wickens & Chang, 2005; Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007), diagnostic automation is seldom perfect. Most often, the information being processed by the automation is imperfect, or the automation must make an assumption about an event that has yet to occur. Because diagnostic automation is often flawed, it is critical to understand how this factor affects the operator's trust and dependence on the system.

Signal detection theory is a useful tool in the understanding of automation errors. According to this theory, there are four possible outcomes of a diagnostic automation: hits (correct warning), misses (incorrect non-warning), false alarms (incorrect warning), or correct rejections (correct non-warning). Among these four possibilities, two of them (misses and false alarms) are automation errors. It may seem like a simple assumption that these errors are equally harmful and produce consequences of equal severity, but research has indicated differently. Data supports the idea that false alarms may be more harmful than misses (e.g. Bliss, 2003), and that the two error types actually cause very different consequences, especially in relation to operator trust (e.g. Dixon & Wickens, 2006; Maltz & Shinar, 2003; Meyer, 2001; 2004; Wickens & Dixon, 2007).

One method of discriminating between false alarms and misses has been offered by Meyer (2001; 2004). According to Meyer, false alarms have a negative effect on compliance, whereas misses have a negative effect on reliance. Compliance refers to how the operator responds to a warning, and reliance is how the operator responds to no warning. It can be inferred that this is an indication of each type of error affecting single but separate cognitive processes, as seen in Figure 1b. While this may seem reasonable, there has been much data indicating that false alarms affect *both* operator reliance and compliance (Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007; Wickens, Dixon, Goh, & Hammer, 2005).

Some of the strongest evidence to date suggests that false alarms have just as much of a negative effect on reliance as do misses (Dixon, Wickens, & McCarley, 2007). This data is conceptualized in Figure 1c.

Finally, further evidence supports yet another model. Rice and McCarley (2008) found that not only do false alarms have an effect on both compliance and reliance, but the same is the case for misses. It was found that misses have a negative effect on both areas as well. This evidence may be conceptualized in two ways, as demonstrated in Figures 1a and 1d.

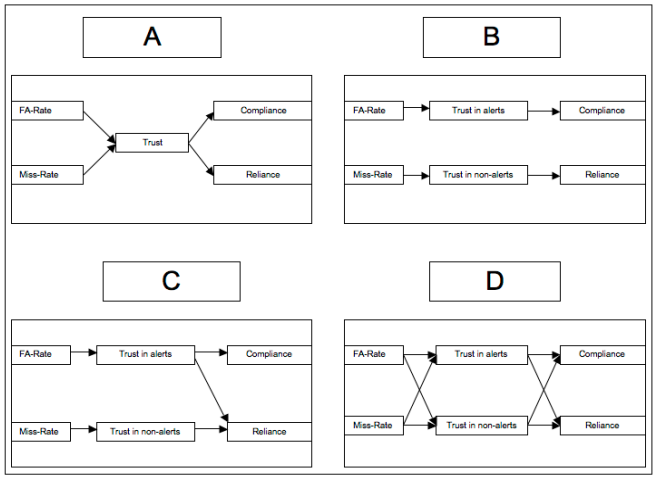

Figure 1. A) Single-process model; B) A selective two-process model; C) Mandler's two-process model; D) non-selective two-process model. Adapted from Dunn & Kirsner (1988).

With so many different models attempting to explain how errors in automation affect the behaviors of operators, there must be a way to distinguish between them and determine which best fits the data. One solution to this problem is to perform a state trace analysis on the data (Bamber, 1979; Dunn & Kirsner, 1988).

A state trace analysis plots two dependent variables against each other. The goal of this comparison is to observe if a monotonic relationship exists within the data. If a monotonic relationship is revealed, any change to one dependent variable will have a similar effect on the other. Should this be the case, a single process model, such as that in Figure 1b, would be supported. However, if the data demonstrates a non-monotonic relationship, (or a reversed association; Dunn & Kirsner, 1988), this would support a multiple process model, which indicates that at least two cognitive processes are in play (Bamber, 1979).

In order to simulate a task common to UAV operators, participants were asked to perform a target detection task by searching a set of aerial images of Baghdad for the presence or absence of an enemy tank. Assistance in this task was provided to participants in the form of a diagnostic aid which ranged in reliability from 95 to 55 percent in 5% increments. The automation was set to produce either false alarms or misses. This yielded 18 conditions total; 9 for false alarms and 9 for misses. Participants were fully informed of the type of errors the automation would make, as well as the reliability of the automation, and they were given clear feedback following each response.

There were four hypotheses related to this study: 1) better overall performance would occur in conditions with higher automation reliability as compared with lower automation reliability; 2) false alarm conditions would have strong selective effects on compliance and weaker non-selective effects on reliance, which would affect both how often the participant agrees with the automation and how quickly this decision is made; 3) miss conditions would have strong selective effects on reliance and weaker non-selective effects on compliance, again affecting both the rates of agreement and the response time; and 4) the use of a state trace analysis would uncover a non-monotonic relationship between compliance and reliance measures, supporting a multiple process theory to explain the outcome.

Method

Participants included 380 undergraduate students from New Mexico State University (230 females, 150 males), with a mean age of 20.3, who were compensated with partial course credit for their participation. All participants were screened for normal or corrected-to-normal eye sight as well as for color vision.

Images were presented on a 1024 x 768 resolution 20 inch flat-screen Dell monitor and computer with a refresh rate of 60 Hz. The monitor was roughly 40 degrees by 40 degrees in visual angle. One hundred images were presented to each participant, half of which consisted of target-absent stimuli and half of target-present stimuli. The target-absent stimuli were made up of 50 unchanged aerial images of Baghdad. The target-present stimuli were made up of the same 50 target-absent images, but with a tank digitally inserted into the image. The visual angle of the tank was roughly 2 degrees by 2 degrees, displayed with the turret facing one of 8 randomly assigned directions, N, NE, E, SE, S, SW, W, or NW. The presentation of images was randomized.

Participants were seated in a chair with their heads positioned by a chin rest 20 inches from the computer monitor. Each participant signed a consent form and proceeded to read directions presented on the screen. Directions included a sample picture of the target tank as well as full information regarding the type of error and reliability of the automation. Participants were asked to proceed quickly and with as much accuracy as possible. When participants felt confident with their instructions, they began the experiment.

As per the instructions, participants were aided by an aid which presented a recommendation before each image appeared on the screen. Twenty participants were randomly assigned to each of the 18 conditions.

Each trial was preceded by a small fixation, displayed for 1000 ms. Next was a screen providing the automation recommendation for the upcoming image, displayed for 1500 ms. The screen was then replaced with a randomly selected image from the previously mentioned set of 100 images. The image remained on the screen until the participant made a decision indicated by pressing the F or J key about the absence or presence of a tank (respectively). Following this, participants were presented with a feedback screen, displaying their accuracy and response time for that decision, as well as their cumulative accuracy for all the images they were previously presented with.

Results

General analyses (*d'*, C, and RT) are offered first, subsequently followed by additional analyses regarding compliance and reliance concerns. Finally, a state trace analysis was performed so as to test the theoretical models represented in Figure 1.

Sensitivity refers to a participant's ability to differentiate target-present from target-absent images, quantified by using the signal detection measure of sensitivity, *d'*. A two-way ANOVA on the imperfectly reliable conditions with Automation Error Type and Reliability as factors indicated that performance increased as the reliability of the automation increased, $F(8, 342) = 14.65$, $p < .001$, but showed no reliable effect of Automation Error Type, $F(1, 342) = 1.23$, $p > .05$, nor a reliable interaction of Automation Error Type and Reliability, $F < 1.0$, $p > .05$. These findings denote that although increased reliability rates did improve overall accuracy, automation error type did not affect accuracy.

Participants' response bias was calculated using the signal detection measure C. A two-way ANOVA on the imperfectly reliable conditions with Automation Error Type and Reliability indicated that participants in the False

Alarm conditions (M = -0.06), had a more liberal response bias than they did in the Miss conditions (M = 0.23), $F(1, 342) = 58.12$, $p < .001$; that is, participants were more apt to indicate that the target was present, regardless of their true performance sensitivity. There was no significant main effect of Reliability on participants' response bias, although it was marginally significant, $F(8, 342) = 1.87$, $p = .063$. An interaction between Automation Error Type and Reliability, $F(8, 342) = 2.77$, $p < .01$, indicated that the False Alarm condition was more likely to generate a liberal response bias among participants as the automation reliability increased.

Agreement rates and RTs were measured with the assumption that when participants trusted the automation, they would respond quickly in agreement. Only RTs from correct trials were integrated into the data analysis.

Compliance rate refers to the frequency in which participants agreed with the automation when it reported that a target was present. A two-way ANOVA performed on the imperfectly reliable conditions, with Automation Error Type and Reliability as factors, revealed a main effect of Automation Error Type, $F(1, 342) = 38.81$, $p < .001$, and a main effect of Reliability, $F(8, 342) = 4.87$, $p < .001$, with a significant interaction, $F(2, 66) = 2.15$, $p < .05$. These results indicate that participants in the False Alarm conditions were less likely than those in the Miss conditions to agree with the automation when it reported that a target was present, particularly when the automation was less reliable.

Reliance rate refers to the frequency in which participants agreed with the automation when it reported that the target was not present. A two-way ANOVA performed on the imperfectly reliable conditions, with Automation Error Type and Reliability as factors, revealed a main effect of Automation Error Type, $F(1, 342) = 25.64$, $p < .001$, but no main effect of Reliability, $F(8, 342) = 1.87$, $p > .05$, and no interaction, $F(2, 66) = 1.14$, $p > .05$. The significant main effect of Error Type indicates that participants in the Miss conditions were less likely than those in the False Alarm conditions to agree with the automation when it reported that a target was not present.

Compliance response time refers to the speed in which participants agreed with the automation when it reported that a target was present. A two-way ANOVA performed on the imperfectly reliable conditions, with Automation Error Type and Reliability as factors, revealed a main effect of Automation Error Type, $F(1, 342) = 48.15$, $p < .001$, no main effect of Reliability, $F(8, 342) = 1.37$, $p > .05$, and no significant interaction between Automation Error Type and Reliability, $F(2, 66) = 1.48$, $p > .05$. These results indicate that participants in the False Alarm conditions were slower to agree with the automation when it reported that a target was present, as compared to those in the Miss conditions.

Reliance response time refers to the speed in which participants agreed with the automation when it reported that a target was not present. A two-way ANOVA performed on the imperfectly reliable conditions, with Automation Error Type and Reliability as factors, revealed a main effect of Automation Error Type, $F(1, 342) = 33.70$, $p < .001$, with no main effect of Reliability, $F(8, 342) = 1.57$, $p > .05$, and no significant interaction between Automation Error Type and Reliability, $F(2, 66) < 1.0$, $p > .05$. These results indicate that participants in the Miss conditions were slower to agree with the automation when it determined that a target was not present, as compared to those in the False Alarm conditions.

The data above expose a pattern of false alarm rates affecting participant compliance more so than reliance, whereas miss rates affected participant reliance more so than compliance. However, in regards to the theoretical models discussed in the Introduction, this behavioral data cannot conclusively determine which model is correct. In order to test these issues, state trace analyses were performed on agreement rates and RTs, as seen in Figure 2a and 2b.

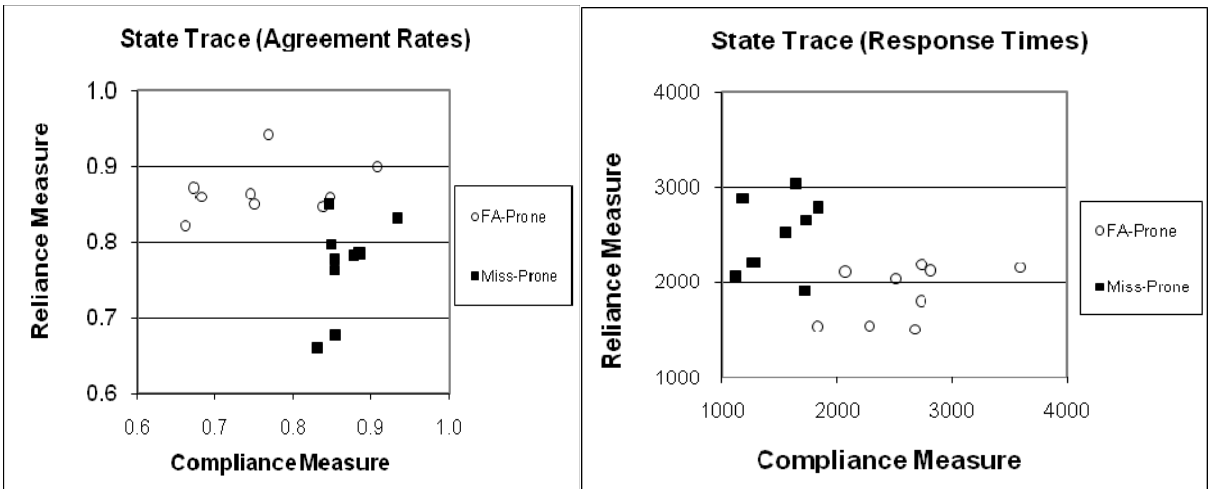

Figure 2. State Trace Analyses on a) Agreement Rates (%); and b) RTs (sec).

This analysis uncovers a non-monotonic relationship between the dependent variables. Spearman rank order correlations on the data revealed an r = -0.28 for Compliance agreement rates against Reliance agreement rates, and r = -0.35 for Compliance RT against Reliance RT. This information conclusively supports the notion that false alarm-prone and miss-prone automation affect at least two different cognitive processes, which translates to a different effect on operator behavior.

Discussion

All four theoretical models described in the introduction assume that there are only two types of automation errors (miss and false alarm) and two types of responses to these errors (compliance and reliance).

Recall the argument by Meyer (2001; 2004) presented in the introduction (Figure 1b). The data in this current study conclusively disconfirm the extreme version of this theory (Note: we agree that Meyer never actually advocated an extreme version). Analyses clearly show that false alarm rates affect both compliance and reliance. Furthermore, the theoretical model in Figure 1c has also been disconfirmed, as miss rates also affected both compliance and reliance.

With the models in Figures 1b and 1c no longer viable, we are left with the task of determining the correct model between Figures 1a and 1d. The only way to distinguish between these two models is with the use of a state trace analysis. As explained in the introduction, a monotonic relationship provides support for a single-process model, while a non-monotonic relationship proves a multiple-process model.

A very clear non-monotonic relationship emerged from the data. In short, an increase in value for one did not result in an equal increase for the other. Clearly, there are at least two separate cognitive processes involved in responding to automation false alarms and automation misses. This finding now disconfirms the theoretical model displayed in Figure 1a and clearly confirms the model in Figure 1d.

As predicted, higher automation reliability yielded better human-automation performance overall, when compared to less reliable automation. This effect is consistent with previous research (Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007). Thus, highly reliable automation should be used whenever possible, as it may be argued that reduced reliability rates could possibly do more harm than good (Wickens & Dixon, 2007).

Designers must take care when establishing the bias (more false alarms vs. more misses) of automation systems. They must not falsely assume that either type of error will produce the same effect. While the current data indicate that overall human-automation performance is not differentially affected by the automation bias, it is clear that performance during target-present or target-absent trials *is* differentially affected. The discrepancy between degradation of compliance (associated more strongly with false alarms) and degradation of reliance (associated more strongly with misses) must be carefully considered when programming the bias of an automated system.

Designers must consider which type of human error is more devastating—missing a target or falsely reporting a target. In a situation like airport security screening, it is much more dangerous to miss a target object than it is to falsely detect a target object. On the other hand, regarding an event such as a fire alarm, it is much more dangerous to have constant false alarms, as people may become subject to the "cry-wolf" effect (Breznitz, 1983). In situations like these, designers must adjust the automation bias according to the least harmful potential outcome.

Acknowledgments

The authors wish to thank Jackie Chavez for her help in collecting data. The authors also wish to thank Jason McCarley for extremely helpful comments and suggestions during the process of this study. This research was funded by an Air Force grant (Index #111915). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author.

References

Bamber, D. (1979). State trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*, 137–181.

Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology, 13*(3), 249-268.

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum.

Dixon, S. & Wickens, C. (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.

Dunn, J.C. & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*(1), 91-101.

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*, 563-572.

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

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors, 39*(2), 230-253.

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

Rice, S. & McCarley, J. (2008). The Effects of Automation Bias and Saliency on Operator Trust. *International Congress of Psychology*.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors, a signal detection analysis. *Human-Computer Interaction, 1*, 49-75.

Wickens, C.D. & Dixon, S. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomic Science, 8*, 201-212.