

Using State Trace Analyses to Measure Effects of
Automation Errors on Cognitive Processes

Stephen Rice¹, Krisstal Clayton¹ & Jason McCarley²

¹ New Mexico State University

² University of Illinois

Direct Correspondence to:

Stephen Rice, Ph.D.

stephenrice@inbox.com

Department of Psychology, MSC 3452

New Mexico State University

PO Box 30001

Las Cruces, NM 88003-8001

Keywords: Automation, Trust, Dependence, State, Trace

ABSTRACT

Objective: To examine the effects of automation errors on operator responses to alerts (compliance) and non-alerts (reliance). **Methods:** Participants performed a simulated combat identification task aided by imperfectly reliable diagnostic automation. The automated aid occasionally erred by producing false alarms or misses. **Results:** Both types of errors produced decrements in both operator compliance and reliance, a finding not seen in previous research. A State Trace Analysis revealed a non-monotonic relationship between the two types of errors. **Conclusions:** At least two cognitive processes are affected by automation false alarms and misses, disconfirming a single-process model of human dependence on automation.

INTRODUCTION

Automation has been described as the use of machines to augment or replace human activity (Wickens & Hollands, 2000). Over the past few decades, automation has become more pervasive in our society, to the point where it is virtually impossible to get through a day without the benefit of automated aids. Automation is used in both the home and the office, with the goal of reducing human workload and allowing humans to focus their cognitive resources on other tasks.

Unfortunately, the spread of automation through our everyday lives has outpaced the development of a theoretical understanding of human-automation interactions (Young, 1969). Although some progress has been made in developing these theories of automation (e.g. Bainbridge, 1983; Lee & Moray, 1994; Parasuraman & Riley, 1997), much remains to be done if we are to fully understand the ramifications of creeping automation. More particularly, it would be desirable to explicate precisely how the spread of automation can both benefit and harm overall human-automation performance. Although automation can improve human performance, much research indicates that automated aids can also lull human operators into complacency or lead them into error (e.g. Dixon & Wickens, 2006; Parasuraman & Riley, 1997; Wickens & Dixon, 2007). It is not safe to assume that automation will improve performance, or that humans should blindly adopt the assistance of each new aid that is offered them. To design automation that truly benefits human users will

require solid research that keeps pace with the emerging machines of the future.

One topic of recent concern has been the influence of automation on the control of perceptual and cognitive resources (Wickens & Holland, 2000), and the potential for automation to relieve these resources so that the operator can reallocate them to another task. Presumably, if automation can take full responsibility for a given task, then the human operator can ignore that task and focus his or her attention on another, thereby allowing the two tasks to be performed in parallel (Dixon, Wickens & Chang, 2005). One area where this could be of great benefit is in visual search, a task common to the combat arena and specifically to a UAV environment, where the payload operator must search photographic images for enemy targets (e.g. Yeh & Wickens, 2001; Maltz & Shinar, 2003). Visual search involves both perceptual and cognitive resources and can be highly demanding, particularly in combat identification tasks, where targets may be very small and/or camouflaged. The visual search paradigm is also useful in that it offers an opportunity to generalize research findings to other paradigms that involve human-automation interaction.

Automation has been described as having four stages, including information synthesis, diagnosis, selection, and execution (Parasuraman, Sheridan & Wickens, 2000). For current purposes, however, we limit our interest to the consideration of diagnostic automation, frequently

referred to as Stage 2 automation (Dixon, Wickens & McCarley, 2007).

Diagnostic automation provides the human operator an assessment as to the state of the world; it does not actually decide what to do, but instead offers a diagnosis to inform the operator's choice of action. For example, in combat identification, a diagnostic aid might alert the operator when it determines that an enemy target is present within a particular scene, leaving it to the operator to choose the appropriate response. Thus, after receiving the diagnosis, the operator can choose to accept it immediately (e.g. if under time pressure), attempt to confirm it by checking the raw data, or simply ignore it. Clearly, this decision is made more difficult when the automation is known to be prone to errors. Intuitively, the more errors the automation makes, the less likely the human operator will be to trust it (Dixon & Wickens, 2006; Wickens & Dixon, 2007), to the point where an abundance of errors might cause the human operator to ignore the aid altogether (Breznitz, 1983; Sorkin & Woods, 1985).

Note that within the framework of signal detection theory (MacMillan & Creelman, 2005) the errors committed by a diagnostic system can take either of two forms: false alarms (FA) and misses. An automation FA occurs when the aid mistakenly diagnoses the presence of a target when none is actually present, whereas a miss occurs when the aid concludes that a target is absent when it is actually present. Although it is tempting to assume that these two types of errors cause equivalent harm, this is not the case. Data suggest not only that automation FAs

cause more overall harm than misses (Bliss, 2003), but also that the two forms of error in fact have qualitatively different effects on operator trust and dependence (Dixon & Wickens, 2006; Maltz & Shinar, 2003; Meyer, 2001; 2004; Wickens & Dixon, 2007).

Meyer (2001, 2004) has proposed that automation FAs degrade compliance, which he defines as the operator's willingness to act on a target-present diagnosis from the aid. In contrast, automation misses affect what Meyer calls reliance, which he defines as the operator's tendency to trust the automation when it is silent, or otherwise indicates that there is no target. In its strongest form, this model essentially holds it may be that FAs affect one underlying cognitive dimension--Trust_{Alert}--that determines the operator's responses to all automation hits and FAs, whereas misses affect a different cognitive dimension--Trust_{Non-alert}--that determines the operator's responses to automation CRs and misses. This model is presented in Figure 1B.

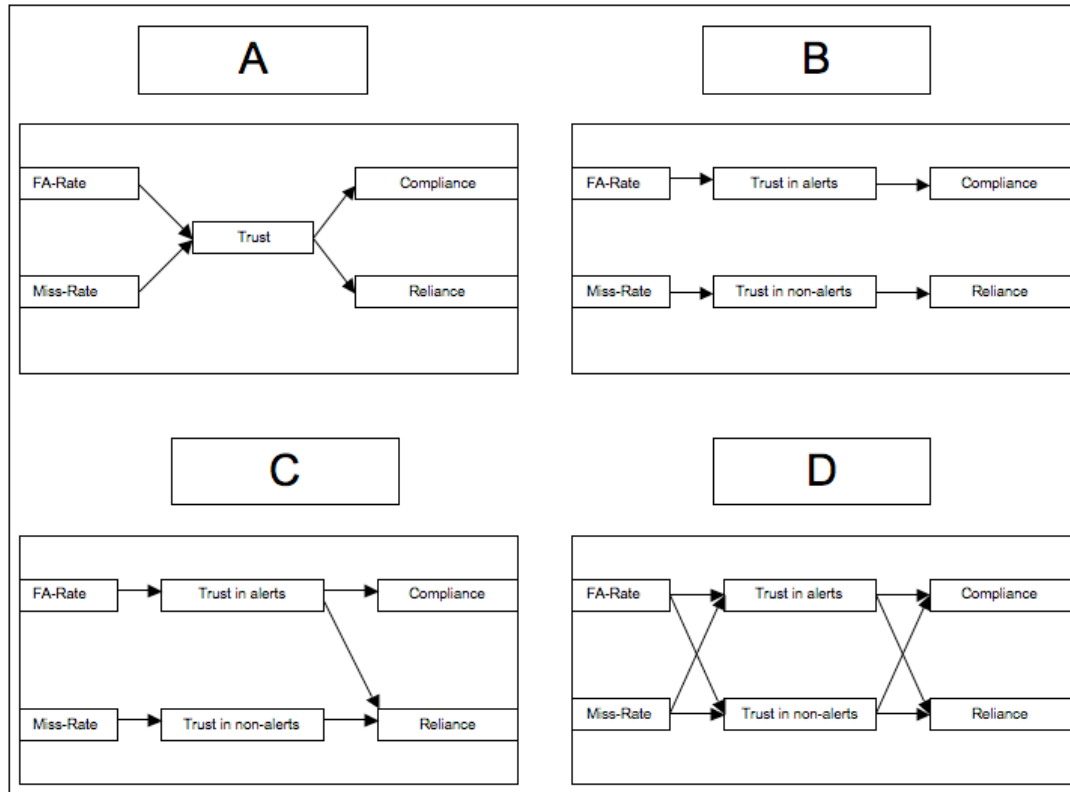


Figure 1. A) Single-process model; B) A selective two-process model; C) Mandler's two-process model; D) non-selective two-process model. Adapted from Dunn & Kirsner (1988).

In this form, the model would predict that compliance and reliance are independent of each other, and that the effects of automations misses and FAs are fully selective; FAs do not affect reliance, and misses do not affect compliance. Data appear to disconfirm this strong form of the compliance/reliance distinction, thereby indicating that automation FAs reduce operator reliance as well as compliance (Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007; Wickens, Dixon, Goh & Hammer, 2005). Figure 1C presents this model. Furthermore, some research has shown that automation misses can also compromise both reliance and

compliance (Rice & McCarley, 2008), which would be represented by Figure 1D.

It remains an open question, though, whether compliance and reliance reflect independent psychological processes. One might assume from the non-selectivity in the Dixon and colleagues studies that FAs and misses affect a single cognitive process, as represented by Figure 1A, which in turn regulates operator compliance and reliance. However, it is alternatively possible that $\text{Trust}_{\text{alert}}$ and $\text{Trust}_{\text{non-alert}}$ indeed exist as independent psychological dimensions, but that each of these dimensions is affected by both forms of automation error (see Figure 1D). If the influence of automation misses and FAs on $\text{Trust}_{\text{alert}}$ and $\text{Trust}_{\text{non-alert}}$ were simply weighted differently, the result would be a pattern like that described above, with FAs degrading compliance more than reliance, and with misses degrading reliance more than compliance. It is the intention of the current study to address these questions with the use of state trace analysis.

In a state trace analysis, one dependent variable is plotted as a function of another. In the current study, for instance, a measure of compliance is plotted as a function of measure of reliance. Of interest is the form of the relationship between the dependent variables if both are mediated by a single, common underlying mental value. Under the very weak assumption of a monotonic relationship between the underlying value and an observable variable, therefore, any manipulation that

increases the underlying value will also increase the value of both dependent variables. Consequently, the relationship between the dependent measures will itself be monotonic. A non-monotonic relationship, sometimes called a reversed association (Dunn & Kirsner, 1988), disconfirms a model based on a single underlying dimension, demonstrating instead that at least two underlying mental values are necessary to account for variation in the dependent measures (Bamber, 1979; Loftus, et al, 2004; see the appendix of Harley et al., 2004 for a brief introduction to state trace analysis).

Current Study. Participants in the current study were required to search aerial images of Baghdad for the presence of a designated target item (an enemy tank). As they performed the task they were assisted by a diagnostic aid that provided an assessment on each trial as to whether or not a target was present. The aid was imperfectly reliable, however, and the operator was free to agree or disagree with the aid as he or she felt was appropriate. The bias and the reliability of the aid were both manipulated, such that the aid either could be FA-prone or miss-prone, with reliability levels of 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55% or 50%. Operators were informed of the aid's reliability and response bias (miss-prone or FA-prone) before beginning the search task. Furthermore, they were given feedback about their accuracy and RT after each trial. We predicted that: a) more reliable automation would result in higher agreement rates and quicker RTs; b) less reliable automation would result

in lower agreement rates and longer RTs; c) state trace analysis would reveal a non-monotonic function, which would indicate that a multiple-dimensional cognitive model is needed to explain the data.

METHOD

Participants. Participants were 400 New Mexico State University undergrads (241 females) who received course credit. The mean age was 20.3. Participants were screened for normal or correct-to-normal visual acuity and color vision.

Apparatus and Stimuli. Stimuli were displayed on a Dell computer with a 20" monitor using 1024 x 768 resolution and a refresh rate of 60 Hz. A set of 100 aerial photographs of Baghdad was created using GoogleEarth. These 100 unaltered images served as target-absent stimuli. Target-present stimuli were created by digitally inserting an image of a tank into each target-absent photograph. Thus, there were 100 target-present and 100 target-absent images. The tank was approximately 2 degrees by 2 degrees in visual angle.

Figure 2 provides a sample image of a target-present trial.



Figure 2. A sample target-present image. The tank is located in the bottom-right quadrant.

Procedure and Design. Participants began by signing a consent form and reading on-screen instructions. The instructions included a picture of the tank and information about the bias and reliability of the automation. Instructions asked the participants to be as accurate as possible in conducting their search task without wasting time. Once participants were comfortable with the instructions, they pressed a key to begin the experiment.

Each trial began with a fixation display, whereby participants were instructed to look at the fixation cross for 1000 ms. This display was then replaced with a display

providing the automation's recommendation for that trial. This display read either, "The automation has detected a tank!" or "The automation has determined that there is no tank!" After 1500 ms, this display was replaced with a stimulus image, which remained until the participant made a response. Responses were made by either pressing the J key for target-present or the F key for target-absent. Following this, a feedback display appeared reporting the participant's accuracy and RT for that trial, along with a measure of cumulative accuracy.

As they performed the task, some participants were aided by a diagnostic aid which provided recommendations before each trial. The aid had a reliability level of 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55% or 50%, and was either FA-prone or miss-prone. Thus, there were a total of 20 conditions. Twenty participants were randomly assigned to each condition in a between-subjects design.

RESULTS

Sensitivity. The signal detection measure of sensitivity, d' , was used to measure participants' ability to discriminate target-present from target-absent images. Data are presented in Figure 3. A two-way ANOVA with Bias and Reliability as factors indicated that performance increased as the reliability of the automated aid increased, $F(8, 342) = 8.81, p < .001$, but showed no reliable effect of Bias, $F < 1.0$, nor a reliable interaction of Bias and Reliability, $F < 1.0, p > .05$.

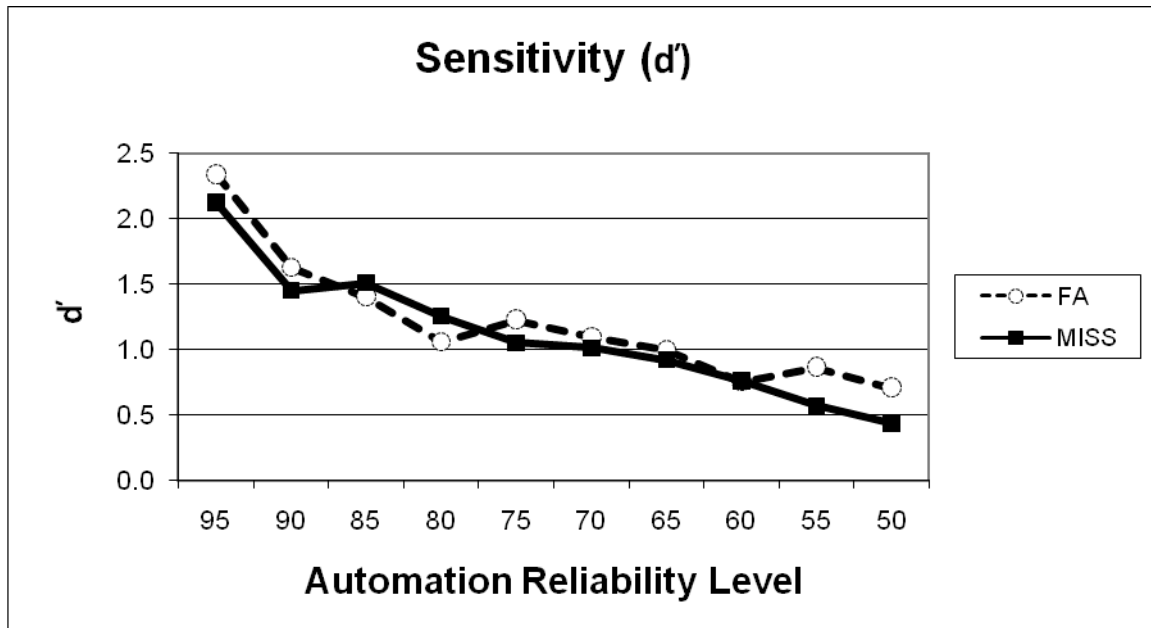


Figure 3. d' as a function of Bias and Reliability.

Bias. The signal detection measure C was used to analyze participants' response bias. A two-way ANOVA with Bias and Reliability indicated that participants in the FA-prone conditions ($M = -0.05$), had a more liberal bias than they did in the Miss-prone conditions ($M = 1.11$), $F(1, 342) = 20.00, p < .001$; that is, they were more likely to say that there was a target present, independent of their actual performance sensitivity. There was no significant main effect of Reliability on response bias, $F(8, 342) = 1.84, p > .05$, nor was there an interaction between Bias and Reliability, $F(8, 342) < 1.0, p > .05$.

Agreement Rates and Response Times (RTs). Agreement rates and RTs were used as a measure of trust in the automation. It assumed that when participants trust the automation, they will quickly agree with the aid. Thus, the following analyses investigated how often participants agreed with the aid and how quickly they did so.

Agreement Rates (Compliance). This measure refers to how often participants agreed with the aid when it indicated that a target was present. A two-way ANOVA with Bias and Reliability as factors performed on data from the imperfect automation conditions revealed a main effect of Bias, $F(1, 342) = 32.10, p < .001$, and a main effect of Reliability, $F(8, 342) = 4.32, p < .001$, with no significant interaction, $F(2, 66) = 1.35, p > .05$. These effects indicate that participants in the FA-prone conditions were less likely than those in the miss-prone conditions to agree with the automation when it judged that a target was present. Participants in both the FA-prone conditions and miss-prone conditions were less likely to comply with the automation when the reliability of the aid was low than when it was high.

Agreement Rates (Reliance). This measure refers to how often participants agreed with the aid when it indicated that a target was absent. A two-way ANOVA with Bias and Reliability as factors performed on data from the imperfect automation conditions revealed a main effect of Bias, $F(1, 342) = 33.89, p < .001$, but no main effect of Reliability, $F(8, 342) = 1.51, p > .05$, and no interaction, $F(2, 66) = 1.91, p > .05$. These results indicate that participants in the miss-prone conditions were less likely than those in the FA-prone conditions to agree with the automation when it judged that a target was absent.

Response Times (Compliance). This measure refers to how quickly participants agreed with the aid when it recommended that a target was present. A two-way ANOVA with Bias and Reliability as factors in the imperfect automation conditions revealed a main

effect of Bias, $F(1, 342) = 30.58, p < .001$, no main effect of Reliability, $F(8, 342) < 1.0, p > .05$, and no significant interaction between Bias and Reliability, $F(2, 66) = 1.21, p > .05$. These data indicate that participants in the FA-prone conditions were slower to agree with the automation when it determined that a target was present, relative to participants in the miss-prone conditions.

Response Times (Reliance). This measure refers to how quickly participants agreed with the aid when it recommended that a target was absent. A two-way ANOVA with Bias and Reliability as factors in the imperfect automation conditions revealed a main effect of Bias, $F(1, 342) = 30.85, p < .001$, with no main effect of Reliability, $F(8, 342) < 1.0, p > .05$, and no significant interaction between Bias and Reliability, $F(2, 66) < 1.0, p > .05$. These data indicate that participants in the miss-prone conditions were slower to agree with the automation when it determined that a target was present, relative to participants in the FA-prone conditions.

State Trace Analyses. The pattern of non-selective effects in the behavioral data (above) indicate that FA-prone automation disrupted operator compliance more than reliance, while miss-prone automation disrupted operator reliance more than compliance. This pattern of effects would be difficult to account for with a single-process model in which automation dependence is regulated by a unitary underlying construct such as a general level of trust. The data thus appear to support a multiple-process model in which different measures of trust regulate compliance and reliance behavior. State trace analyses were

conducted to confirm this interpretation. For these analyses, agreement rates and RTs were used, as seen in Figure 4.

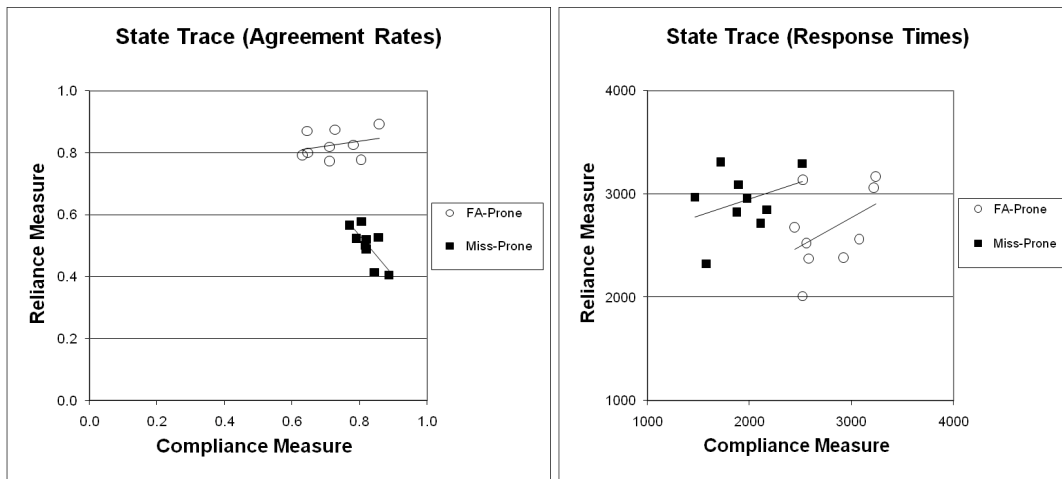


Figure 4. State Trace Analyses on a) Agreement Rates (%); and b) RTs (sec). Solid squares indicate miss-prone conditions, while clear circles indicate FA-prone conditions. Trend-lines are included in the graphs.

The figures clearly reveal a non-monotonic relationship between the dependent variables. This provides strong evidence against a single-process account of automation dependence. FA-prone and miss-prone automation appear to affect at least two different cognitive processes, which then in turn differentially affect operator behavior.

DISCUSSION

Benefits of Automation. As expected, highly reliable automation increased human performance in comparison to automation that was less reliable. Although past research has found that FA-prone automation

causes a decrease in human-automation performance in comparison to miss-prone automation (e.g. Maltz & Shinar, 2003; Dixon, Wickens & McCarley, 2007), no such asymmetry was demonstrated in the current data. Instead, d' and RTs were similar across the two bias manipulations.

Compliance and Reliance. Using operator agreement rates and RTs as an index of the effects of error-prone automation on operator compliance and reliance demonstrated that automation reliability did indeed have strong effects on operator behavior. Data revealed that the effects of FA- and miss-prone automation on operator compliance were asymmetrical but not fully selective; FA-prone automation severely compromised operator compliance but also produced a weak decrease in reliance, while miss-prone automation strongly degraded reliance but simultaneously reduced compliance as well. Therefore, these results argue against the strongest form of a two-process model of automation dependence (Maltz & Shinar, 2003; Meyer, 2001; 2004).

Regardless of these results, it is important to mention that non-selective effects of automation FAs and misses do not necessarily disconfirm the compliance/reliance distinction. They can be easily reconciled, rather, with a two-process model like that of either Figure 1D, in which two distinct forms of trust influence dependence and reliance and behavior. State-trace analyses conducted over measures of reliance and compliance produced results consistent with this possibility, demonstrating a non-monotonic relationship between measures. This

non-monotonic relationship precludes a model, like that of Figure 1A, which holds that both types of automation errors affect a single cognitive process that in turn affects operator compliance and reliance behaviors. As such, the data in total indicate that the multi-trust model which fits the current data most closely is the generalized multiple-process model illustrated in Figure 1D in the Introduction. In the current task, as described in the Introduction, the model would hold that automation errors affect two separate cognitive processes, which we call $\text{Trust}_{\text{Alert}}$ and $\text{Trust}_{\text{Non-alert}}$, which differentially influence compliance, the operator's response to an alert from the aid, and reliance, the operator's response to an "all clear" judgment from the aid. Results thus provide strong support for Meyer's (2001; 2004) compliance/reliance distinction.

Conclusions. In summary, data lead to a number of conclusions. First, highly reliable automation produces better human-automation performance than does less reliable automation (Wickens & Dixon, 2007). Second, the effects of automation misses and FAs and misses on compliance and reliance are not fully selective. Though FA-prone automation strongly reduces operator compliance, it also has weak effects on operator reliance. Automation misses, conversely, strongly compromise reliance, but also reduce compliance. Third, despite the non-selective effects of automation misses and FAs on operator dependence, the two types of automation errors affect at least two different cognitive processes, as revealed by the state trace analysis.

The practical implications of these findings are that designers must be aware of the differential effects of FA-prone and miss-prone automation on human dependence and behavior. Although intuition may suggest that FA-prone and miss-prone automation have equal effects on operator dependence and performance, there is in fact an important distinction to be made between the two types of automation errors and how they affect operator dependence and behavior.

Moreover, system designers may well face a tradeoff in designing an automated system for optimal human performance. Consider, for example, a case in which the automation designer wishes to minimize the possibility that a human operator will miss a target. The designer's understandable inclination may be to establish a liberal response bias for the automated aid, ensuring that the automation maintains a high hit rate. This hit rate, however, will come at the cost of a high FA rate. An unfortunate consequence of the aid's frequent FAs, in turn, will be a decrease in the operator's willingness to comply with the aid's alerts. The designer will thus have created a situation in which the aid is likely to detect a target but the operator, ironically, is unlikely to act on the aid's alerts. In establishing the response criterion for an automated diagnostic aid, therefore, the designer's goal should not be simply to optimize behavior of the aid itself, but to elicit an optimal pattern of dependence from the aid's user.

Acknowledgments

The authors wish to thank Amy Wells, Gayle Hunt, and Jackie Chavez for their help in collecting data. 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 authors.

REFERENCES

- Bainbridge, L. (1982). *Ironies of automation*. In G. Johanssen et al. (Eds.), *Analysis, design and evaluation of man-machine systems* (pp. 151–157). Pergamon.
- Bamber, D. (1979). State trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.
- 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 than 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.

- Harley, E. M., Dillon, A. M., & Loftus, G. R. (2004). Why is it difficult to see in the fog? How stimulus contrast affects visual perception and visual memory. *Psychonomic Bulletin & Review*, *11*(2), 197–231.
- Lee, J.D. & Moray, N. (1994). Trust, self-confidence, and operator's 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.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- 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.
- Wickens, C. D. & Hollands, J. G. (2000). Engineering Psychology and Human Performance, 3rd Edition. Upper Saddle River, NJ: Prentice Hall.
- Wickens, C.D., Dixon, S.R., Goh, J. & Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights: an attentional visual scanning analysis. *In Proceedings of the 13th Annual International Symposium of Aviation Psychology*. Dayton, Ohio.
- Yeh, M., & Wickens, C.D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors, 43*(3), 355-365.
- Young, L. R. (1969). On adaptive manual control. *Ergonomics, 12*(4), 635-675.