

HOW DO AUTOMATION FALSE ALARMS AND MISSES AFFECT OPERATOR COMPLIANCE AND RELIANCE?

Stephen R. Dixon, Christopher D. Wickens, and Jason S. McCarley
University of Illinois, Department of Psychology
Champaign, Illinois

Participants performed a tracking task along with a system and flight parameter monitoring task while aided by diagnostic automation. The goal of the study was to examine operator compliance and reliance as affected by automation failures, and to clarify claims regarding independence of these two constructs. Background data revealed a trend towards non-independence of the compliance-reliance constructs. Thirty-two undergraduate students performed a simulation that presented the visual display and collected dependent measures. False-alarm prone automation hurt overall performance more than did miss-prone automation, while also clearly affecting both operator compliance and reliance. Miss-prone automation only appeared to affect operator reliance.

INTRODUCTION

Diagnostic automated aids are designed to assist or replace human operators in a variety of tasks. They are useful for both civilian and military aviation environments. For example, the Traffic alert and Collision Avoidance System (TCAS) is a diagnostic aid that alerts pilots of a potential collision threat. Diagnostic alerts can also be used to warn pilots of low fuel conditions, sudden or unexpected altitude changes, attitude changes, etc. When operating reliably, these warning systems can relieve pilots of visual and cognitive demands, freeing resources that can be allocated to other tasks.

While reliable automation can have many benefits to overall performance in aviation (e.g. Dixon, Wickens, & Chang, 2005), automation is often imperfect because it must diagnose events based on imperfect probabilistic information in a changing world (Wickens & Dixon, 2006). For example, a TCAS alert indicates that a potential collision is possible if both aircraft follow the current flight path. However, sometimes awkward geometry is problematic for the TCAS, as seen in the “Dallas Bump”. In this situation, ascending aircraft are seen as a threat to the TCAS, when in reality, the ascending aircraft will level off before they become a real hazard. Thus, in this situation, the TCAS may be perceived as being erroneous.

When automation fails, it produces one of two types of errors. A false alarm (FA) is an incorrect indication of an event, while a miss is a failure of the automation to notice an event. In the case that the sensitivity of the automation cannot be improved, designers are responsible for appropriately setting the threshold of the aid, such that there are either more FAs or more misses. In aviation, there are many contexts in which alert thresholds might be varied by a designer, including the Automatic Target Recognition system in UAVs, collision alerts, and system alerts. It is critical for aviation designers to understand the consequences of automation errors on operator behavior when adjusting the threshold setting for each of these aids. In aviation settings, it appears that automation FAs may be more damaging than misses (Bliss, 2003); however, these effects may be modulated by cost/benefits, base rates, and training.

Meyer (2001, 2004) posited that automation FAs and misses have qualitatively different effects on operator dependence. He argues for two types of operator dependence: compliance and reliance. Compliance is how the operator responds to an automation alert, while reliance is how the operator responds to an automation non-alert. An increase in automation FAs will reduce operator compliance, while an increase in automation misses will reduce operator reliance. Thus, compliance and reliance are different states of operator dependence that are modulated by the threshold setting of the automated

device, and have implications on the behavior of the typical pilot.

One implication of the Meyer model is that compliance and reliance are independent constructs; that is, FAs selectively affect compliance, while misses selectively affect reliance (Meyer, 2004). However, data from recent aviation experiments using imperfect diagnostic automation reveals some inconsistency with this presumption. For example, Dixon and Wickens (2006) had pilots fly unmanned aerial vehicle (UAV) missions, while searching for targets and monitoring for system failures. The latter task was augmented by an imperfect diagnostic aid, which had various levels of reliability. Their results indicated that automation misses selectively affected operator reliance, while FAs non-selectively affected both compliance and reliance. Wickens, Dixon, Goh and Hammer (2005) replicated this study using an eye tracker, and found similar evidence in both behavioral and eye data. Recently, Wickens, Dixon and Johnson (2005) also found evidence for non-selective effects of FAs. The weakness in these studies, however, was that low experimental power prevented the investigators from making strong claims.

The current study moves away from the high fidelity simulations in order to provide a continuous and sensitive measure of concurrent task performance, thus allowing greater statistical power and experimental control. Participants were required to perform a difficult tracking task (simulating flight control in heavy turbulence) while simultaneously performing a gauge monitoring task. A context-free task was intentionally chosen in order to mimic the high attention and cognitive demands of a range of such cognitive monitoring tasks that might confront the pilot. Participants were aided by either a perfectly reliable diagnostic aid, or an imperfect aid that either failed by producing FAs or misses.

METHODS

Thirty-two students from the University of Illinois were paid \$9 per hour, plus bonuses for the top three performers in each condition. Performance bonuses were based equally on tracking task and systems monitoring performance. The simulation ran on a Dell computer with a 21" monitor, using 1280x1024 resolution. Figure 1 presents the experimental display.

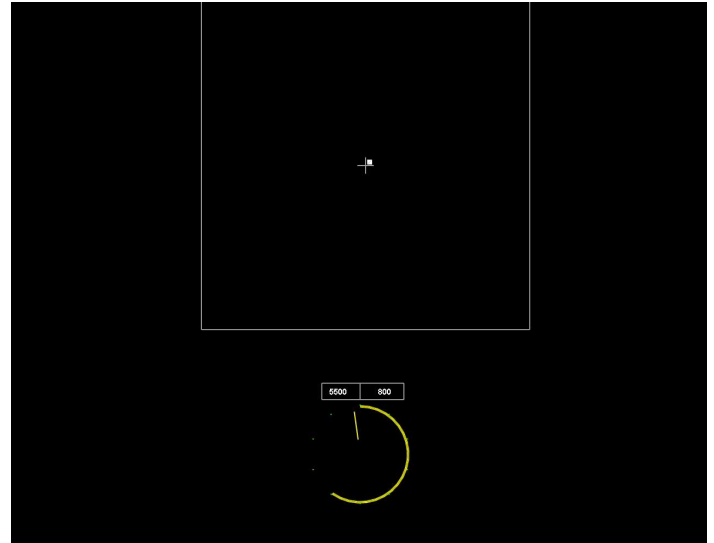


Figure 1. Sample screenshot of experimental display.

In the top portion of the display was the tracking task. Using a joystick, participants were required to keep the ball as close to the middle crosshairs as possible. In the bottom portion of the display was the system monitoring task. The gauge represented a generic real-world variable. If the gauge went outside an acceptable range, then participants were required to respond as quickly as possible to this “system failure (SF)”. Participants were told that if a system failure went undetected, then it was tantamount to allowing the aircraft to crash.

The system monitoring task was augmented by a diagnostic aid that sounded an auditory alert when a SF occurred. Expressed in signal detection theory, the aid could provide a hit, miss, false alarm, or correct rejection. The aid was either 100% reliable (A100), 60% reliable with false alarms (FA60), or 60% reliable with misses (M60). A final condition with no automation provided baseline performance. Participants were told that the automation would either be perfectly reliable or “not perfectly reliable”, and in the latter case, whether the automation would produce false alarms or misses.

There were 80 trials in total, all of which lasted exactly 30 seconds regardless of in-trial events. When a trial began, the target value (in the left numeric box above the SF gauge) changed to a new random value between 1000 and 9000, rounded to the nearest 100. The target range (in right numeric box above the SF gauge) changed to a new random value between 100 and 900, rounded to the nearest 100. A system failure occurred on half the trials, randomly ordered. Failures occurred within 5-12 seconds into a trial, and there

was never more than one SF per trial. Participants were allowed only one response per trial. At the end of each trial, a green light flashed for correct answers and a red light flashed for incorrect answers.

RESULTS

One subject in the M60 condition was dropped due to unusually poor performance. Analysis entailed a one-way omnibus ANOVA, followed by three planned comparisons: a) Baseline vs. A100, b) Baseline vs. the combination of FA60 and M60 in a planned comparison (i.e. weights of -1, 0.5, 0.5), and c) FA60 vs. M60. Because only three a priori comparisons were made, familywise error rates were not adjusted (see Keppel, 1982, for more details). Any post-hoc tests used a Bonferroni correction.

Tracking Error. Tracking error was calculated only during the period of time between the beginning of a trial and the onset of either a system failure or an automation false alarm, since this was the period of time where variations in attentional reliance caused by the different conditions were expected. The white bars in Figure 2 present these data as a function of condition.

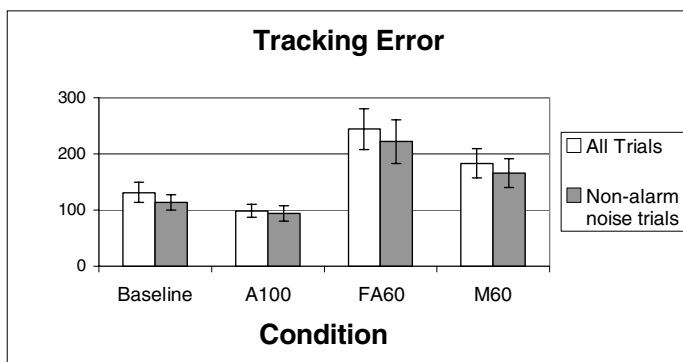


Figure 2. Tracking error as a function of condition. The white bars represent all trials in the experiment, while the grey bars represent only non-alarm noise trials. SE bars are included.

A one-way ANOVA revealed a reliable main effect of condition, $F(3,27) = 6.64$, $p < .01$. Planned comparisons between the Baseline condition ($M = 131$) and the A100 condition ($M = 98$) approached significance, $t(14) = 1.53$, $p = .07$. The Baseline condition showed better tracking performance relative to the average of the two unreliable conditions, $t(14) = 2.55$, $p = .01$, while the difference between the FA60 ($M = 243$) and M60 conditions ($M = 182$) was not statistically significant, $t(13) = 1.32$, $p > .10$.

A separate analysis done only on trials in which there was no SF and the automation was silent revealed a main effect of condition, $F(3,27) = 5.09$, $p < .01$. These data are shown as grey bars in Figure 2. A post hoc comparison between the FA60 and the A100 conditions, $t(14) = 3.78$, $p < .01$, revealed that tracking performance in the false alarm prone condition was worse than in the perfectly reliable condition. This indicates that, in the false alarm condition, participants did not rely on the automation as much as their counterparts did in the condition where the automation never failed. Theoretically, when the automation was silent, participants should never have looked at the system gauge because they knew that the automation never missed. Only automation alerts should have warranted an inspection of the system gauge. The fact that tracking error suffered so much during non-alert trials indicates that operator reliance was negatively affected; that is, participants were inspecting the system gauge even during trials when the automation was silent.

SF Detection Rate. For all detection rate analyses, the signal detection measure d' was used. A one-way ANOVA revealed a main effect of condition, $F(3, 27) = 8.84$, $p < .001$. Planned comparisons revealed no significant difference between the Baseline condition ($M = 3.03$) and A100 condition ($M = 3.20$), $t(14) < 1.0$. The Baseline condition produced higher detection rates than the average of the two unreliable conditions, $t(14) = 2.43$, $p = .01$. The FA60 condition ($M = 2.04$) produced lower detection rates than the M60 condition ($M = 2.61$), $t(13) = 3.08$, $p < .01$. Post hoc tests revealed that the Baseline condition produced higher detection rates relative to the FA60 condition, $t(14) = 3.15$, $p < .01$, but did not differ significantly from the M60 condition, $t(13) = 1.38$, $p > .10$.

When there was an automation alert, all groups tended to agree with automation, but less so with FA-prone automation ($M = .93$) than with miss-prone automation ($M = 1.00$), $t(14) = 3.75$, $p < .01$. In contrast, when the automation was silent, the operator was more likely to incorrectly false alarm in the miss-prone condition than in the FA-prone condition, $t(13) = 2.14$, $p < .05$.

SF Response Times. SF response times are presented in Figure 3. The white bars represent all

trials in the experiment while the grey bars represent *only true-alarm signal* trials (to be discussed below).

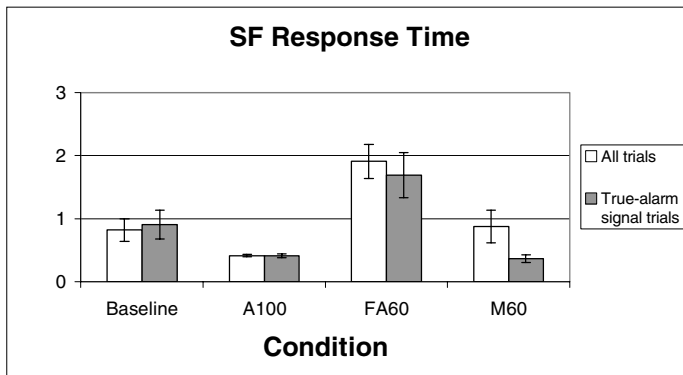


Figure 3. SF response times as a function of Condition. The white bars represent all trials in the experiment, while the grey bars represent only true-alarm signal trials. SE bars are included.

A one-way ANOVA on the data for all trials revealed a main effect of condition, $F(3, 27) = 9.85$, $p < .001$. Planned comparisons revealed that participants in the Baseline condition ($M = 0.82$ s) responded more slowly than those in the A100 condition ($M = 0.42$ s), $t(14) = 2.26$, $p < .05$, but more quickly than the average of the two unreliable automation conditions, $t(14) = 2.48$, $p = .01$. The FA60 condition ($M = 1.91$) produced longer response times than the M60 condition ($M = 0.88$), $t(13) = 2.71$, $p < .01$. Post hoc tests revealed that response times in the Baseline condition were faster than in the FA60 condition, $t(14) = 3.35$, $p < .001$, but was not significantly different from that in the M60 condition, $t(13) < 1.0$.

A one-way ANOVA on trials where the automation sounded a true alert revealed a main effect of condition, $F(3, 27) = 7.54$, $p < .01$. Planned comparisons between the M60 condition ($M = 0.37$) and A100 condition ($M = 0.41$), $t(13) < 1.0$, revealed that the misses in the M60 condition did not appear to affect operator compliance, and that response times in the M60 condition following a true alarm were faster than those in the Baseline condition ($M = 0.91$), $t(13) = 2.10$, $p < .05$. As expected, the FA60 condition ($M = 1.69$) did degrade operator compliance, $t(14) = 3.53$, $p < .01$.

DISCUSSION

The current study replicated the finding that perfect diagnostic automation is beneficial to overall

human-automation performance (Dixon, Wickens, & Chang, 2005). Miss-prone automation harmed the tracking task by causing operators to shift attention away from that task in order to catch the potential automation misses (Dixon and Wickens, 2006).

False-alarm prone automation damaged the systems monitoring task by reducing operator compliance, as both the SF detection rates and response times suffered relative to the perfectly reliable automation condition, and even dropped below Baseline performance. Importantly, automation false alarms also adversely affected operator reliance, confirming the non-selective affects that Wickens et al. (2005) and Dixon and Wickens (2006) suggested based on trends seen in their data. When the automation was silent, operators in the false-alarm condition should have completely ignored the systems monitoring gauge and focused their entire attention on the tracking task. Instead, the data revealed that the tracking task performance in the false-alarm condition was worse than the reliable automation condition. This implies that automation errors may not selectively affect operator compliance and reliance.

Thus, our prediction that the FA-prone condition would be more harmful to overall performance relative to the miss-prone condition, proved to be correct both qualitatively and quantitatively. First, the FA-prone automation adversely affected both operator compliance and reliance, while the miss-prone automation only appeared to reduce operator reliance. Second, FA-prone automation hurt performance more on the automated task than did miss-prone automation, and hurt performance at least as much as miss-prone automation on the concurrent task.

The purpose of the current study was to expand on the compliance-reliance constructs posited by Meyer (2001; 2004) by providing a context-free simulation that could provide a more sensitive analysis of the qualitative differences between automation false alarm and misses. By moving away from the high-fidelity aviation simulations conducted in previous studies, the current data allow stronger conclusions to be made regarding the impact of automation false alarms and misses on operator performance when deciding where to set the bias threshold in future aviation systems.

ACKNOWLEDGMENTS

The authors wish to acknowledge the research support of contract ARMY MAD 6021.000-01 from Micro-Analysis and Design and the Army Research Laboratory. Dave Dahn and Marc Gacy were the scientific/technical monitors. The opinions expressed in this chapter are those of the authors and do not necessarily reflect those of the US Army.

REFERENCES

- Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology*, 13(3), 249-268.
- Dixon, S.R. & Wickens, C.D. (2006). Automation Reliability in Unmanned Aerial Vehicle Flight Control: Evaluating a Model of Automation Dependence in High Workload. *Human Factors*.
- 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.
- Levinthal, B. & Wickens, C.D. (2005). *Supervising Two Versus Four UAVs With Imperfect Automation: A Simulation Experiment*. (AHFD-05-24/MAAD-05-7). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- 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.
- Wickens, C.D. & Dixon, S.R. (2006). Is There a Magic Number 7 (to the Minus 1)? The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature. *Theoretical Issues in Ergonomics Science*.
- 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*.
- Wickens, C.D., Dixon, S.R., & Johnson, N.R. (2005). *UAV Automation: Influence of Task Priorities and Automation Imperfection in a Difficult Surveillance Task*. (AHFD-05-20/MAAD-05-6). Savoy, IL: University of Illinois, Aviation Human Factors Division.