

AUTOMATION FAILURES ON TASKS EASILY PERFORMED BY OPERATORS UNDERMINES TRUST IN AUTOMATED AIDS

Poornima Madhavan, Douglas A. Wiegmann, and Frank C. Lacson
Aviation Human Factors Division, Institute of Aviation
University of Illinois at Urbana-Champaign

Automation users often distrust diagnostic aids that are imperfectly reliable. The extent to which users' trust in automation is influenced by the simplicity of automation errors was explored. Participants ($n = 30$) performed 200 trials of a target-detection task using an imperfectly reliable automated aid. For the "easy-miss" group ($n = 15$) the automated aid missed targets only on easy trials and was accurate on difficult trials. For the "difficult-miss" group ($n = 15$) the aid missed targets only on difficult trials and was accurate on easy trials. A control group ($n = 15$) performed the task unaided. The easy-miss group trusted the aid less, was less accurate, and contradicted the aid more than the difficult-miss group, even on trials when the easy-miss aid was more reliable than the difficult-miss aid. Results suggest that the "easiness" of the aid's errors for the easy-miss group undermined automation trust and reliance. Potential future directions include examining whether easy false alarms affect user trust in a manner akin to easy misses.

INTRODUCTION

Automated diagnostic aids are increasingly being incorporated in complex systems such as aviation, nuclear power, and healthcare. The extent to which automated aids will actually improve performance is difficult to predict given such aids are often less than 100% reliable and, as such, operators may not trust them. Indeed, trust in automation is one of the fundamental elements governing human-automation interaction (Sheridan & Farrell, 1974). Imperfectly reliable automation is often under-trusted and disused even under circumstances where aided performance is better than unaided performance (Parasuraman & Riley, 1997; Wickens & Hollands, 2000), suggesting a complex and multifaceted relationship between automation trust and reliance (Lee & Moray, 1992; Wiegmann, Rich & Zhang, 2001; Sheridan, 2002).

When operators trust an automated diagnostic aid that is more reliable than manual (i.e., unaided) performance, they are more likely to rely on the automated aid than on their own diagnoses. Similarly, when operators distrust a diagnostic aid that is less reliable than manual performance, they are more likely to ignore the aid and rely on themselves to diagnose a situation. Appropriate reliance occurs

in both cases (Dzindolet, Peterson, Pomranky, Pierce, & Beck, in press). However, over-reliance or misuse can occur when an operator over-trusts an aid that is less reliable than unaided performance. Likewise, when operators under-trust an aid that is more reliable than manual performance, under-reliance or disuse of automation can occur (Parasuraman & Riley, 1997).

According to signal detection theory (Green & Swets, 1988; Swets & Pickett, 1982), operators' reliance on automation depends on the quantitative weighting of the probability that the information presented by the automated aid is truly representative of the actual state of the system. The choice of a utilization strategy depends on the user's perception of the trade-off between the true-positive proportion (hit rate), and the false-positive proportion (false alarm rate) (e.g., Getty, Swets, Pickett, & Gonthier, 1995; Elvers & Elrif, 1997; Meyer, 2001; Maltz & Meyer, 2002). However, accurate weighting of the reliability of automation rarely occurs.

While operators sometimes under-estimate the reliability of imperfect automation (Wickens & Hollands, 2000; Wiegmann, 2002), the extent to which perceived reliabilities deviate from true reliabilities may depend upon the type of error made by an automated aid. For example, when the base rate

of a real world event is low (e.g., a weapon in a suitcase), the potential for false alarms by an automated aid is high, even for very sensitive systems. Automated aids that produce a high number of false alarms (i.e., repeatedly “cry wolf”; Breznitz, 1983) create under-trust in automation, resulting in operators ignoring automation alerts.

Another factor undermining trust is the *conspicuity* of automation failures. According to Dzindolet, Pierce, Beck, Dawe & Anderson (2001) the perception of the reliability of an automated aid is filtered through the operator’s “perfect automation schema,” or the expectation that automation will perform at near perfect rates. This expectation leads operators to pay too much attention to errors made by automation (Dzindolet, Pierce, Beck, and Dawe, 2002), thereby triggering a rapid decline in trust when diagnostic aids make errors (e.g., Lee & Moray, 1992, 1994; Dzindolet, et al., 2001; Wiegmann, et al, 2001).

Indeed, researchers in cognitive psychology have found that information inconsistent with expectations (e.g., schemas) is likely to be well remembered and plays an unduly large role in information processing (Smith & Graesser, 1981; Ruble & Stangor, 1986). Therefore, obvious errors made by an automated aid might dramatically reduce trust in an automated diagnostic aid. Consistent with the obvious-error hypothesis, a recent study by Dzindolet, et al (in press) revealed that automation users who observed the errors made by an imperfect diagnostic aid trusted the aid less and were more likely rely on their own decisions, compared to those who did not have the opportunity to view the automated aid’s errors.

A factor related to error conspicuity that may also adversely affect trust is the *simplicity* of errors made by automated aids. For example, Dzindolet et al (in press) noted that several participants justified their lack of trust in the aid by stating, “The computer didn’t earn my complete trust because I swear I saw someone when the computer said there was no one there.” And “There were a few times that I’m pretty sure I saw a soldier but the program said he was absent.” Such statements suggest that when operators find a target “easy” to detect, yet the automation fails to detect it, trust is severely undermined, even when the aid is on average more reliable than the human operator is. In addition, when operators catch easy mistakes made by an automated aid, it may serve to irrationally bolster their self-confidence in their own ability to outperform the aid. This “easy-error” hypothesis, however, has yet to be empirically tested.

Purpose of the Present Study

The purpose of the present study was to test the hypothesis that automation failures on tasks easily performed by operators will undermine trust in automation that reliably performs difficult tasks. Specifically, participants performed a signal detection task in which they were required to find target letters embedded in an array of letter distracters. On some trials, the number of distracters was small, making the target easy to detect, whereas on other trials numerous distracter letters were present, making the target difficult to detect. Two groups of participants utilized a diagnostic aid that had a .2 miss rate and a .4 false alarm rate. However, for half the participants, the aid only missed targets on easy trials. For the other half, the aid only missed targets on difficult trials. A third unaided group served as a control.

Based on the easy-error hypothesis, we predicted that compared to an aid that generated difficult errors alone, an aid that generated easy misses would (1) have a greater impact on subjective trust and perceived reliability of the aid, and (2) lead to lower reliance on the aid (i.e., greater disagreement) and higher confidence estimates on difficult trials. The appropriateness of this reduced reliance, however, can only be determined by comparing the performance of aided groups with the unaided control group.

METHOD

Participants

Forty-five students from the University of Illinois completed the experiment. Participants were paid \$ 8 for their participation and participation time did not exceed 1 hour.

Tasks and Procedures

Participants performed 200 trials of a computer simulation task that required them to detect the presence of a target “X” among an array of alphabets on a screen. The simulation was developed using Visual Basic for MS-DOS and presented on a desktop computer equipped with a 22-inch color monitor and standard keyboard. Out of the 200 trials, 50% (n = 100) were “target trials”

where the “X” was present among the noise alphabets, while the other 50% (n = 100) were “noise trials” wherein the “X” was absent. Out of the 100 target trials, 50% (n = 50) were “easy” (90 targets embedded in 1000 noise alphabets), and 50% (n = 50) were “difficult” (5 targets embedded in 1000 noise alphabets).

The trials flashed on the screen in random order, each for a duration of 750 milliseconds. After a delay of 5 seconds, aided participants (n = 30) were presented with the decision of a diagnostic aid to help determine the presence or absence of the “X”. The likelihood that the aid made an error (i.e., miss/false-alarm) was .20 and .40, respectively. On the 100 noise trials, the aid was programmed to commit 40 false alarms and 60 correct-rejections for all participants. Participants were not informed of the reliability of the aid before testing.

For the “difficult-miss” group (n = 15) the aid committed 20 misses and 30 hits on the 50 *difficult-target trials* (60% accurate) and was 100% accurate on the 50 easy-target trials. Conversely, for the “easy-miss” group (n = 15), the aid committed 20 misses and 30 hits on *easy-target trials* (60% accurate) but was 100% accurate on the 50 difficult-target trials. After receiving the aid’s diagnosis participants indicated whether or not they thought the target was present. A third “unaided control group” (n = 15) performed the task without the aid. All participants indicated their confidence in each diagnosis on a scale ranging from 1 (no confidence) to 5 (very confident). Feedback was then given as to whether the diagnosis was correct.

Following completion of 200 trials, aided participants estimated the diagnostic aid’s reliability using a percentage scale that ranged from 0 to 100, and their trust in the aid using a scale that ranged from 0 (did not trust at all) to 8 (trusted all the time) on a post-experimental questionnaire.

RESULTS

Performance Measures

Easy target trials. As expected, target detection accuracy of all three groups was perfect (M = 100%) on easy-target trials. In addition, the difficult-miss group never disagreed with the aid given the aid made no errors on easy-target trials.

The easy-miss group appropriately agreed with the aid on roughly 60% of the easy-target trials (M = 59%, SD = .008), given their aid was accurate on only 60% of these trials. Confidence did not differ significantly across groups: easy-miss (M = 3.71, SD = .34), difficult-miss (M = 3.69, SD = .53) or control (M = 3.67, SD = .4).

Difficult target trials. As illustrated in Figure 1 (a), accuracy of participants in the unaided control group was 50% (SD = .13), confirming that targets were difficult to detect. Accuracy of the difficult-miss group (M = 61%, SD = .12) was significantly higher than that of the control group $t(28) = 2.33$, $p < .05$, and roughly equivalent to the 60% accuracy of their automated aid. Consistent with the easy-error hypothesis, accuracy of the easy-miss group (M = 50, SD = .12) was identical to that of controls and significantly less than that of the difficult-miss group, $t(28) = 2.01$, $p = .05$, though the aid utilized by the easy-miss group made no errors on difficult target trials. In fact, subjects in the easy-miss group agreed with their automated aid on only 50% of difficult-target trials (SD = .16), which was significantly lower than the difficult-miss group (M = 64%, SD = .12) whose aid was only 60% reliable, $t(28) = 2.66$, $p < .001$. Contrary to expectations, confidence in the easy-miss (M = 1.89, SD = .55) and difficult-miss groups (M = 1.92, SD = .45) did not differ significantly from one another, $t(28) = .17$, $p = .87$, and both were significantly less than that of controls (M = 3.41, SD = .57), $t(28) = 7.45$, $p < .001$, $t(28) = 7.93$, $p < .001$.

Noise trials. As depicted in Figure 1 (b), accuracy of the unaided control group on noise trials was 81% (SD = .13). Accuracy of the difficult-miss group (M = 70%, SD = .1) was less than controls, but significantly higher than their 60% reliable aid, $t(14) = 3.89$, $p < .005$. Consistent with the easy-error hypothesis, accuracy of the easy-miss group (M = 55%, SD = .18) was significantly lower than that of control, $t(28) = 3.18$, $p < .005$, and difficult-miss groups $t(28) = 2.71$, $p < .01$, even though their aid was equally reliable. In fact, accuracy of the easy-miss group was significantly lower than the 60% accuracy of their aid, $t(14) = 1.23$, $p = .05$. The easy-miss group (M = 62%, SD = .14) agreed with the aid significantly less than the difficult-miss group (M = 71%, SD = .009), $t(28) = 2.04$, $p = .05$, but was

more confident ($M = 3.36$, $SD = .5$) than both difficult-miss ($M = 2.82$, $SD = .74$), $t(28) = 2.3$, $p < .05$, and control groups ($M = 2.68$, $SD = .67$), $t(28) = 3.12$, $p < .005$. Confidence did not differ between difficult-miss and control groups.

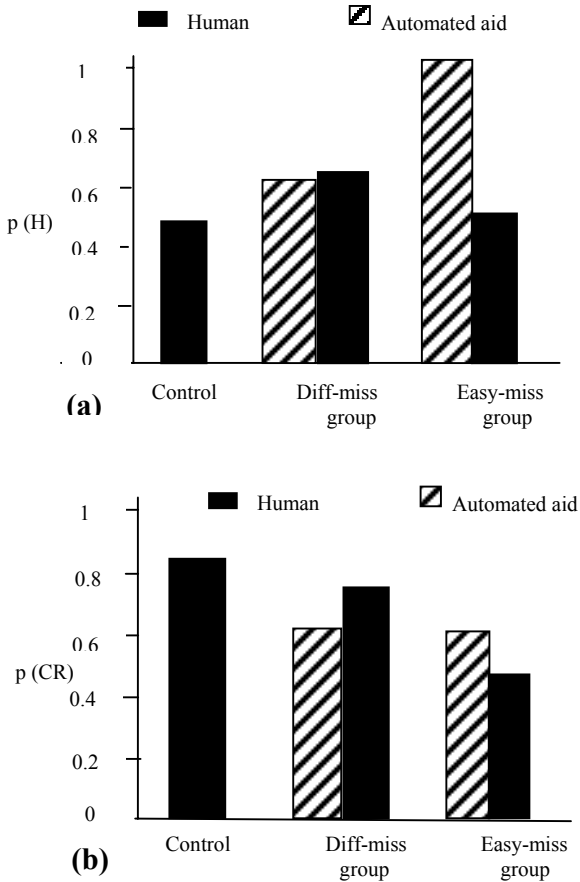


Figure 1. Accuracy of each group and diagnostic aid on (a) difficult-target trials, and (b) noise trials.

Subjective Measures

As expected, trust in the aid was lower in the easy-miss group ($M = 2.67$, $SD = 1.72$) than in the difficult-miss group ($M = 3.87$, $SD = 2.1$), $t(28) = 1.71$, $p = .07$. The former also had significantly lower reliability estimates ($M = 53.2$, $SD = 18.96$) than the latter ($M = 69.67$, $SD = 15.98$), $t(28) = 2.57$, $p < .05$. The easy-miss group significantly underestimated the true 70% reliability of the aid ($M = 53.2\%$, $SD = 18.96$), $t(14) = 10.73$, $p < .001$, while the difficult-miss group was generally accurate.

DISCUSSION

The results of the study support the “easy errors” hypothesis that automation failures on tasks easily performed by operators severely undermines automation trust and reliance. Participants utilizing an aid that missed easy targets had lower estimates of trust and aid reliability than participants whose aid missed only difficult targets. The easy-miss group also exhibited automation under-reliance on difficult target trials, disagreeing with the aid approximately 50% of the time, even though the aid was 100% accurate on these trials. As a result, their target detection performance was virtually equivalent to the unaided control group and significantly less than that of participants in the difficult-miss group, whose aid was only 60% reliable on difficult-target trials. This is in keeping with findings by Dzindolet, et al (in press) that automated aids that make even half as many errors as human operators lead to a rapid decline in automation trust and a less than average rating of the aid’s trustworthiness.

Participants in the easy-miss group also disagreed with the aid more often on noise trials than did the difficult-miss group, even though the aids in both groups were equally reliable. Although this may seem justifiable, given the aid was less accurate than controls on noise trials, participants in the easy-miss group actually performed worse than automation, suggesting that they were intentionally contradicting the aid. Furthermore, the easy-miss group had significantly higher confidence ratings than the difficult-miss and control groups, suggesting that they were overconfident in their abilities. Such over-confidence coupled with the tendency to intentionally contradict the aid may reflect automation *defiance* rather than the more traditional automation under-reliance. This overconfidence is likely to have been a result of users being able to detect easy automation errors more frequently than difficult errors, thereby raising their confidence in their ability to appropriately calibrate their reliance on automation with the actual reliability of the aid.

The results of the present study cannot be attributed solely to the conspicuity of errors committed by automation (Dzindolet et al., in press). Both the easy-miss and difficult-miss

groups observed all types of errors made by automation during testing. Therefore, it appears that the “easiness” of the error committed by an aid is a factor affecting trust above and beyond conspicuity (albeit conspicuity is necessary for operators to know the easiness of the error).

In the real world, such “easy” misses of targets may occur when the algorithm used by the aid to discriminate between noise and signal is insufficient to capture all instances of a particular target. For example, there are extreme differences between the physical characteristics of certain categories of weapons, such as the differences between a sword and a folded pocketknife or a derringer pistol and a machine gun. While clear images of all of these weapons would likely be obvious to a human operator, they may be unclear to an automated aid that functions on rigid algorithms thereby increasing the probability of automation errors that appear “easy” to the human operator.

Conclusions

The results of the present study suggest that users’ interaction with an aid that makes errors on easy tasks results in a greater reduction in trust and reliance, than when interacting with an aid that makes errors only on difficult tasks while reliably performing easy tasks. Several questions, however, remain concerning the effects that easy errors have on operators’ trust in automated diagnostic aids. In the present study, the only misses made by the aid in the easy-miss group were easy. However, such is not likely to be the case in the real world, where automated aids are more likely to miss difficult targets and only occasionally miss easy targets. Do such “occasional” easy misses also undermine trust when the aid makes more difficult than easy misses? Also, what effect will easy false alarms have on automation trust (e.g., an aid inaccurately diagnosing a hair-dryer for a handgun)? Clearly, further research is needed to address these issues.

REFERENCES

Breznitz, S. (1983). *Cry-wolf: The Psychology of False Alarms*. Mahwah, NJ: Erlbaum.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). Perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79-94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147-164.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (in press). The role of trust in automation reliance. *International Journal of Human-Computer Studies*.

Elvers, G. C., & Elrif, P. (1997). The effects of correlation and response bias in alerted monitor displays. *Human Factors, 39* (4), 570-580.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1* (1), 19-33.

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human machine systems. *Ergonomics, 22* (6), 671-691.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors, 43* (2), 217-225.

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

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

Ruble, D. N., & Stangor, C. (1986). Stalking the elusive schema: Insights from developmental and social-psychological analyses of gender schemas. *Social Cognition, 4* (2), 227-261.

Sheridan (2002). Human performance in relation to automation. In T. B. Sheridan (Ed.), *Human and Automation: System Design and Research Issues*. (pp. 69-89). Santa Monica, CA: John Wiley.

Sheridan, T., & Farrell, W. (1974). *Man-machine Systems: Information, Control, and Decision Models of Human Performance*. Cambridge, MA: MIT Press.

Smith, D. A., & Graesser, A. C. (1981). Memory for actions in scripted activities as a function of typicality, retention interval, and retrieval task. *Memory and Cognition, 9* (6), 550-559.

Swets, J. A. & Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance (3rd ed.)*. Upper Saddle River, NJ: Prentice Hall.

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users’ concurrence strategies. *Human Factors, 44*(1), 44-50.

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.