

Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids

Poornima Madhavan, Carnegie Mellon University, Pittsburgh, Pennsylvania, and Douglas A. Wiegmann and Frank C. Lacson, University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois

Objective: We tested the hypothesis that automation errors on tasks easily performed by humans undermine trust in automation. **Background:** Research has revealed that the reliability of imperfect automation is frequently misperceived. We examined the manner in which the easiness and type of imperfect automation errors affect trust and dependence. **Method:** Participants performed a target detection task utilizing an automated aid. In Study 1, the aid missed targets either on easy trials (easy miss group) or on difficult trials (difficult miss group). In Study 2, we manipulated both easiness and type of error (miss vs. false alarm). The aid erred on either difficult trials alone (difficult errors group) or on difficult and easy trials (easy miss group; easy false alarm group). **Results:** In both experiments, easy errors led to participants mistrusting and disagreeing more with the aid on difficult trials, as compared with those using aids that generated only difficult errors. This resulted in a downward shift in decision criterion for the former, leading to poorer overall performance. Misses and false alarms led to similar effects. **Conclusion:** Automation errors on tasks that appear “easy” to the operator severely degrade trust and reliance. **Application:** Potential applications include the implementation of system design solutions that circumvent the negative effects of easy automation errors.

INTRODUCTION

The use of automated diagnostic aids is becoming increasingly common in complex systems such as aviation, nuclear power, and health care as technology becomes more readily available. However, the extent to which automated aids will actually improve performance is difficult to predict, given that these aids are unlikely to be 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 & Ferrell, 1974). Nonetheless, the relationship between trust and operators' reliance upon imperfectly reliable automation is often complex and multifaceted (e.g., Lee & Moray, 1992, 1994; Lerch, Prietula, & Kulik, 1997; Muir, 1987, 1994; Sheridan, 2002; Wiegmann, Rich, & Zhang, 2001).

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 an automated 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 (i.e., self-reliance). In both cases, appropriate reliance occurs (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). However, overreliance or misuse can occur when an operator overtrusts an aid that is less reliable than unaided performance. Likewise, when operators undertrust an aid that is more reliable than manual performance, underreliance or disuse of automation can occur (Parasuraman & Riley, 1997).

According to the theory of signal detection (Green & Swets, 1988; Swets & Pickett, 1982),

two important stages of information processing in detection tasks are the aggregation of sensory evidence concerning the presence or absence of a signal and the decision about whether or not the sensory evidence indicates a signal. The availability of decision-aiding technology in recent times has provided operators with automated diagnostic aids that help perform both of these stages of information processing (Parasuraman, Sheridan, & Wickens, 2000). However, diagnostic aids are often imperfect, and decisions to utilize such aids depend on the user's quantitative weighting of the probability that the information presented by the aid is truly representative of the actual state of the system. Specifically, research has demonstrated that the choice of a utilization strategy depends on the user's perception of the trade-off between the true positive proportion, or hit rate, and the false positive proportion, or false alarm rate, of the aid (e.g., Elvers & Elrif, 1997; Getty, Swets, Pickett, & Gonthier, 1995; Maltz & Meyer, 2001; Meyer, 2001).

Previous research has demonstrated that accurate weighting of the actual reliability of automated aids rarely occurs (e.g., Wiegmann, 2002; Wiegmann et al., 2001). Many researchers have observed that operators underestimate the true reliability of imperfectly reliable automation (Wickens & Hollands, 2000; Wiegmann, 2002), and the extent to which perceived reliabilities deviate from true reliabilities often depends upon the type of error made by an 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") create undertrust in automation, resulting in operators not responding to alarms or other automation alerts (Parasuraman, Hancock, & Olofinboba, 1997). This underestimation of the aid's reliability and consequent loss of trust in the diagnostic aid might lead to a shift in criterion setting or decision bias for the human operator, as well as changes in sensitivity (Green & Swets, 1988), thereby affecting overall diagnostic performance.

A recent study by Cotte, Meyer, and Coughlin (2001) revealed that humans react differentially to automation errors depending on whether the errors are misses or false alarms. Specifically,

systems with frequent false alarms and relatively few misses led to a greater user propensity to agree with the aid when it said "no" (i.e., target absent) because of the aid's low miss rate and to disagree when the aid said "yes" (i.e., target present) because of the aid's high false alarm rate. Conversely, diagnostic aids that generated few false alarms but a large number of misses led users to disagree with the aid when it said "no" or "target absent" because of the aid's high miss rate and to agree when the aid said "yes" or "target present" because of the aid's low false alarm rate.

Another factor undermining trust is the conspicuity of automation failures. According to Dzindolet, Pierce, Beck, Dawe, and Anderson (2001), perceptions of the reliability of an automated diagnostic aid are filtered through the operator's "perfect automation schema," or the expectation that automation will perform at nearly perfect rates. This expectation leads operators to pay too much attention to errors made by an aid (Dzindolet, Pierce, Beck, & Dawe, 2002), thereby triggering a rapid decline in trust when diagnostic aids make errors (e.g., Dzindolet et al., 2001; Lee & Moray, 1992, 1994; Wiegmann et al., 2001). Indeed, researchers in cognitive psychology have found that information that contradicts expectations (e.g., schemas) is likely to be well remembered and plays an unduly large role in information processing (Ruble & Stangor, 1986; Smith & Graesser, 1981). Therefore, obvious errors made by an automated aid might dramatically reduce trust in the aid. Consistent with this obvious-error hypothesis, a recent study by Dzindolet et al. (2003) revealed that automation users who observed the errors made by an imperfect diagnostic aid trusted the aid less and were more likely to rely on their own decisions, as compared with 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. (2003) 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 that of the human operator.

Such statements may be attributable to the finding that humans in general hold exaggerated opinions of their own competence and greatly overestimate their own expertise while performing seemingly easy tasks (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Moreover, the mistakes of others are often attributed to their incompetence at performing a particular task (Oesch & Murnighan, 2003). In the context of automated aids, it is possible that automation errors that are perceived as easy by the human operator might lead to a rapid fall in the perceived competence of automation, particularly in the face of one's own exaggerated assessments of self-competence. Furthermore, the failure of automation to perform a task that appears easy or intuitive to the human operator would likely confuse operators, thereby making it difficult or impossible to comprehend the reasons for the "behavior" of automation. Such incomprehensibility of aid behavior is likely to result in a reduction in operator trust in the aid. Consequently, when operators catch easy mistakes made by an automated aid, it may serve to bolster their self-confidence in their own ability to outperform the aid while reducing trust in the aid, even on difficult tasks. As a result, inappropriate self-reliance or automation underreliance may occur. This "easy-error" hypothesis, however, has yet to be empirically tested.

STUDY 1

Easy Misses Versus Difficult Misses

The purpose of the first 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. We tested this hypothesis by using a generic target detection paradigm. This generic task was designed to mimic the basic characteristics of most real-world tasks that require some form of target detection performance by the operator. For example, airline luggage inspectors are required to detect the presence of contraband substances by observing X-ray images of passenger baggage, and quality control inspectors need to inspect an array of objects to detect the presence of a flaw in

manufacturing. The logic underlying such tasks, (i.e., detecting a target feature or object among several nontargets or distractors) is very similar to that underlying the simple generic target detection paradigm used here.

Specifically, participants performed a signal detection task in which they were required to find target letters embedded in an array of letter distractors. On some trials, the number of distractors was small, making the target easy to detect, whereas on other trials numerous distractor 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 missed targets on only easy trials. For the other half, the aid missed targets on only difficult trials. A third, unaided, group served as a control.

Based on the easy-error hypothesis, we predicted that compared with an aid that generated difficult errors alone, an aid that generated easy misses would (a) have a greater impact on subjective trust and perceived reliability of the aid and (b) lead to higher confidence estimates on difficult trials and lower reliance on the aid (i.e., greater disagreement), leading to poorer target detection performance than was ultimately possible.

Method

Participants. Forty-five undergraduate and graduate students from the University of Illinois completed all phases of the experiment. Participants were paid \$8.00 for their participation, and participation time did not exceed 1 hr.

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 alphanumeric characters on a screen. The simulation was developed using Visual Basic for MS-DOS and presented on a desktop computer equipped with a 22-inch (55.9-cm) color monitor and standard keyboard. Out of the 200 trials, 50% ($n = 100$) were target trials, in which the *X* was present among the noise characters, whereas the other 50% ($n = 100$) were noise trials, in which the *X* was absent. Out of the 100 target trials, 50% ($n = 50$) were easy (90 targets embedded in 1000 distractor characters) and 50% ($n = 50$) were difficult (5 targets embedded in 1000 distractor characters). In other words, each screen contained 1000 letters of the alphabet

painted in white over a black background, out of which the number of *X* symbols was 90, 5, or none, depending on trial type. Each letter was 6 × 6 mm in size. The relative clutter of symbols on the screen was similar for all types of trials, and the only feature that distinguished the three types of trials was the number of *X* symbols painted on the stimulus screen. The 200 trials flashed on the screen in random order. Participants were presented a brief demonstration of each type of trial as well as verbal descriptions of the same.

At the onset of each trial, the screen with letters flashed on the monitor for 750 ms, after which the stimulus screen was masked by a blank screen. After a delay of 5 s, aided participants ($n = 30$) were presented with the decision of a diagnostic aid to help them determine the presence or absence of the *X*. The likelihood that the aid made an error was .20 and .40 for misses and false alarms, respectively. Participants were not informed as to the actual reliability of the aid before testing.

For the *difficult miss group* ($n = 15$), the aid generated 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 used the *Y* and *N* keys on the computer keyboard to indicate whether or not they thought the target was present in the array. A third group, the unaided *control group* ($n = 15$), performed the same task without the diagnostic aid. On each trial, participants indicated their confidence in their diagnosis on a scale ranging from 1 (*no confidence*) to 5 (*very confident*).

Feedback was presented in the form of single line on the screen indicating whether or not the target was present. The main purpose of providing feedback was to make automation errors more salient, so as to ensure that participants did not fail to notice errors made by the aid and be unduly influenced by the conspicuity of errors instead of their nature or frequency. However, the nature of feedback was such that it provided only an assessment of whether the participant's final response was correct or incorrect, along with the participant's cumulative score from which he or she was expected to draw inferences about the

accuracy of the aid. Ten points were awarded for correct decisions and subtracted for incorrect answers. After feedback, participants could activate the appearance of the next stimulus screen by pressing the Enter key. Although they were allowed to pause between trials (i.e., after feedback but before pressing Enter to activate the next trial), they were instructed not to pause during any particular trial. Participants were instructed that accuracy was more important than speed and that the goal was to earn as many points as possible.

Following completion of the 200 trials, aided participants were administered a postexperimental questionnaire that required them to estimate the diagnostic aid's overall reliability as well as reliability on each type of trial using a percentage scale that ranged from 0 to 100 and to indicate their trust in the aid using a scale that ranged from 0 (*did not trust at all*) to 8 (*trusted all the time*).

Results

Trial types (easy target, difficult target, noise) were randomly distributed during testing. However, for analysis purposes, data for computing target detection accuracy and agreement with the aid (for aided groups) were grouped into blocks according to the type of trial. Overall, accuracy and agreement scores were computed for the three types of trials: easy target trials ($n = 50$), difficult target trials ($n = 50$), and noise trials ($n = 100$).

Planned comparisons between the two automated aid groups were performed using two-tailed *t* tests. Although some of these comparisons were hypothesized to be in a particular direction, others were not. Consequently, we chose to use two-tailed rather than one-tailed tests in order to maintain consistency and reduce the likelihood of making a Type I error. For variables in which multiple planned comparisons were performed between the automation and control groups, a Scheffe test was utilized in order to control for Type I error rates across all possible comparisons. One-sample *t* tests (Milton & Corbet, 1979) were also used to compare performance data with an expected value (e.g., actual aid reliability). An overall alpha rate of $p = .05$ was used as a criterion for all tests and is implied when differences are referred to as statistically significant. Effect sizes are given as Cohen's *d* with values of

TABLE 1: Subjective Trust and Reliability Estimates of the Automated Diagnostic Aid in Study 1

Variable	Group	
	Easy Miss (n = 15)	Difficult Miss (n = 15)
Trust rating	2.67 (1.72) ^a	3.87 (2.10)
Aid reliability on		
Easy target trials		
Actual	60.00	100.00
Estimated	60.00 (29.46) ^a	86.27 (20.7) ^b
Difficult target trials		
Actual	100.00	60.00
Estimated	51.87 (20.18) ^{a,b}	62.33 (23.82)
Noise trials		
Actual	60.00	60.00
Estimated	50.67 (21.2)	56.07 (23.02)

Note: Standard deviations in parentheses.

^aIndicates that the two groups differ significantly based on a two-tailed t test, $p \leq .05$. ^bIndicates that the value differs significantly ($p \leq .05$) from the known reliability of the aid based on a one-sample t test.

.20, .50, and .80 reflecting small, medium, and large effects, respectively (Cohen, 1988).

Trust and perceived reliability. Trust in the diagnostic aid was assessed on an 8-point scale in which 1 indicated *did not trust at all* and 8 indicated *trust all the time*. See Table 1 for means and standard deviations. As expected, participants' estimates of trust in the diagnostic aid were generally lower in the easy miss group than in the difficult miss group, $t(28) = 1.71, p < .05, d = 0.65$. In addition to trust, participants also estimated their perceived reliability of the aid on each of the three types of trials (i.e., easy targets, difficult targets, and noise trials). Means and standard deviations for these estimates are also presented in Table 1. A 2 (group: difficult miss and easy miss) \times 3 (trial type: easy target, difficult target, and noise) analysis of variance (ANOVA) on perceived reliability ratings did not reveal a significant interaction between group and trial type, $F(2, 56) = 1.93, p = .16$. However, there were significant main effects for trial type, $F(2, 56) = 7.19, p < .01$, as well as for group, $F(1, 28) = 6.22, p < .05$.

Perceived aid reliabilities of each group were then compared with the true reliability of their respective diagnostic aids on each type of trial using one-sample *t* tests. On easy target trials, the

difficult miss group underestimated the true 100% reliability (i.e., perfect hit rate) of their diagnostic aid, $t(14) = 2.57, p < .05, d = 0.94$, whereas the easy miss group accurately estimated the hit rate of their aid to be roughly 60%, $t(14) = 0, ns, d = 0$. In contrast, on difficult target trials, the easy miss group significantly underestimated the hit rate of their diagnostic aid, indicating that the aid was roughly 52% accurate when it was actually 100% accurate on these trials, $t(14) = 9.24, p < .01, d = 3.37$, whereas the difficult miss group accurately perceived the 60% reliability of their aid, $t(14) = .38, ns, d = 0.14$. Both aided groups were reasonably accurate in their perceptions of the 60% correct rejection rates of their respective aids on the noise trials.

Aid utilization. The percentage of trials in which participants in the aided groups agreed with their diagnostic aid across the different types of trials is presented in Table 2. As expected, on easy target trials, the difficult miss group never disagreed with the automated aid, given that the aid never made an error on these trials. Participants in the easy miss group appropriately agreed with the aid on roughly 60% of the easy target trials, given that their aid was accurate on only 60% of these trials. However, on difficult target trials, participants in the easy miss group agreed

TABLE 2: Percentage of Trials in which Participants Agreed with the Automated Diagnostic Aid in Study 1

Trial Type	Group	
	Easy Miss (n = 15)	Difficult Miss (n = 15)
Easy target trials	59 (0.008) ^a	100 (0.0)
Difficult target trials	50 (0.16) ^a	64 (0.12)
Noise trials	62 (0.14) ^a	71 (0.009)

Note: Standard deviations in parentheses.

^aIndicates that the two groups differ significantly based on a two-tailed t test, $p \leq .05$.

with their automated aid on only 50% of the trials, even though their aid was 100% accurate on those trials. Their agreement rate was significantly less than that of participants in the difficult miss group, whose aid was only 60% reliable, $t(28) = 2.66, p < .01, d = 0.99$. Analysis of the noise trials revealed a smaller yet significant difference between aided groups. Participants in the easy miss group agreed with the aid significantly less on noise trials than did participants in the difficult miss group, although both groups utilized aids of equal (60%) correct rejection rates, $t(28) = 2.04, p < .05, d = 0.91$.

Performance accuracy and bias. Analyses of sensitivity (d') and bias (β) scores indicated that the type of errors committed by the diagnostic aid

had little effect on participants' sensitivities but did have a large impact on decision biases (see Table 3). Participants in the easy miss group had a somewhat greater tendency to say "no" when the aid indicated that a target was present and to say "yes" when the aid indicated that the target was absent, relative to the other two groups. As a result, the easy miss group had fewer hits on difficult target trials as compared with the difficult miss group and had more false alarms on noise trials than did both the difficult miss and control groups (Table 3). This resulted in a lowered criterion setting (β) for the easy miss group relative to that for the difficult miss ($d = .31$) and control groups ($d = .29$) as indicated by Scheffe tests. Hit rates did not differ across groups for easy target trials.

TABLE 3: Sensitivity, Bias, and Accuracy Scores Across Experimental Groups in Study 1

Variable	Group		
	Easy Miss (n = 15)	Difficult Miss (n = 15)	Control (n = 15)
Sensitivity (d')	1.49 (0.32)	1.52 (0.36)	1.68 (0.39)
Bias (β)	0.53 (3.2) ^{a,b}	1.24 (0.63)	1.23 (1.36)
Easy target hits	1.00 (0)	1.00 (0)	1.00 (0)
Difficult target hits	0.50 (0.12) ^a	0.61 (0.12) ^c	0.50 (0.13)
False alarms	0.45 (0.18) ^{a,b}	0.30 (0.10) ^c	0.19 (0.13)

Notes. Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the difficult miss group. ^bIndicates that the easy miss group differs significantly from the control group. ^cIndicates that the difficult miss group differs significantly from the controls group. Significance levels are calculated using a two-tailed Scheffe test controlling for Type I error rate at $p = .05$ across all possible between-groups comparisons for an individual variable.

TABLE 4: Confidence Estimates Across Experimental Groups in Study 1

Variable	Group		
	Easy Miss (n = 15)	Difficult Miss (n = 15)	Control (n = 15)
Easy target trials	3.71 (0.34)	3.69 (0.53)	3.67 (0.40)
Difficult target trials	1.89 (0.55) ^b	1.92 (0.45) ^c	3.41 (0.57)
Noise trials	3.36 (0.50) ^{a,b}	2.82 (0.74)	2.68 (0.67)

Note: Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the difficult miss group. ^bIndicates that the easy miss group differs significantly from the control group. ^cIndicates that the difficult miss group differs significantly from the control group. Significance levels are calculated using a two-tailed Scheffe test controlling for Type I error rate at $p = .05$ across all possible between-groups comparisons for an individual variable.

Confidence estimates. Confidence in decisions was assessed on a 5-point scale in which 1 indicated *not confident at all* and 5 indicated *very confident*. See Table 4 for means and standard deviations. Scheffe tests were used to analyze differences in these scores across groups for each type of trial. Results indicated that confidence estimates on easy target trials were generally high and did not differ significantly across groups. Analysis of confidence estimates on difficult target trials revealed no difference between the two aided groups. However, both aided groups were significantly less confident than were the control participants (difficult miss group: $d = 2.9$; easy miss group: $d = 2.71$). Analysis of confidence scores on noise trials revealed that the easy miss group had significantly higher confidence scores than did participants in both the difficult miss and control groups. Confidence scores on noise trials did not differ significantly between the difficult miss and control groups.

Discussion

The results of Study 1 support the easy-error hypothesis that automation failures on tasks easily performed by operators undermine trust and automation reliance. Participants who utilized an aid that missed only easy targets had significantly lower estimates of trust and aid reliability than did participants whose aid missed only difficult targets. Participants in the easy miss group also exhibited automation underreliance 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 that of 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.

The general inclination of the easy miss group to disagree with the aid on difficult (target and noise) trials resulted in a downward shift in bias, more so than in a decrease in sensitivity. As a result, the easy miss group had a relatively equal probability of a hit versus a false alarm on difficult trials, which differed from the other two groups, who had generally more hits than false alarms. Consequently, the easy miss group had lower overall hit rates on difficult target trials as well as higher false alarm rates on noise trials, leading to poorer target detection performance than was ultimately possible. Conversely, there is a hint of evidence suggesting that when aids are 100% reliable on easy tasks, overtrust occurs, resulting in overreliance on difficult tasks. This is evidenced by the fact that the difficult miss group overrelied on their 60% reliable aid on noise trials (though the aid was less accurate than the average unaided participant), ultimately generating a larger number of false alarms as compared with the unaided control participants.

In keeping with the findings by Cotte et al. (2001), participants in both aided groups were expected to disagree with the aid when it said "target present" and to agree with the aid when it said "target absent," owing to both diagnostic aids' relatively high (40%) false alarm rates. However, participants in the easy miss group also disagreed with the aid more often on noise trials than did

participants in the difficult miss group, even though the aids in both groups were equally reliable. Although this behavior may seem justifiable, given that the aid was ultimately less accurate than the unaided controls on noise trials, participants in the easy miss group actually performed worse than the automation, suggesting that they were intentionally contradicting the aid.

In addition, participants' confidence ratings in the easy miss group were significantly higher than those in both the difficult miss and control groups, suggesting that the easy miss participants were overconfident in their abilities. Such overconfidence coupled with the tendency to repeatedly contradict the aid may actually reflect automation *defiance* rather than the more traditional automation underreliance. Such an effect has rarely been substantiated empirically in the literature. However, before the findings of this study can be interpreted, several questions must be answered concerning the effects that easy errors have on operators' trust in and reliance on automated diagnostic aids. These concerns were addressed in Study 2.

STUDY 2

Occasional Easy Misses Versus False Alarms

Study 1 demonstrated that aids that generate errors (misses) only on tasks that are perceived as easy are more likely to be mistrusted and underrelied on than are aids that generate errors only on difficult tasks. However, in Study 1 the only misses made by the aid in the easy miss group were easy targets. 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. Therefore, the main purpose of Study 2 was to test the hypothesis that occasional easy errors interspersed among difficult errors will undermine trust in automation more than when automation fails only on difficult tasks.

In addition to misses, research has also shown that automated aids that produce a large number of false alarms (i.e., that "cry wolf"; Breznitz, 1983) create undertrust in automation (Gupta, Bisantz, & Singh, 2001; Parasuraman et al., 1997) and consequently affect user compliance with automated aids. Although relatively little research appears to have examined the relative conse-

quences of false alarms versus misses in influencing human trust and reliance on automated aids, a few recent studies suggest that false alarms may indeed be more degrading of trust than are misses (Cotte et al., 2001; Gupta et al., 2001; Maltz & Meyer, 2001). In the context of automated alarm systems, a false alarm directs the operator's attention away from other important tasks, ultimately resulting in the operator wasting time and effort in dealing with the alarm. Therefore operators are likely to lose trust in such a system that demands this extra effort (Thomas, Wickens & Rantanen, 2003).

Furthermore, the consequences of false alarms are often immediate, whereas the consequences of misses are generally delayed. For instance, in the context of airline luggage screening, a false alarm generated by misdiagnosing a harmless object as a weapon will likely be apparent immediately, when the luggage is rechecked, whereas a failure to diagnose (or miss) the presence of a weapon may not be apparent until much later. Consequently, false alarms have been shown to have a greater effect than misses on operator trust in difficult tasks. However, the effect of *easy* false alarms (e.g., an aid inaccurately diagnosing a hair dryer as a handgun) on trust relative to *easy* misses remains unexplored. Therefore, the second objective of Study 2 was to extend the findings of Study 1 by examining whether easy false alarms might affect user trust and reliance in a manner akin to easy misses.

In order to address these issues, we used a paradigm similar to that used in Study 1. Specifically, participants performed a signal detection task in which some trials were easy and some were difficult. Three groups of participants utilized a diagnostic aid that had a .3 miss rate and a .3 false alarm rate. For one group, the aid made errors only on difficult trials (difficult noise and target trials). For the other two groups the aid made errors across a mixture of difficult and easy trials, with easy errors being either solely misses or solely false alarms. A fourth, unaided, group served as a control.

Based on the easy-error hypothesis, as in Study 1, we predicted that (a) participants who utilized aids that generated easy as well as difficult errors would have lower levels of trust in and perceived reliability of the aid than those using an aid that generated only difficult errors, and (b)

participants in the two easy errors groups would rely less on the aid (i.e., disagree more) and have higher confidence ratings on difficult trials (target and noise) as compared with the difficult errors only group.

Based on Cotte et al.'s (2001) finding on differential user reactions to automation misses and false alarms, we predicted that (a) participants using the aid that generated false alarms on easy trials would disagree more frequently with the aid's diagnoses of "target present" (i.e., demonstrating lower compliance) than would participants using an aid that generated only misses on easy trials, and (b) consequently, the easy false alarm group would be less accurate than the easy miss group on difficult target trials and more accurate on difficult noise trials. Similar to Study 1, here the appropriateness of these reliance strategies can be determined only by comparing the performance of aided groups with the unaided control group.

Method

Participants. Sixty students from the University of Illinois completed all phases of the experiment. Participants were paid \$8.00 for their participation, and participation time did not exceed 1 hr.

Tasks and procedures. Participants performed 200 trials of a computer simulation task similar to that in Study 1. Out of the 200 trials, 50% ($n = 100$) were target trials and the other 50% ($n = 100$) were noise trials. Out of the 100 target trials, 50% ($n = 50$) were easy (90 targets embedded in 1000 distractor characters) and 50% ($n = 50$) were difficult (5 targets embedded in 1000 distractor characters). Likewise, out of the 100 noise trials, 50% ($n = 50$) were easy (5 distractor characters scattered on the screen) and 50% ($n = 50$) were difficult (1000 distractor characters closely spaced on the screen). The relative clutter on the stimulus screen was similar for the target trials and the difficult noise trials. The only exceptions were the easy noise trials, which were characterized by fewer and more sparsely distributed noise characters on the screen. The 200 trials flashed on the screen in random order, and participants were apprised of the four different types of trials in a manner similar to that in Study 1.

Aided participants ($n = 45$) were presented with the decision of a diagnostic aid (in the same

manner as in Study 1) to help determine the presence of the X. The likelihood of the aid presenting either a correct (hit/correct rejection) or an incorrect diagnosis (miss/false alarm) was .70 and .30, respectively. The reliability of the aid in each of the four types of trials varied as per experimental group: (a) For the *difficult errors group* ($n = 15$), the aid committed 30 misses on the 50 difficult target trials (40% accuracy rate) and 30 false alarms on the 50 difficult-noise trials (40% accuracy rate) but was 100% accurate on the 50 easy target and 50 easy-noise trials. (b) For the *easy miss group* ($n = 15$), the aid committed 20 misses on the 50 difficult target trials (60% accuracy rate), 10 misses on the 50 easy target trials (80% accuracy rate), and 30 false alarms on the 50 difficult noise trials (40% accuracy rate), but it was 100% accurate on easy noise trials. (c) For the *easy false alarm group* ($n = 15$), the aid committed 30 misses on the 50 difficult target trials (40% accuracy rate), 10 false alarms on the 50 easy noise trials (80% accuracy rate), and 20 false alarms on the 50 difficult noise trials (60% accuracy rate), but it was 100% accurate on easy target trials. (d) The *control group* ($n = 15$) performed the task without the diagnostic aid.

Aided participants estimated the aid's reliability and their trust in the aid on a postexperimental questionnaire using scales similar to those used in Study 1.

Results

Similar to Study 1, data for estimated target detection accuracy and agreement with the aid (for aided groups) were grouped into blocks according to the type of trial. Thus, accuracy scores and agreement scores were grouped into four blocks: easy target trials ($n = 50$), difficult target trials ($n = 50$), easy noise trials ($n = 50$), and difficult noise trials ($n = 50$).

Trust and perceived reliability. Similar to Study 1, trust in the diagnostic aid was assessed on an 8-point scale in which 1 indicated *did not trust at all* and 8 indicated *trust all the time*. Means and standard deviations of subjective trust and perceived reliability are presented in Table 5. Scheffe tests indicated that as expected, both easy errors groups trusted the aid significantly less than did the difficult errors group (easy miss group: $d = 0.96$; easy false alarm group: $d = 0.79$). Trust did not differ significantly between the easy miss and

TABLE 5: Subjective Trust and Reliability Estimates of the Automated Diagnostic Aid in Study 2

Variable	Group		
	Difficult Errors (<i>n</i> = 15)	Easy Miss (<i>n</i> = 15)	Easy False Alarm (<i>n</i> = 15)
Trust rating	5.20 (1.9)	3.27 (2.12) ^a	3.60 (2.29) ^b
Aid reliability on			
Easy target trials			
Actual	100.00	80.00	100.00
Estimated	88.33 (13.84) ^c	82.00 (21.36)	82.93 (19.82) ^c
Easy noise trials			
Actual	100.00	100.00	80.00
Estimated	77.67 (22.59) ^c	90.00 (18.89) ^c	66.47 (29.97)
Difficult target trials			
Actual	40.00	60.00	40.00
Estimated	51.33 (21.91)	45.33 (24.31) ^c	66.53 (19.45) ^c
Difficult noise trials			
Actual	40.00	40.00	60.00
Estimated	51.33 (18.07) ^c	42.67 (24.78)	62.87 (19.61)

Note: Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the difficult errors group; ^bindicates that the easy false alarm group differs significantly from the difficult errors group, based on a two-tailed *t* test, $p \leq .05$. ^cIndicates that the value differs significantly ($p \leq .05$) from the known reliability of the aid based on a one-sample *t* test.

easy false alarm groups. As in Study 1, a 3 (group: difficult errors, easy miss, and easy false alarm) \times 4 (trial type: easy target, difficult target, easy noise, and difficult noise) ANOVA on perceived reliability ratings revealed a significant interaction between group and trial type, $F(6, 126) = 5.09$, $p < .01$, as well as a significant main effect for trial type, $F(3, 126) = 34.9$, $p < .01$. However, the main effect for group failed to reach significance, $F(2, 42) = 0.4$, $p = .67$.

These ANOVA results were further examined by comparing perceived reliability estimates with true aid reliabilities on each type of trial using one-sample *t* tests. The diagnostic aid was 100% reliable on easy target trials for the difficult errors and easy false alarm groups. However, both of those groups significantly underestimated the aid's accuracy rate – difficult errors group: $t(14) = 3.26$, $p < .01$, $d = 1.19$; easy false alarm group: $t(14) = 3.34$, $p < .01$, $d = 1.22$ – whereas the easy miss group accurately perceived the 80% reliability of their diagnostic aid on easy target trials, $t(14) = 0.36$, *ns*, $d = 0.13$. Contrary to the results for easy target trials, on easy noise trials the easy miss group significantly underestimated the 100% reliability of their aid, $t(14) = 2.05$, $p < .05$,

$d = 0.75$, whereas participants in the easy false alarm group were generally accurate in their estimates of the aid's 80% reliability, $t(14) = 1.75$, *ns*, $d = 0.64$. The difficult errors group underestimated the aid's 100% reliability on easy noise trials, $t(14) = 3.83$, $p < .01$, $d = 1.4$. Similar to the easy noise trials, on the difficult target trials the easy miss group significantly underestimated the true 60% reliability of their aid, $t(14) = 2.34$, $p < .05$, $d = 0.85$, whereas the easy false alarm group overestimated the accuracy of their 40% reliable aid, $t(14) = 5.29$, $p < .01$, $d = 1.93$. The difficult-errors group was generally accurate in estimating the reliability of their 40% accurate aid. On difficult noise trials, the easy miss and easy false alarm groups accurately estimated the reliabilities of their aids, which were 40% and 60% accurate, respectively. However, the difficult errors group significantly overestimated the 40% reliability of their aid, $t(14) = 2.43$, $p < .05$, $d = 0.87$.

Aid utilization. The percentage of trials in which aided groups agreed with their diagnostic aid across four types of trials is presented in Table 6. As expected, the difficult errors group never disagreed with the aid on easy trials (target and noise), given that their aid never made errors on

easy trials. Likewise, the easy miss group never disagreed with the aid on easy noise trials, and the easy false alarm group never disagreed with the aid on easy target trials, given that their aids never made errors on these trials. The easy miss group and easy false alarm groups appropriately agreed with the aid on roughly 80% of the easy target trials and easy noise trials, respectively, as their aids were 80% accurate on these trials.

As expected, the easy false alarm group demonstrated a slightly different pattern of compliance with their aid as compared with the easy miss group, although these differences were not statistically significant. The easy false alarm group agreed (i.e., complied) with the aid less than did the easy miss group, $t(28) = 1.9, ns, d = 0.72$, on difficult target trials. Although this may seem justifiable given that the easy false alarm aid was less accurate than the easy miss aid on difficult target trials, the easy false alarm group also complied with the aid less, $t(28) = 1.7, ns, d = 0.64$, than did the easy miss group on difficult-noise trials, despite the easy false alarm aid being more accurate on these trials. The statistical nonsignificance of these results suggests that the compliance strategies of the two groups did not differ significantly from chance expectations. However, medium to large effect sizes indicate that the magnitude of the difference between groups (i.e., the relative degree of compliance with the aid) was large, thereby suggesting potentially important (albeit nonsignificant) behavioral differences between the easy miss and easy false alarm groups.

Performance accuracy and bias. Similar to Study 1, analysis of sensitivities did not reveal a significant difference between any of the experimental or control groups (see Table 7). Again similar to Study 1, the tendency of participants in the easy miss group to say “no” when the aid indicated that a target was present, along with their tendency to say “yes” when the aid indicated that the target was absent, resulted in a significant shift in decision criterion (β) for the easy miss group relative to the control group, $d = 0.93$, as indicated by a Scheffe test. As a result, the easy miss group had more hits on difficult target trials and more false alarms on noise trials than did the control group (see Table 7). The opposite was true for the easy false alarm group, which had fewer hits on difficult target trials (owing, in part, to the fact that their aid was less accurate on these trials) and fewer false alarms on difficult noise trials relative to the easy miss group. Hit rates did not differ across groups for easy target trials.

Confidence estimates. As in Study 1, confidence in decisions was assessed on a 5-point scale in which 1 indicated *not confident at all* and 5 indicated *very confident*. See Table 8 for means and standard deviations. Scheffe tests were used to analyze differences in these scores across groups for each type of trial. Results indicated that confidence estimates on easy target and easy noise trials were generally high and did not differ significantly across groups. Analysis of confidence estimates on difficult target trials revealed that confidence estimates of the easy miss group were significantly higher than those of the difficult

TABLE 6: Percentage of Trials in Which Participants Agreed with the Automated Diagnostic Aid in Study 2

Trial Type	Group		
	Difficult Errors (n = 15)	Easy Miss (n = 15)	Easy False Alarm (n = 15)
Easy target trials	100 (0)	78 (0.006) ^{a,b}	100 (0)
Easy noise trials	100 (0)	100 (0)	79 (0.006) ^c
Difficult target trials	54 (0.12)	79 (0.17) ^a	66 (0.21)
Difficult noise trials	53 (0.008)	61 (0.008) ^a	56 (0.11)

Note: Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the difficult errors group. ^bIndicates that the easy miss group differs significantly from the easy false alarm group. ^cIndicates that the easy false alarm group differs significantly from the difficult errors group, based on a two-tailed t test, $p \leq .05$.

TABLE 7: Sensitivity, Bias, and Accuracy Scores Across Experimental Groups in Study 2

Variable	Group			
	Difficult Errors	Easy Miss	Easy False Alarm	Control
Sensitivity (d')	1.85 (0.29)	1.80 (0.19)	1.79 (0.27)	1.93 (0.38)
Bias (β)	1.29 (0.63)	0.97 (0.87) ^b	1.43 (1.2)	1.84 (1.01)
Easy target hits	1.00 (0)	1.00 (0)	1.00 (0)	1.00 (0)
Easy noise false alarms	0 (0)	0 (0)	0 (0)	0 (0)
Difficult target hits	0.60 (0.15)	0.69 (0.14) ^{a,b}	0.58 (.1)	0.52 (0.01)
Difficult noise false alarms	0.33 (0.17)	0.43 (1.0) ^b	0.35 (0.17)	0.26 (0.15)

Note: Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the easy false alarm group. ^bIndicates that the easy miss group differs significantly from the control group. Significance levels are calculated using a two-tailed Scheffe test controlling for Type I error rate at $p = .05$ across all possible between-groups comparisons for an individual variable.

errors, $d = 0.68$, and control groups, $d = 0.71$. The easy false alarm group did not differ in confidence from any of the other groups. Similarly, analysis of confidence scores on noise trials revealed that the easy false alarm group was significantly more confident than were the difficult errors, $d = 0.85$, and control groups, $d = 0.80$. Confidence of the easy miss group did not differ significantly from that of the other three groups.

Discussion

The results of Study 2 partially support the easy-error hypothesis. Compared with participants utilizing automation that made only difficult errors, participants using an aid that generated easy errors trusted the aid less, underestimated its reliability, and were highly self-confident regardless of the accuracy of their diagnoses. Moreover, as expected, participants using an aid that generated easy false alarms demonstrated lower compliance with the aid on difficult trials, thereby lowering accuracy on difficult target trials and helping accuracy on difficult noise trials, relative to participants using the easy miss aid. These findings corroborate the findings of Study 1 that easy errors made by a diagnostic aid undermine automation trust and reliance. In addition, the results of Study 2 extend the findings of Study 1 by demonstrating that easy errors undermine trust and reliance even when the easy errors (a) are occasional and (b) consist of easy false alarms.

According to expectations, the two easy errors groups, on average, were significantly more confident than the difficult errors and control groups, although this inflated confidence was not justified by significantly higher target detection performance for either easy error group. Such overconfidence could result from participants overestimating their ability to detect automation errors on difficult trials when the aid committed obvious errors on corresponding easy trials. Thus, the easy miss and easy false alarm groups trusted the aid less and were overconfident in their target detection abilities when they believed they could easily “catch” automation errors.

Although the results of both studies revealed that operators tend to become more confident in their own decisions when the aid makes easy errors, there appear to be disparities in the manner in which decision confidence was affected by the simplicity of automation errors. In the first study, the easy miss group was significantly more confident than both the difficult miss and unaided groups on noise trials, despite being significantly less accurate than the latter groups, whereas this was not true on difficult target trials. Similarly, both easy errors groups were more confident than the difficult errors and control groups in Study 2, although this inflated confidence was not justified by greater target detection accuracy for the easy errors groups. This overconfidence is likely to have been a result of users being able to detect easy automation errors more frequently than difficult

TABLE 8: Participants' Confidence Estimates Across Experimental Groups in Study 2

Variable	Group			
	Difficult Errors	Easy Miss	Easy False Alarm	Control
Easy target trials	4.86 (0.005)	4.85 (0.007)	4.79 (0.14)	4.85 (0.007)
Easy noise trials	4.27 (0.57)	4.25 (0.63)	4.32 (0.54)	4.21 (0.62)
Difficult target trials	3.22 (0.62)	3.57 (0.43) ^{a,b}	3.50 (0.50)	3.27 (0.45)
Difficult noise trials	2.54 (0.74)	2.91 (0.56)	3.12 (0.66) ^{c,d}	2.64 (0.59)

Note: Standard deviations in parentheses.

^aIndicates that the easy miss group differs significantly from the difficult errors group. ^bIndicates that the easy miss group differs significantly from the control group. ^cIndicates that the easy false alarm group differs significantly from the difficult errors group. ^dIndicates that the easy false alarm group differs significantly from the control group. Significance levels are calculated using a two-tailed Scheffe test controlling for Type I error rate at $p = .05$ across all possible between-groups comparisons for an individual variable.

errors, thereby raising their confidence in their ability to appropriately calibrate their reliance on automation with the actual reliability of the aid.

In keeping with findings by Cotte et al. (2001), participants in the easy false alarm group demonstrated lower compliance with the aid (or lower tendency to agree with the aid when it said "target present"), resulting in reduced accuracy on difficult target trials and improved accuracy on difficult noise trials as compared with the easy miss group. This is corroborated by the recent findings of Meyer (2004), who suggested that responses to warnings or alarming systems can be dichotomized into "compliance" strategies (saying "yes" when the aid says "yes") and "reliance" strategies (saying "no" when the aid says "no"), and it is possible for either strategy to be independently affected by the properties of the aid. In the present experiment, perhaps compliance with the aid for the two easy errors groups on difficult trials was influenced by the probability of their having to disagree with the aid when it made obvious errors on corresponding easy trials. Furthermore, data revealed that easy misses led to participants adopting a decision criterion lower than that for easy false alarms, although the decision criteria were not significantly different for the two groups. A lower decision bias for the easy miss group could likely have led easy miss participants to outperform the easy false alarm group on target trials while performing worse than the latter on noise trials. The easy miss group's underestimation of aid reliability was the only significant subjective difference between the easy miss and easy false

alarm groups, suggesting that automation errors on tasks easily performed by humans undermine subjective trust regardless of whether these easy errors constitute misses or false alarms.

The costs associated with misses and false alarms in the present study were identical (i.e., participants lost 10 points for each error generated regardless of error type). Such is rarely the case in real-world scenarios such as airline luggage screening, wherein a miss is invariably costlier than a false alarm. Perhaps the lack of differential costs for errors in the present study led to participants perceiving and reacting to easy automation misses in a manner indistinguishable from easy false alarms. Furthermore, research has documented that in real-world complex systems, misses have a smaller effect on trust than do false alarms (see Thomas et al., 2003), as misses are more likely to go undetected for longer periods. Such was not the case in the present study, as participants received immediate and direct feedback regarding their target detection performance, which in the present context might have further nullified the differential effects of easy misses versus false alarms on automation trust and reliance.

Nonetheless, the lack of an effect of easy errors relative to difficult errors on overall agreement with the aid suggests that occasional easy errors distributed among many difficult errors, as is characteristic of most real-world tasks, undermine trust and reliance in a manner different from that found when all the errors committed are easy in nature. Specifically, participants in the easy miss

and easy false alarm groups distrusted the aid and were overconfident on difficult trials, suggesting a negative impact on automation trust. However, the two groups were appropriately calibrated with the aid's true reliability, as they attained more or less the same level of accuracy as the unaided participants on difficult trials, suggesting the lack of an undermining effect on automation reliance. Therefore, participants in Study 2 failed to demonstrate the automation defiance observed in Study 1, in which the easy miss group repeatedly disagreed with the automated aid while completely disregarding its true 100% accuracy rate. An alternative explanation for the lack of a strong defiance effect in Study 2 is that a smaller number of easy errors led to a proportionately small effect on trust and reliance. This is evidenced by the fact that the degree to which the easy miss group erred in their judgment of the aid's reliability was similar for both studies.

Furthermore, estimates of trust in the aid were substantially different across the two studies, particularly for the groups that utilized aids that generated only difficult errors. Participants in the difficult errors group in Study 2 had higher estimates of trust in the aid than did their counterparts in Study 1 (the difficult miss group). This dramatic difference in trust was a likely consequence of the distribution of easy and difficult trials across the two studies. Although the aids used by the two difficult-errors-only groups in both studies were 100% accurate on easy trials, there were fewer easy trials in Study 1 ($n = 50$, consisting of only easy target trials) than in Study 2 ($n = 100$, consisting of 50 target and 50 noise trials). Therefore, the larger number of trials might have given participants in Study 2 many additional opportunities to observe the 100% reliable performance of their diagnostic aid, ultimately raising their level of trust in the aid.

GENERAL DISCUSSION

The results of these two studies support the easy-error hypothesis that the simplicity of errors made by automation is an important factor that undermines users' trust in automation. This lowered trust, in turn, adversely affects automation reliance and inflates confidence in one's own ability to perform the task unaided. This is in keeping with findings by Dzindolet et al. (2003)

that automated aids that even make 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.

On the one hand, the effect of easy errors on the human operator is likely to be exacerbated by the fact that errors that are simple are also generally more conspicuous (see Dzindolet et al., 2003) or obvious to the human than are errors that are more difficult to detect. Indeed, it is possible that the "easiness" of errors may have been confounded with conspicuity of errors, as a miss generated by the aid on an easy trial may have been more conspicuous than that on a difficult trial. However, the results of the present studies cannot be attributed solely to the conspicuity of errors committed by the automated aid as defined by Dzindolet et al. (2003). Participants in both studies observed all types of errors and received immediate feedback regarding the errors generated by the aid during testing. Indeed, participants in the difficult miss group in Study 1 were generally accurate in their estimates of the overall reliability of the aid, even though they were not provided with such information prior to testing. Therefore, it appears that the easiness of the error committed by an aid is a factor affecting trust above and beyond conspicuity (although conspicuity is necessary in order for operators to know the easiness of the error).

That the easiness of automation errors affects trust and reliance has important implications for the development of an integrative model of trust in and reliance on automated aids that is based on earlier research findings on people's perceptions of automation characteristics and human decision-making biases. Previous research suggests that decision makers often hold exaggerated notions of their own competence, particularly when the task to be performed appears easy to them (Dunning et al., 2003) as well as several biases about the "behavior" of automated aids, such as schemas of perfection and credibility (Dzindolet et al., 2001). Easy errors generated by an automated aid further inflate users' preconceived notions of their own self-competence relative to automation.

According to Dzindolet et al. (2002), perceptions of the reliability of an automated diagnostic aid are filtered through the operator's "perfect automation schema," or the expectation that automation will perform at nearly perfect rates.

This expectation leads operators to pay too much attention to errors made by an aid, thereby triggering a rapid decline in trust when diagnostic aids make obvious errors. Research on responses to warnings has shown that operators are sensitive to the predictive value of a system (Meyer & Bitan, 2002) and are less likely to respond (e.g., Bliss, Gilson, & Deaton, 1995) and assign less weight to information from a system (e.g., Maltz & Meyer, 2001; Meyer, 2001; Robinson & Sorkin, 1985) when the system is perceived as low on informative value. Therefore, it follows that an aid that generates easy errors is perceived as less informative or less capable than the human is of performing the same task unaided. This lowered perceived utility of automation combines with users' heightened notion of self-competence, ultimately bolstering the operators' degree of overconfidence in their own ability to outperform the diagnostic aid, even on difficult tasks when human diagnostic accuracy may not necessarily be higher than automation accuracy. This inflated self-confidence eventually leads to a shift in decision bias and to a reduction in operator trust in and reliance on automation.

Conclusions and Implications

The present studies highlight the simplicity of automation errors as one key factor that affects human trust in and reliance on automation. However, this factor is likely to interact with the user's preconceived notions and biases about automation, as well as with several psychological characteristics of the decision maker such as perceived self-competence, self-confidence, and decision bias, creating a complex network of variables that might affect automation trust, depending on the weights attached to each of these factors by operators in specific contexts. As is evident in the present studies, significant reductions in automation trust as a consequence of easy errors ultimately lead to inappropriate reliance on reliable automated aids that are primarily designed to assist and improve human performance in complex systems.

The results of the present study primarily imply that the avoidance of easy automation errors is a commendable goal in system design. However, easy misses of targets or easy false alarms are inevitable in the real world when the algorithms used by an automated aid to discriminate between noise and signal are insufficient to capture all instances

of a particular target. For example, there are extreme differences in the physical characteristics of certain categories of weapons, such as the differences between a sword and a folded pocket-knife or a derringer pistol and a machine gun. Although 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.

Humans tend to respond to automation in a social manner (Lee & See, 2004; Madhavan & Wiegmann, in press) and it appears from the research that trust guides the choice of automation dependence strategies when unanticipated errors by automated aids make accurate comprehension of machine behavior difficult. Under such circumstances, providing human operators with clear and specific information regarding the functional limitations of automation (i.e., informing operators about reasons a machine might err; see Dzindolet et al., 2001; Lee & See, 2004) will make automation more "trustable" and give operators a better opportunity to appropriately discriminate between the relative accuracy of automation and their own abilities. Such system-specific information (and/or training) will eventually lead to the choice of better automation utilization strategies and improved human-machine relations in the long run.

REFERENCES

- Bliss, J. P., Gilson, R., & Deaton, I. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38*, 2500-2512.
- Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Mahwah, NJ: Erlbaum.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cotte, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger drivers' reliance on collision warning systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83-87.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*, 697-718.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*, 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.

- Elvers, G. C., & Elrif, P. (1997). The effects of correlation and response bias in alerted monitor displays. *Human Factors*, 39, 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, 19–33.
- Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. New York: Wiley.
- Gupta, N., Bisantz, A. M., & Singh, T. (2001). Investigation of factors affecting driver performance using adverse condition warning systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 1699–1703). Santa Monica, CA: Human Factors and Ergonomics Society.
- Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human machine systems. *Ergonomics*, 22, 671–691.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153–184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 50–80.
- Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). The Turing effect: The nature of trust in expert system advice. In P. J. Feltovich & K. M. Ford, (Eds.), *Expertise in context: Human and machine*. (pp. 417–448). Cambridge, MA: MIT Press.
- Madhavan, P., & Wiegmann, D. A. (in press). A review of operator trust in automated aids: Is trust in machines comparable to trust in humans? *Theoretical Issues in Ergonomics Science*.
- Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors*, 43, 217–225.
- 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, 196–204.
- Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors*, 44, 343–353.
- Milton, J. S., & Corbet, J. J. (1979). *Applied statistics with probability*. New York: Van Nostrand.
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527–539.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37, 1905–1922.
- Oesch, J. M., & Murnighan, J. K. (2003). Egocentric perceptions of relationships, competence, and trustworthiness in salary allocation choices. *Social Justice Research*, 16, 53–78.
- Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision-warning systems. *Ergonomics*, 40, 390–399.
- Parasuraman, R., & Riley, V. (1997). Humans and automation use: Use, misuse, disuse, abuse. *Human Factors*, 39, 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 – Part A: Systems and Humans*, 30, 286–297.
- Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. In R. Eberts & C. G. Eberts (Eds.), *Trends in ergonomics/human factors* (Vol. 2, pp. 75–82). Amsterdam: North Holland.
- Ruble, D. N., & Stangor, C. (1986). Stalking the elusive schema: Insights from developmental and social-psychological analyses of gender schemas. *Social Cognition*, 4, 227–261.
- Sheridan, T. B. (2002). Human performance in relation to automation. In T. B. Sheridan (Ed.), *Humans and automation: System design and research issues* (pp. 69–89). Santa Monica, CA: Human Factors and Ergonomics Society.
- Sheridan, T. B., & Ferrell, 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, 550–559.
- Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.
- Thomas, L. C., Wickens, C. D., & Rantanen, E. M. (2003). Imperfect automation in aviation traffic alerts: A review of conflict detection algorithms and their implications for human factors research. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 344–348). Santa Monica, CA: Human Factors and Ergonomics Society.
- 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, 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, 352–367.

Poornima Madhavan is a postdoctoral fellow in the Dynamic Decision Making Laboratory within the Department of Social and Decision Sciences at Carnegie Mellon University. She received her Ph.D. in engineering psychology (human factors) from the University of Illinois at Urbana-Champaign in 2005.

Douglas A. Wiegmann is an associate professor of aviation human factors within the Institute of Aviation at the University of Illinois at Urbana-Champaign, where he holds appointments in the Department of Psychology and the Beckman Institute for Advanced Science and Technology. He is also currently a visiting research scientist in the Division of Cardiovascular Surgery at the Mayo Clinic College of Medicine. He received his Ph.D. in experimental psychology from Texas Christian University in 1992.

Frank C. Lacson is a graduate student in the human factors program at the University of Illinois at Urbana-Champaign, where he received his baccalaureate degree in psychology in 2002.

Date received: January 9, 2004

Date accepted: February 14, 2005