Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task

Ericka Rovira, Kathleen McGarry and Raja Parasuraman

Human Factors: The Journal of the Human Factors and Ergonomics Society 2007 49: 76
DOI: 10.1518/001872007779598082

Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task

Ericka Rovira, Kathleen McGarry, and **Raja Parasuraman,** The Catholic University of America, Washington, D.C.

Objective: Effects of four types of automation support and two levels of automation reliability were examined. The objective was to examine the differential impact of information and decision automation and to investigate the costs of automation unreliability. **Background:** Research has shown that imperfect automation can lead to differential effects of stages and levels of automation on human performance. **Method:** Eighteen participants performed a "sensor to shooter" targeting simulation of command and control. Dependent variables included accuracy and response time of target engagement decisions, secondary task performance, and subjective ratings of mental workload, trust, and self-confidence. **Results:** Compared with manual performance, reliable automation significantly reduced decision times. Unreliable automation led to greater cost in decision-making accuracy under the higher automation reliability condition for three different forms of decision automation relative to information automation. At low automation reliability, however, there was a cost in performance for both information and decision automation. **Conclusion:** The results are consistent with a model of human-automation interaction that requires evaluation of the different stages of information processing to which automation support can be applied. **Application:** If fully reliable decision automation cannot be guaranteed, designers should provide users with information automation support or other tools that allow for inspection and analysis of raw data.

INTRODUCTION

Military command and control (C^2) typically requires strategic and tactical decisions to be made in a timely manner based on multiple sources of information. A major goal for C^2 systems is to shorten the targeting cycle or tighten the *sensor to shooter* loop (i.e., execute the process more quickly than the enemy; Adams, 2001). The sensor to shooter cycle includes processes from the time a sensor spots a target until a shooter locks on it. Shortening the cycle by necessity requires that the human operators of C^2 systems be supported in some manner. Automation is one major form of support.

An important design issue for automated support tools is how much authority to assign to the automation. Factors influencing this decision include how well the automation supports the operator's situation awareness as well as the type and reliability of the automation (Barnes, 2003). Unfortunately, automated aids have not always enhanced system performance, primarily because of problems in their use by human operators or unanticipated interactions with other subsystems. Problems in human-automation interaction have included unbalanced workload, reduced system awareness, decision biases, mistrust, overreliance, and complacency (Parasuraman & Riley, 1997). There is thus a need for designing automation that supports military operators in command and control in ways that avoid such negative influences.

Parasuraman, Sheridan, and Wickens (2000) proposed a taxonomy that can guide automation design. They identified four stages of human information processing that may be supported by automation: information acquisition (Stage 1), information analysis (Stage 2), decision and action

Address correspondence to Ericka Rovira, 267C Thayer Hall/Engineering Psychology, Department of Behavioral Sciences and Leadership, U. S. Military Academy, MADN-BS&L/646 Swift Rd., West Point, NY 10996; Ericka.Rovira@usma.edu.
HUMAN FACTORS, Vol. 49, No. 1, February 2007, pp. 76–87.

selection (Stage 3), and action implementation (Stage 4). Each of these stages may be supported by automation to varying degrees, between the extremes of manual performance and full automation (Sheridan & Verplank, 1978). Because they deal with distinct aspects of information processing, the first two (information acquisition and analysis) and the last two stages (decision selection and action implementation) are sometimes grouped together and referred to as *information* and *decision* automation, respectively (see also Billings, 1997; Lee & Sanquist, 1996). Information automation can assist the user in inference and recommend possible courses of action (Jones et al., 2000). Decision automation involves selecting from various decision options and is generally based on an inference of the state of the world (information automation); however, different outcomes may be given certain values, hence making automation of this stage distinct from information automation (Wickens, 2000).

The sensor to shooter loop can be directly mapped to the Parasuraman et al. (2000) model: Target information is first acquired from sensors, subsequently analyzed further before a course of action is decided on, and finally acted upon. One example of a fielded C^2 system in which automation is applied to different stages and at different levels is the Theater High Altitude Area Defense (THAAD) system. THAAD, a system used to intercept ballistic missiles (Department of the Army, 2003), has relatively high levels of information acquisition, information analysis, and decision selection; however, action implementation automation is low, giving the human control over the execution of a specific action.

Several studies have examined the differential effects of stages and levels of automation on human performance (Crocoll & Coury, 1990; Endsley & Kaber, 1999; Galster, Bolia, & Parasuraman, 2002; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002; Sarter & Schroeder, 2001; Wickens & Xu, 2002). Crocoll and Coury (1990) examined three versions of an automated decision aid in an air defense targeting (identification and engagement) task. The automation provided (a) status information about a target (information automation), (b) recommendation concerning its identification (decision automation), or (c) both (information + decision automation). Crocoll and Coury found that with an imperfect aid, participants were better able to recover when provided with only status information. Sarter and Schroeder (2001) conducted a similar study involving automated decision aiding in a flight simulation environment. The decision aid provided status aiding (information automation) or command aiding (decision automation) regarding in-flight icing. The results confirmed the findings of Crocoll and Coury (1990) that imperfect automation can lead to worse performance when it provides decision support than when it gives only information support.

The present study compared information and decision automation but differed from previous studies in the following. First, we investigated automated aiding in a command and control environment to examine the replicability and generality of the previous results to a different simulated task domain. Second, the level of automation reliability was fixed in the previous studies. We varied automation reliability over two levels, which allowed an assessment of the range of automation reliability effects on decision making as well as a comparison with previous work on automation complacency in which automation reliability was varied and operator reliance on automated aiding was evaluated (May, Molloy, & Parasuraman, 1993; Parasuraman, Molloy, & Singh, 1993). Our hypothesis for examining this factor was guided by one of the paradoxes of automation: that when automation is imperfect, greater imperfection may lead to a lowering of the cost of unreliability. We therefore predicted that the differential cost of unreliable decision automation would be greater for high than for low automation reliability. Third, we developed three different kinds of decision automation to examine the generality of the phenomenon of the unreliability of decision automation.

Finally, in the present study we examined the *first automation failure effect,* which refers to the influence on user performance of the initial failure of previously perfect automation. The assumption is that although operators may be complacent as a result of exposure to apparently perfect automation and not respond appropriately to the first occurrence of a failure, further experience with imperfect automation may lead to appropriate calibration of trust and, subsequently, improved performance. In an examination of this effect, Merlo, Wickens, and Yeh (2000) measured target detection times in a cued search task and introduced several cuing errors in the final experimental block. Participants improved their detection of the uncued high-priority target after experiencing the first

cuing error (accuracy after was 91% vs. 50% before), suggesting that participants used a different strategy after some exposure to imperfect automation, thus improving their performance on later trials. Consequently, one additional objective of our study was to examine first automation failure effects with information and decision automation. We hypothesized that the first time automation failed operators would perform poorly but that after some experience they would be able to calibrate their trust accordingly, thus reducing complacency and improving performance.

METHODS

Participants

Eighteen undergraduate students (9 men and 9 women) aged 18 to 22 years (M = 19.83, SD = 1.25) from The Catholic University of America volunteered and were paid $15/hr for their participation.

Apparatus: The Sensor to Shooter Task Environment

A PC-based low-fidelity software simulation of a sensor to shooter targeting system (STS) was developed. The program was written in Java code for a Pentium 2 processor with a 17-inch (43-cm) monitor. A mouse was used as the input device.

The STS simulation consisted of three components shown in separate windows: a terrain view, a task window, and a communications module. The right portion of the screen was dedicated to a two-dimensional terrain view of a simulated battlefield. The window showed three red enemy units (labeled E1, E2, and E3), three yellow friendly battalion units (B1, B2, and B3), six green friendly artillery units (A1, A2, A3, A4, A5, and A6), and one blue friendly headquarter unit (HQ). A second window, the task window, was where the user made enemy-friendly engagement selections. The participants were required to identify the most dangerous enemy target and to select a corresponding friendly unit to engage in combat with the target, this was known as an enemy-friendly engagement selection. Figure 1 shows a static view of the STS simulation.

The bottom left portion of the task window was allocated to the automation support. (a) *Information automation* provided the participant with a complete list of all possible engagement combinations, including the distances between enemy targets, friendly units, and headquarters. Because no explicit pointer to decision selection was provided, this corresponds to information automation

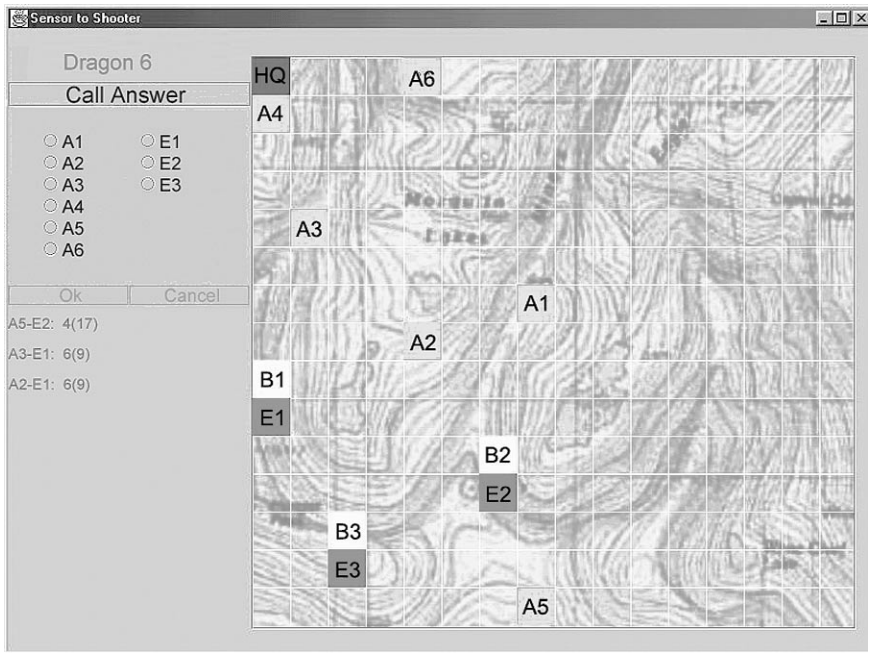

Figure 1. Sensor to shooter display with medium decision automation.

(Stage 2) in the Parasuraman et al. (2000) taxonomy. The remaining types represented different varieties of decision support: (b) *Low decision automation* gave a complete list of all possible engagement combinations, including the distances between enemy targets, friendly units, and headquarters; however, in this instance the listings were prioritized with the best selection first and the worst choice last, thereby making this a form of decision automation (Stage 3). (c) In the *medium decision automation* condition, the participant was provided the top three options for engagement, including the distances between enemy targets, friendly units, and headquarters. (d) The *high decision automation* condition recommended the best enemy-friendly engagement. Although distance information between enemy targets, friendly units, and headquarters was computed by the automation algorithm, it was not readily accessible to the operator. Thus the highest form of decision automation removed some of the "raw" data.

The communications task window, a secondary task, was located in the top left portion of the screen. A call sign appeared every 5 s and remained displayed until the next call sign. A call for communications was randomly distributed once over a 50-s period.

Task Procedures and Design

An enemy-friendly engagement selection was required within 10 s. Additionally, participants were required to click on the call answer button every time their call sign appeared in the communications window.

A 4 (type of automation) × 2 (overall automation reliability) × 2 (trial reliability) within-subjects design was used. The four automation support conditions included information automation and three different forms of decision automation: low, medium, and high. Overall automation reliability was varied across two values (60% and 80%) by manipulating the proportion of trials in which a correct assessment or decision was provided. In the 60% overall automation reliability condition, 24 trials were reliable (correct automated assessment) and 16 trials were unreliable (incorrect automated assessment). Similarly, in the 80% overall automation reliability condition, 32 trials were reliable and 8 trials were unreliable. Reliability was manipulated for each of the four automation support conditions. *Trial reliability* referred to a correct automated assessment (reliable) versus an incorrect automated assessment (unreliable). Participants were informed that although the automation was highly reliable, it was not 100% reliable all of the time. However, no further information on the actual reliability was given.

Each participant first performed the task manually (20 trials) with no automation support. Each participant experienced 40 trials of each of the eight conditions: information automation at (a) 60% and (b) 80% reliability; low decision automation at (c) 60% and (d) 80% reliability; medium decision automation at (e) 60% and (f) 80% reliability; and high decision automation at (g) 60% and (h) 80% reliability. The order of the four automation support tools was counterbalanced between participants. Participants completed both overall automation reliability conditions (60% and 80%) for each automation support tool before a new automation support tool was introduced. Overall automation reliability was counterbalanced within each type of automation. Prior to the start of each automation condition, participants completed 20 trials of 100% reliability automation; these trials were given to allow operators to recover trust in the automation but were not part of the data analysis. In all, each participant completed 500 trials: 20 manual trials, 320 experimental automation trials, and 160 "recovery" (100% reliability) trials.

Dependent variables included the accuracy and speed of enemy-friendly engagement selections. Accuracy was calculated by the percentage of trials in which the participant correctly selected the most dangerous enemy target and a corresponding friendly unit to engage in combat. Secondary task measures of performance included accuracy and response time to the communications (call sign) task. To obtain subjective measures of mental workload, participants were asked to fill out the NASA-Task Load Index (Hart & Staveland, 1988). Subjective ratings of trust in automation and operator self-confidence were obtained on a Likert rating scale ranging from 0 to 100, built after scales used by Lee and Moray (1994). All subjective measures were administered after each automation condition.

RESULTS

Manual Control Versus Automation Support

Decision-making accuracy. The mean rate of correct enemy-friendly engagement selections

(decision-making accuracy) was computed for each participant under both manual control and with automation support. Under imperfect automation (60% and 80% reliability), engagement selection rates were computed separately for reliable and unreliable trials within each condition. A paired samples *t* test of engagement selection rates showed that there was no difference in decision accuracy between manual ($M = 89.4\%$) and reliable automation ($M = 88.4\%$), $t(17) = 0.62$, $p = .541$. However, there was a significant difference in decision accuracy between the manual and the unreliable automation support conditions, $t(17) = 6.9$, $p < .001$, with accuracy declining to 70% under unreliable automation. In general, there was no difference in accuracy performance between manual and reliable automation, but accuracy declined under unreliable automation.

Response times. Enemy-friendly engagement response times (RTs) were reduced significantly from the manual condition when participants were given reliable automation support, $t(17) = 5.3$, $p < .001$. Under the highest form of decision automation, participants had only to click "OK" if they agreed with the automation, and this may have led to a difference in RTs simply because of the design of the interface and the requirement to make a single motor response, rather than to the functionality of the automation. Therefore, we conducted a paired samples *t* test of engagement selections between manual and reliable automation, excluding the trials with the highest form of decision automation. This revealed that RTs were nonetheless reduced by the reliable automation, $t(17) = 2.37$, $p = .030$. Although RTs increased under unreliable automation, they were not significantly different from the manual condition, $t(17) = –0.49$, $p = .633$.

Effects of Automation Support and Reliability on Decision-Making Accuracy

The mean rate of correct enemy-friendly engagement selections (decision accuracy) was computed for each participant. A 4 (type of automation: information automation and low, medium, or high decision automation) × 2 (overall automation reliability: 60% or 80%) × 2 (trial reliability: reliable or unreliable) repeated measures ANOVA was used to examine decision accuracy. The main effects of type of automation, $F(3, 51) = 2.13$, $p = .108$, and of overall automation reliability were not significant, $F(1, 17) = 1.32$, $p = .267$, but that of trial reliability was, $F(1, 17) = 34.54$, $p < .001$. Mean accuracy rates for reliable and unreliable trials were 88.5% and 70.0%, respectively. The interaction between type of automation and overall automation reliability was significant, $F(3, 51) = 11.18$, $p < .001$. Neither the interaction between automation type and trial reliability, $F(3, 51) = 1.34$, $p = .273$, nor that between overall automation reliability and trial reliability was significant, $F(1, 17) = 0.46$, $p = .508$. However, there was a significant three-way interaction between type of automation, overall automation reliability, and trial reliability, $F(3, 51) = 13.962$, $p < .001$. To examine these effects more closely, we conducted individual comparisons between levels of trial reliability (reliable vs. unreliable) across all types of automation support (information automation and low, medium, and high decision automation) and levels of overall automation reliability (60% and 80%).

For the 80% overall automation reliability conditions, the effect of trial reliability was not significant for information automation, $t(17) = 0.141$, $p = .89$, but was significant for each of the three decision automation conditions. Accuracy was higher on reliable trials than on unreliable trials for low ($M = 88.62\%$ vs. $M = 66.4\%$), $t(17) = –2.78$, $p = .013$, medium ($M = 86.14\%$ vs. $M = 59.17\%$), $t(17) = –3.47$, $p = .003$, and high decision automation ($M = 95.72\%$ vs. 64.67%), $t(17) = –4.01$, $p = .001$. As the top panel of Figure 2 shows, trial unreliability did not degrade performance in the information automation condition but did for all three decision automation conditions. As shown in the bottom panel, however, for the 60% overall automation reliability condition there was a cost of trial unreliability for information automation ($M = 83.51\%$ reliable, 53.06% unreliable), $t(17) = –7.16$, $p < .01$, and two of the three decision automation conditions: medium decision automation ($M = 91.44\%$ reliable, 72.97% unreliable), $t(17) = –3.15$, $p = .006$, and high decision automation ($M = 89.64\%$ reliable, 73.39% unreliable), $t(17) = –2.62$, $p = .018$. The effect of trial reliability was not significant for low decision automation, $t(17) = –0.97$, $p = .347$.

Effects of Automation Support and Reliability on Response Time

Enemy-friendly engagement response times were examined using a 4 (type of automation:

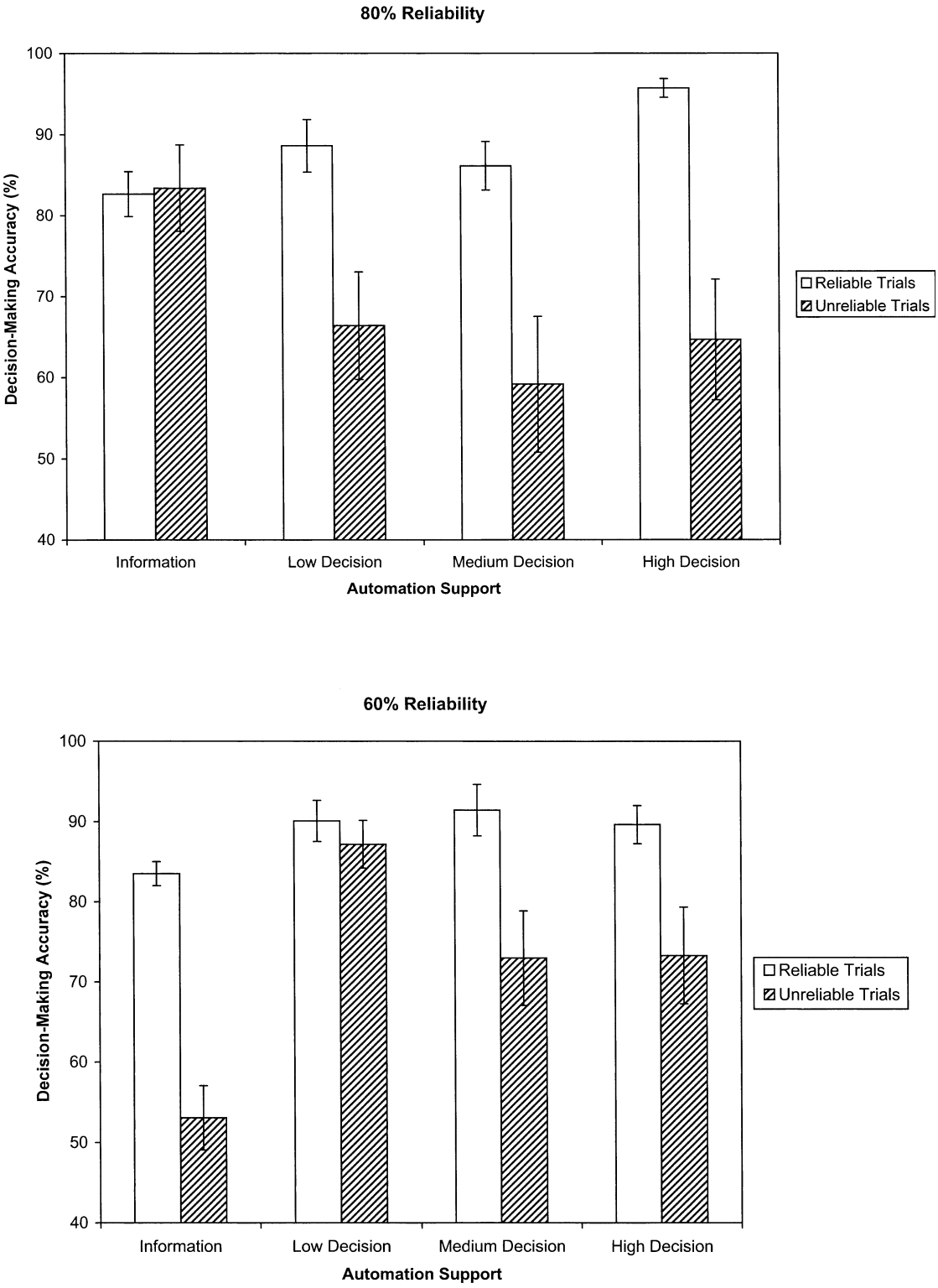

Figure 2. Decision-making accuracy for reliable and unreliable trials as a function of type of automation support with (top panel) 80% and (bottom panel) 60% overall automation reliability.

information automation and low, medium, or high decision automation) × 2 (overall level of automation reliability: 60% or 80%) × 2 (type of trial reliability: reliable or unreliable) repeated measures ANOVA. The main effect of type of automation was significant, $F(3, 51) = 8.61$, $p < .001$. The mean enemy-friendly engagement RTs for information automation and for low, medium, and high decision automation were 6037, 5795, 6172, and 5210 ms, respectively. The main effect of overall automation reliability was marginally significant, $F(1, 17) = 3.33$, $p = .086$, the mean enemy-friendly engagement response times for 60% and 80% reliability being 5904 and 5703 ms, respectively. The main effect of trial reliability was significant, $F(1, 17) = 43.65$, $p < .001$, with mean enemy-friendly engagement response times for reliable and unreliable trials being 5382 and 6225 ms, respectively. The interaction between automation type and overall automation reliability was significant, $F(3, 51) = 9.38$, $p < .001$. The interaction between automation type and trial reliability, $F(3, 51) = 17.52$, $p < .001$, and the three-way interaction among automation type, overall automation reliability, and trial reliability, $F(3, 51) = 2.99$, $p = .04$, were significant.

For the 80% overall automation conditions, the effect of trial reliability was not significant for the information automation condition, $t(17) = –1.52$, $p = .147$, but was significant for two of the three decision automation conditions. The effect of trial reliability was significant for low decision automation, $t(17) = 3.18$, $p = .005$, with shorter response times on reliable trials ($M = 5881$ ms) than on unreliable trials ($M = 6719$ ms). The effect of trial reliability was also significant for high decision automation ($M = 3689$ ms reliable, $M = 6059$ ms unreliable), $t(17) = 6.13$, $p < .001$, but only marginally significant for medium decision automation ($M = 6045$ ms reliable, $M = 6691$ ms unreliable), $t(17) = 1.85$, $p = .081$. For the 60% overall automation reliability condition, however, there was a cost of trial unreliability for information automation ($M = 5794$ ms reliable, $M = 6203$ ms unreliable), $t(17) = 3.11$, $p = .006$, and two of the three decision automation conditions: medium decision automation ($M = 5515$ ms reliable, $M = 6437$ ms unreliable), $t(17) = 5.03$, $p < .001$, and high decision automation ($M = 4450$ ms reliable, $M = 6540$ ms unreliable), $t(17) = 6.34$, $p < .001$. The effect of trial reliability was not significant for low decision automation, $t(17) = 0.54$, $p = .596$.

First Failure Analysis

The mean decision-making accuracy across participants was first calculated for each automation failure in the medium and high decision automation conditions. We then plotted the data to assess each participant's engagement decision following the first failure in each condition. We found that only 2 participants made correct decision engagement responses to all first failures in each automated system. Further, we found that regardless of whether participants made correct or incorrect decision responses to the initial first failure in the experiment, they still made incorrect decision responses to first failures in subsequent automated conditions.

Using the binomial distribution, we determined estimates of the mean, π, and standard deviation, $\sigma = \sqrt{[\pi(1 - \pi/N)]}$, of the probability of making a correct decision following the first automation failure (Ott, 1993, pp. 366–367), in which π = the number of participants making a correct decision/total number of participants, N, on that automation failure trial. Next, we used π and σ to compute the 90% confidence intervals for correct response to the first failure and then examined which of the correct response rates to subsequent failures lay outside the interval. If the correct response rate for a particular failure number was above the upper interval value, one can conclude that decision accuracy improved as compared with the first failure. For this analysis, we collapsed across automation reliability conditions (60% and 80%). Failure Numbers 4, 5, and 8 showed significant improvement over the first failure correct response rate with high decision automation support, whereas only Failure Number 7 did with medium decision automation support. There was no consistent first failure effect with either low decision automation or information automation support. There appeared to be first failure effects for two out of our three decision automation support tools. Figure 3 shows clear trends for the first failure effect for both high decision and medium decision automation.

Secondary Task Performance (Communications)

Accuracy. A 4 (type of automation: information automation and low, medium, or high decision automation) × 2 (overall automation reliability: 60%

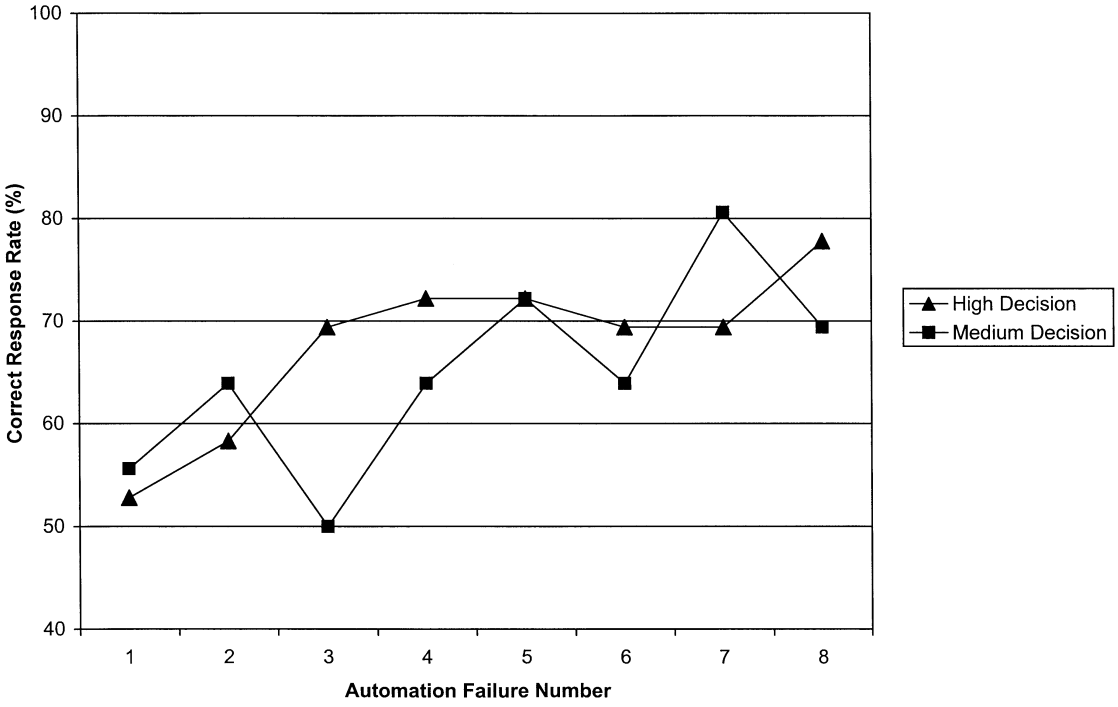

Figure 3. The proportion of participants who made a correct decision following an automation failure, averaging across automation reliability conditions (60% and 80%); medium and high decision automation conditions are shown.

or 80%) repeated measures ANOVA was used to examine communication task accuracy rates. The only significant effect was the interaction between automation type and overall automation reliability, $F(3, 51) = 6.07$, $p < .001$. The effect of overall automation reliability for information automation, $t(17) = 0.14$, $p = .888$, and low decision automation was not significant, $t(17) = 0.68$, $p = .506$, but was significant for medium decision automation, $t(17) = –4.99$, $p = .001$. Mean accuracy rates indicated better performance under low ($M = 89.7\%$) than under high ($M = 83.6\%$) overall automation reliability. The effect of overall automation reliability in the high decision condition was also significant, $t(17) = 2.60$, $p = .019$, and showed accuracy to be better under high ($M = 89.1\%$) than under low ($M = 83.9\%$) overall automation reliability.

Response times. For RTs in the communications task, a 4 (type of automation: information automation and low, medium, or high decision automation) × 2 (overall automation reliability: 60% or 80%) repeated measures ANOVA was used. The main effect of type of automation, $F(3, 51) = 7.87$, $p < .001$, was significant, with high decision automation leading to lower RTs as compared with the other three conditions. No other effects were significant.

Subjective Ratings of Mental Workload

Using a 4 (type of automation: information automation and low, medium, or high decision automation) × 2 (overall automation reliability: 60% or 80%) repeated measures ANOVA there was a significant effect of type of automation, $F(3, 51) = 4.66$, $p = .006$. Subjective ratings of workload were highest for information automation ($M = 38$), followed by medium decision ($M = 36$), high decision ($M = 34$), and low decision automation ($M = 29$). The effect of overall automation reliability on subjective workload ratings was also significant, $F(1, 17) = 7.89$, $p < .012$, with higher ratings for the 80% overall automation reliability condition ($M = 35$) than for the 60% overall automation reliability condition ($M = 33$).

Subjective Ratings of Trust and Self-Confidence

A 4 (type of automation: information automation and low, medium, or high decision automation) × 2 (overall automation reliability: 60% or 80%) repeated measures ANOVA was used to

examine subjective ratings of trust. There was a significant effect of type of automation on ratings of trust, $F(3, 51) = 3.39$, $p = .025$. Subjective ratings of trust were highest for low decision automation ($M = 53$), followed by high decision ($M = 48$), medium decision ($M = 41$), and information automation ($M = 34$). Unexpectedly, overall automation reliability had no significant effect on trust, $F(1, 17) = 0.14$, $p = .710$. The interaction between type of automation and overall automation reliability was also not significant, $F(3, 51) = 1.51$, $p = .222$. Finally, there were no significant effects of any factor on subjective ratings of self-confidence.

DISCUSSION

The purpose of this study was to examine the differential impact of information and decision automation on human performance and to investigate the costs of automation reliability in a simulated command and control task. Three different types of decision automation were examined. When automation was reliable it enhanced performance. In particular, decision automation significantly reduced the time for target engagement decisions. This is important for reducing the overall sensor to shooter time in tactical C^2 operations. However, the automation was imperfect, and when it provided an incorrect assessment the accuracy of target engagement decisions declined. When overall automation reliability was 80%, this cost of automation reliability was greater for the three decision automation support tools than for the information automation condition. The accuracy of engagement decisions was significantly reduced and the decision time increased, confirming the previous studies of Crocoll and Coury (1990) and Sarter and Schroeder (2001). The results demonstrate the replicability of the earlier findings in a simulated command and control task and for different forms of decision automation.

The differential cost of automation unreliability for the three forms of decision automation, as compared with information automation, confirmed our hypothesis regarding the effects of automation unreliability. However, these effects were found only for 80% overall automation reliability. At 60% overall automation reliability, both information and decision automation reduced performance on unreliable trials. This finding suggests that below a certain threshold level of reliability, automation imperfection leads to poorer performance irrespective of the type of automation. In a recent review, Wickens and Dixon (2005) suggested that human performance is sensitive to the level of automation imperfection. Our data support this view, as we found worse performance for information automation when the overall automation reliability was 60% relative to 80%.

These findings may be interpreted via the automation model proposed by Parasuraman et al. (2000). Information automation can give the operator status information, can integrate different sources of data, and may recommend possible courses of action (Jones et al., 2000). However, this form of automation typically does not give values to the possible courses of action, whereas decision automation does (Wickens, 2000). Therefore, it is possible that when the automation is highly reliable yet imperfect, performance is better with an information support tool because the user continues to generate the values for the different courses of action and, hence, is not as detrimentally influenced by inaccurate information. Additionally, a user of decision automation may no longer create or explore novel alternatives apart from those provided by the automation, thus leading to a greater performance cost when the automation is unreliable.

The use of two levels of overall automation reliability allowed for a comparison with previous work on automation complacency in which automation reliability was varied (May et al., 1993; Parasuraman et al., 1993). The results showed that there was a greater cost of unreliable decision automation with high (80%) than with low (60%) overall automation reliability. This finding indicates that participants rely to a greater extent on automation when it is more reliable, even though it is imperfect. This is also consistent with a study by May et al. (1993), who concluded that the more reliable the automation, the more operators had reason to believe the automation to be dependable. As a result, when the automation was performing reliably, complacency increased, leading to poorer operator detection rates when it failed. The results of the present study parallel this finding, and both are examples of one of the paradoxes of automation. The more reliable the automation, the greater its detrimental effect when it fails.

One problem with this interpretation, however, is that the subjective ratings of participant trust in automation did not provide supporting evidence. While trust ratings were greater for decision automation than for information automation, somewhat

surprisingly, they did not vary with overall automation reliability as expected. Wiegmann, Rich, and Zhang (2001) also observed a disassociation between subjective ratings of automation reliability and objective performance. It was suggested that the two measures might not consistently reflect the same underlying construct of automation trust (Wiegmann et al., 2001). The relationship between subjective ratings of trust and automation usage is complex and depends on a range of environmental and individual factors (Dzindolet, Pierce, Beck, & Dawe, 2001; see review by Lee & See, 2004). Further research on the relationship between subjective trust measures and actual automation usage would be helpful in understanding why trust ratings and automation reliance sometimes do and at other times do not dissociate.

An additional feature of the present study is that the first automation failure effect was examined and tested statistically. Based on previous research (Merlo et al., 2001; Stanton, Young, & McCaulder, 1997), it was hypothesized that participants would perform poorly the first time the automation failed but that after some experience with imperfect automation, participants would be able to calibrate their trust accordingly, thus reducing complacency and improving performance. The results showed that performance changed after exposure to unreliable automation. In general, the reduction in decision accuracy with unreliable automation was less marked with each subsequent failure; this change was significant for both medium and high decision automation. However, we did not obtain statistically significant evidence for increased correct response rate for all failures subsequent to the first failure in all conditions. The reason for this could simply be the low power associated with analyzing data from single trials.

These results partially support the findings of Wickens and Xu (2002). Prior literature on the first failure effect is somewhat contradictory. Some studies have found little or no evidence of such effects (Davison & Wickens, 1999; Wickens, Gempler, & Morphew, 2000). One reason may be the use of displays that integrate raw data, thus supporting the operator better and reducing complacency effects (Molloy & Parasuraman, 1994; Wickens et al., 2000). Another reason for not finding first automation failure effects, as proposed by Wickens and Xu (2002), pertains to the instructions that participants receive. Prior instructions that the automation may be imperfect, as well as experience with unreliable automation in practice trials, may minimize the human performance consequences of automation failure (Wickens, Helleberg, & Xu, 2002). Participants in this study did not have experience with the unreliable trials during practice blocks, but they were specifically instructed that the automation was highly reliable but not 100% reliable, and yet first automation failure effects were still found.

The greater detrimental impact on user performance of imperfect yet highly reliable decision automation, as compared with information automation, may be particularly apparent in high-workload, time-stressed work environments, as in the present simulation study, and in conditions of high-tempo engagement in the battlefield. The target engagement task was relatively complex, had to be time shared with a communications task, and had to be completed within 10 s. Such conditions may have encouraged reliance on the decision choice suggested by the automation, particularly if there was insufficient time for the user to check the information sources to verify the automated advice. When the conditions permit such verification, however, the costs of imperfection in decision automation may be reduced. Lorenz et al. (2002) found that a high level of decision automation did not lead to poorer decision-making performance (as compared with a lower or moderate level of automation) when an automated fault management system failed. They attributed this lack of a cost of imperfection in high-level decision automation to the ability of the users to query the system, inspect the raw information sources, and verify or negate the automated advice.

Limitations

The sensor to shooter microworld used in this study may not be fully representative of command and control automated environments. Participants did not have to account for real-world stressors such as fatigue, environmental factors, and fatal implications. Furthermore, the cognitive demands involved in the communications task may not have been sufficiently challenging. In the simulated environment participants simply had to click on a call answer button, whereas in military environments communication involves more detail; for instance, operators may need to tune a

radio to a particular frequency. The limited complexity of the sensor to shooter task may explain the lack in accuracy differences between manual and reliable automation conditions. Further, participants performed better with less reliable automation, which may be a result of rejecting the automation entirely and opting for manual control. It may be difficult for operators of future automated military systems to discard automation because of the vast amount of information that would need to be gathered and processed.

Another possible limitation of the study was the decision to include perfect trials between the two varying levels of overall automation reliability. We deliberately made this choice in our experimental design despite the danger that presenting 100% reliable automation followed by either 60% or 80% reliability could have led to a possible "averaging" effect of automation reliabilities – for example, (60 + 100)/2 versus (80 + 100)/2 = 80% versus 90% – such that any difference between the two unreliable levels would be potentially obscured. However, we felt it was more important to allow operators to recover in an effort to encourage automation usage and, hence, preceded each unreliable automation block with perfectly reliable automation. Additionally, it could be that simply providing participants with a block of reliable trials may not have been sufficient to rebuild trust in the automation. If this were the case, we would have expected to find differences between responses to the first automation failure in the experiment and responses to first failures occurring in subsequent conditions. We found that regardless of whether the participants correctly or incorrectly responded to the initial first failure in the experiment, they still incorrectly respond to first failures in subsequent automated conditions.

Practical Implications

The current study has potentially important implications for automation use and the design of real-world decision support tools. Human automation teams may break down when automation is unreliable because an operator may no longer experience benefits from the automation. Thresholds for automation trust need to be better understood, specifically regarding the differential impact of information and decision automation. The present research, in conjunction with previous studies (Crocoll & Coury 1990; Sarter & Schroeder, 2001), suggests that there may be a larger gap between the acceptability of information and decision automation when the system is not 100% reliable. The problem is the trend in automation tools that support the operator by completing the integration process and providing a solution with much of the raw data removed. This is problematic because it removes the operator's understanding of the information and moves the operator further away from the decision-making process. As a result, when the automation is unreliable, system performance degrades particularly when decision automation is provided. Therefore, if reliable decision automation cannot be guaranteed, it is recommended that information automation be provided to support the user, or at least a low level of decision automation versus a highly autonomous decision support tool. Future research in this area may allow for more sophisticated support tools to be used in dynamic environments while minimizing performance costs and improving safety.

ACKNOWLEDGMENTS

This work was supported by Grant DAAD17-00-p-0366 from the Army Research Laboratory, Ft. Huachuca, AZ; by Grant GSRP ORG 364230 from NASA Langley Research Center; and by the Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Army Research Office under Grant DAAD19-01-1-0621. The views expressed in this work are those of the authors and do not necessarily reflect official U.S. Army policy. Many thanks to Ulla Metzger, Bernd Lorenz, Marla Zinni, and Peter Squire. Lastly, thanks to Xiong Jiang for programming the sensor to shooter simulation.

REFERENCES

Adams, T. K. (2001). Future warfare and the decline of human decision-making. *Parameters* (Winter 2001–2002), 57–71.

Barnes, M. J. (2003). *The human dimension of battlespace visualization: Research and design issues* (Tech. Rep. ARL-TR-2885). Aberdeen Proving Ground, MD: Army Research Lab.

Billings, C. (1997). *Aviation automation: The search for a human-centered approach.* Hillsdale, NJ: Erlbaum.

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1524–1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Davison, H., & Wickens, C. D. (1999). *Rotorcraft hazard cueing: The effects on attention and trust* (Tech. Rep. ARL-99-5/NASA-99-1). Savoy: University of Illinois, Aviation Research Lab.

Department of the Army. (2003). *THAAD theatre high altitude area defense missile system, USA.* Retrieved November 17, 2003, from http://www.army-technology.com/projects/thaad/

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2001). *A framework of automation use* (Tech. Rep. ARL-CR-2412). Aberdeen Proving Ground, MD: Army Research Lab.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42,* 462–292.

Galster, S. M., Bolia, R. S., & Parasuraman, R. (2002). Effects of information and decision-aiding cueing on action implementation in a visual search task. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 438–442). Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier Science.

Jones, P. M., Wilkins, D. C., Bargar, R., Sniezek, J., Asaro, P., Danner, N., et al. (2000). CoRAVEN: Knowledge-based support for intelligent analysis. In *Federated Laboratory 4th Annual Symposium: Advanced Displays and Interactive Displays Consortium* (pp. 89–94). University Park, MD: U.S. Army Research Lab.

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., & Sanquist, T. F. (1996). Maritime automation. In R. Parasuraman & M. Mouloua (Eds.), *Human performance in automated systems: Recent research and trends* (pp. 365–384). Hillsdale, NJ: Erlbaum.

Lee, J., & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. *Human Factors, 46,* 50–80.

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault management in a simulated space flight microworld. *Aviation, Space, and Environmental Medicine, 73,* 886–897.

May, P. A., Molloy, R., & Parasuraman, R. (1993, October). *Effects of automation reliability and failure rate on monitoring performance in a multi-task environment.* Paper presented at the Human Factors and Ergonomics Society 37th Annual Meeting, Seattle, WA.

Merlo, J. L., Wickens, C., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. In *Proceedings of the 4th Annual Army Federated Laboratory Symposium* (pp. 27–31). College Park, MD: Army Research Laboratory.

Molloy, R., & Parasuraman, R. (1994). Automation-induced monitoring inefficiency: The role of display integration and redundant color coding. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Recent research and trends* (pp. 224–228). Hillsdale, NJ: Erlbaum.

Ott, R. L. (1993). *An introduction to statistical methods and data analysis.* Belmont, CA: Wadsworth.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology, 3,* 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: 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.

Sarter, N., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors, 43,* 573–583.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators.* Cambridge: Massachusetts Institute of Technology, Man Machine Systems Laboratory.

Stanton, N. A., Young, M., & McCaulder, B. (1997). Drive-by-wire: The case of mental workload and the ability of the driver to reclaim control. *Safety Science, 27,* 149–159.

Wickens, C. D. (2000). *Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness* (ARL-00-10/NASA-00-2). Moffett Field, CA: NASA Ames Research Center.

Wickens, C. D., & Dixon, S. R. (2005). *Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature* (Tech. Rep. AHFD-05-01/MAAD-05-1). Savoy: University of Illinois, Aviation Research Lab.

Wickens, C. D., Gempler, K., & Morphew, M. E. (2000). Workload and reliability of traffic displays in aircraft traffic avoidance. *Transportation Human Factors Journal, 2,* 99–126.

Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver, choice, and workload in free flight. *Human Factors, 44,* 171–188.

Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention* (Tech. Rep. AHFD-02-14/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab.

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.

Ericka Rovira is an assistant professor in the Engineering Psychology Program of the Behavioral Sciences and Leadership Department at the United States Military Academy, West Point, NY. She received her Ph.D. in applied experimental psychology from The Catholic University of America in 2006.

Kathleen McGarry is a Ph.D. candidate in the Applied Experimental Psychology Program at The Catholic University of America and a research affiliate at the Arch Lab, George Mason University. She received her M.A. in psychology from The Catholic University of America in 2004.

Raja Parasuraman is a professor of psychology at the Arch Lab, George Mason University. He received his Ph.D. in psychology from Aston University, U.K., in 1976.

Date received: February 17, 2004
Date accepted: October 24, 2005