THE EFFECTS OF AUTOMATION EXPERTISE AND SYSTEM CONFIDENCE ON TRUST BEHAVIORS

Randall D. Spain
James P. Bliss
Old Dominion University
Department of Psychology
Norfolk, Virginia

Trust in automation is more likely to be appropriate when information about the automation's capability is available. The goal of this study was to determine how automation expertise and system confidence affected automation trust behaviors. Forty-one participants completed a target detection task while receiving advice from an imperfect diagnostic aid that varied in expertise (expert vs. novice) and confidence (75% vs. 50% vs. 25%, no aid). Results showed that participants were more willing to comply with the highly confident expert aid than the highly confident novice aid. Furthermore, participants were more apt to generate false alarms as system confidence increased. These results suggest that, similar to interpersonal relationships, humans appraise automation features such as confidence and expertise when deciding to comply with automation. Implications and direction for future research are discussed.

INTRODUCTION

Technological advancements have allowed engineers to introduce automation into complex technical systems. Automation is technology that gathers, filters, and organizes information, makes decisions, and carries out actions that a human would otherwise execute (Parasuraman & Riley, 1997). Parasuraman, Sheridan, and Wickens (2000) have identified four stages of automation that parallel the stages of human information processing. These stages are information acquisition, diagnosis, action selection, and execution. Information acquisition automation assists operators by selecting, organizing, highlighting, and filtering information. Diagnostic automation assists operators by performing cognitive operations such as integration and assessment. Action selection and execution automation assist operators by generating decision alternatives and executing actions on behalf of the operator.

Diagnostic Automation

Automation in this category, including alarm systems and decision support systems, possesses several features that are relevant to the study of human-machine 'team' performance. First, diagnostic aids are based on imperfect algorithms and function in an uncertain world. Therefore, automation failures are likely to occur. Automation failures can take two forms: misses and false alarms. Empirical evidence suggests that automation false alarms may at times be more damaging than misses (Bliss, 2003) and that the two types of errors affect trust related behaviors differently (Rice, 2009). Second, diagnostic aids are opaque. Rarely do operators have access to the raw data the aid is diagnosing. When raw data are not available, or when the data are too difficult to interpret, the operator has only the automation's recommendation to base his or her judgments (Sorkin & Woods, 1985). In these situations, an operator's decision to rely on an automated aid will likely depend on a number of factors including his or her trust in the diagnostic aid.

Trust in Automation

Trust is an attitude that guides automation dependence (i.e., reliance and compliance). The role of trust in human-computer interaction has been the focus of much research over the past two decades (see Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Madhavan & Wiegmann, 2007; Lee & Moray, 1992). Trust largely depends on perceptions of the capability of automation (Sheridan & Parasuraman, 2006). Therefore, trust in and compliance with an automated system are more likely to be appropriate when information about the automation's capability is available.

Purpose of the Current Study

The purpose of this research was twofold. First, we sought to determine if presenting users with system confidence information would influence trust related behaviors. As McGuirl and Sarter (2006) note, system confidence may promote appropriate trust by providing information about the aid's situational decision making accuracy. Rather than blindly following an aid's advice, individuals may be able to use confidence information to guide their response decisions on case-by-case basis. However, this assumption is subject to empirical investigation, as few studies have directly examined the influence of system confidence on trust behaviors.

The second goal of this research was to determine if automation expertise moderated the effects of system confidence on trust behaviors. We were particularly interested to see if operators would weigh confidence information from expert and novice systems differently. Existing research indicates that in human-human teams, credibility and confidence influence the acceptance of advice (Sniezek & Von Swol, 2001). A similar interaction strategy could exist in human-automation teams. However, this is subject to empirical investigation.

Hypotheses: Based on the literature we expected participants' compliance rates to calibrate to the system's level of confidence (McGuirl & Sarter, 2006). We also expected an interaction between system confidence and automation expertise. Specifically, we predicted that participants would comply more often with the expert aid than the novice aid when system confidence was high. Furthermore, we predicted that participants would commit more false alarms when they interacted with the highly confident expert aid because the aid's expertise would serve as a cue to accelerate compliance (Dijkstra, 1999). These hypotheses were examined within a simulated military target detection task in which participants searched for covert enemy targets and received diagnostic support from a detection aid that varied in expertise and confidence.

METHOD

Experimental Design

A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, unaided) x 2(image quality: high, low) mixed subjects design was used for this experiment. Automation expertise served as the between subjects variable and was manipulated by providing participants with information concerning the expertise and performance capability of the diagnostic aid. The expert system, SUPER CONTRAST DETECTOR, was described as being a highly advanced target detection system with sophisticated algorithms and superior test scores. The novice system, CONTRAST DETECTOR, was described as being an obsolete target detection system with out-dated algorithms and inferior test scores.

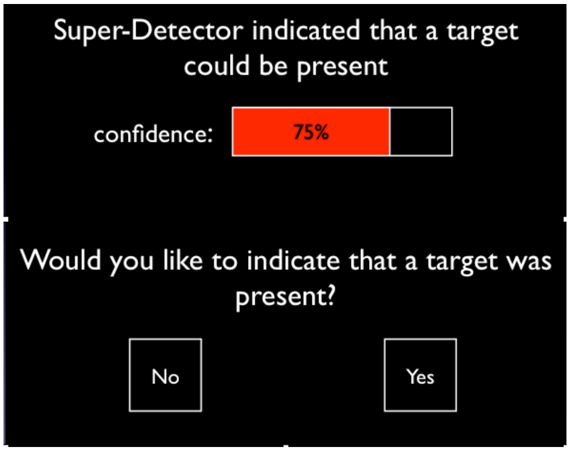
System confidence ratings were presented numerically and graphically. A 75% rating was displayed with a red bar three-fourths the size of the horizontal indicator with the rating superimposed in black font (Figure 1). A 50% rating was displayed with an orange bar, one-half the size of the horizontal indicator, with the rating superimposed in black font. A 25% rating was displayed with a yellow bar, one-fourth the size of a horizontal indicator, with the rating superimposed in black font. On trials with no aid, participants did not receive diagnostic advice from the system. Each participant received 24 high confidence, 24 neutral confidence, 24 low confidence trials, and 24 no aid trials. In these four conditions, the base rate of a target being present was .75, .50, .25, and .50, respectively.

Participants

Forty-one undergraduate students (males = 15, females = 26) from Old Dominion University participated in this study. The average participant age was *M* = 21.75 (*SD* = 5.41). Participants received 1.5 credit points for participating. These points could be used for extra credit or to meet course requirements. Participants were screened for normal or corrected-to-normal

Apparatus

A simulated military target detection scenario served as the primary task. Using Visual Basic 6.0™ software, this scenario was displayed on a 17-inch monitor connected to IBM compatible 3.20 GHz Intel Pentium D computer hosting Windows XP. The scenario required participants to view simulated synthetic aperture radar (SAR) images and detect a covert enemy target. The simulated SAR images contained 10 randomly dispersed stimuli. Approximately half of the images contained a target. The placement of the target varied across images. System confidence and image quality served as within subject variables.



Super-Detector indicated that a target could be present

confidence: 75%

Would you like to indicate that a target was present?

Figure 1: Screen shot of the simulation interface for 75% confident expert system.

Tasks and Procedure

During the scenario, participants played the role of an inspector who was responsible for searching SAR images for enemy targets. At the onset of each trial, the SAR image appeared on the screen for one second. After the image disappeared, participants received diagnostic advice from the detection aid in the form of a text message. This message included the aid's confidence (refer to Figure 1) assessment concerning the presence of an enemy target. Participants were informed that the confidence estimates were based on how well the information collected from system's detection algorithms matched the enemy template located in the system's target database. After reviewing the aid's advice, participants indicated whether they thought a target was present and their decision confidence. Then participants received feedback concerning the accuracy of their decision.

After completing the first session, participants took a short break. The second session followed the same procedures as the first session; however, participants who originally interacted with the expert aid and viewed low quality images then viewed high quality images, and vice versa. Participants completed both sessions (i.e., 192 trials total) in approximately one hour.

RESULTS

Prior to computing inferential statistical analyses, we screened the data set for missing data, unequal sample sizes, and outliers. We computed descriptive statistics for each variable to ensure that the statistical assumptions for each of the analyses were not violated. Unless otherwise noted, all analyses were computed using a critical value of $\alpha = .05$.

Compliance

A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, no aid) x 2(image quality: high, low) mixed factorial ANOVA was calculated to assess main and interaction effects of image quality, automation expertise, and system confidence on advice compliance. Because the system was false alarm prone (i.e. it did not commit any misses), compliance was defined as the number of trials during which the participant reported a target being present.

Results indicated significant main effects for image quality *F*(1, 39) = 8.71, *p* < . 01, partial ŋ^2 = .18, and system confidence *F*(3, 117) = 97.53, *p* < .001, partial ŋ^2 = .71, and a significant interaction between image quality, automation expertise, and system confidence *F*(3, 117) = 3.46, *p* < .05, partial ŋ^2 = .08. Compliance was always higher in the low image quality condition (*M* = 12.75, *SE* = .43) than the high image quality condition (*M* = 14.17, *SE* = .50). Therefore, we analyzed the data within levels of image quality.

High image quality. A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, no aid) mixed ANOVA indicated that the interaction of automation expertise and system confidence and the main effect of automation expertise on compliance failed to reach significance (*p* > .05). However, there was a significant main effect for system confidence, *F*(3, 117) = 90.85, *p* < .001, partial ŋ^2 = .70. Post hoc analysis indicated that participants were more likely to indicate that a target was present when the aid was 75% (*M* = 18.39, *SD* = 3.80) confident than when it was 50% (*M* = 12.51, *SD* = 3.93) or 25% (*M* = 8.20, *SD* = 3.63) confident. The difference in compliance between 50% and 25% confidence was also statistically significant.

Low image quality. A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, no aid) mixed ANOVA indicated a significant main effect of system confidence on compliance *F*(3, 117) = 54.12, *p* < .001, partial ŋ^2 = .58, and a significant interaction between system confidence and automation expertise on compliance *F*(3, 117) = 3.79, *p* < .05, partial ŋ^2 = .09.

As shown in Figure 1, when the aid was 25% confident, automation expertise did not significantly affect compliance with the novice (*M* = 10.09, *SD* = 4.72) or expert diagnostic aid (*M* = 8.95, *SD* = 4.52). However, when the aid was 75% confident, participants who interacted with the expert aid were more likely to report that a target was present (*M* = 20.34, *SD* = 3.76) than participants who interacted with the novice aid (*M* = 17.82, *SD* = 4.47), *F*(1, 39) = 3.84, *p* = .05. A similar effect occurred when the aid was 50% confident; participants were more likely to comply with the expert aid (*M* = 15.74, *SD* = 4.60) than the novice aid (*M* = 12.81, *SD* = 4.05), *F*(1, 39) =6.91, *p* < .05.

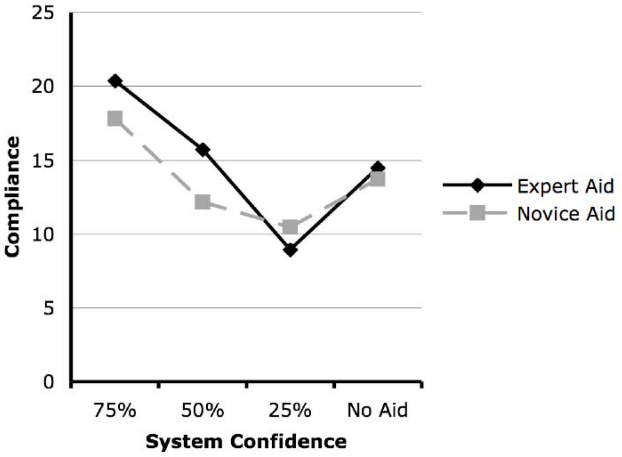


Figure 2: Compliance rates as a function of automation expertise and system confidence in the low image quality condition.

Probability of Committing a False Alarm

A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, no aid) x 2(image quality: high, low) mixed factorial ANOVA was calculated to assess main and interaction effects of image quality, automation expertise, system confidence, on the probability of committing a false alarm (pFA). Results indicated significant main effects for image quality *F*(1, 39) = 28.51, *p* < . 001 partial ŋ^2 = .42, and system confidence, *F*(3, 117) = 51.04, *p* < .001, partial ŋ^2 = .58. The interaction between image quality, system confidence, and automation expertise approached significance *F*(3, 117) = 2.39, *p* = .07 partial ŋ^2 = .06, power = .59. Because of the interaction, we analyzed the data within levels of image quality.

High image quality. A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, none) mixed ANOVA revealed a significant main effect for system confidence. Post hoc analysis indicated that participants were more likely to commit a false alarm when the was aid 75% (*M* = .63, *SD* = .31) confident than when the aid was 50% (*M* = .40, *SD* = .25) or 25% (*M* = .26, *SD* = .19) confident, or when they did not receive help from the aid (*M* = .36, *SD* = .22), *F*(3, 117) = 30.10, *p* < .001, partial ŋ^2 = .44, power = 1.00. All other effects were non-significant.

Low image quality. A 2(automation expertise: expert, novice) x 4(system confidence: 75%, 50%, 25%, none) mixed ANOVA indicated a significant main effect of system confidence *F*(3, 117) = 36.73, *p* < .001 partial ŋ^2 = .49, and a significant interaction of system confidence and automation expertise on pFA, *F*(3, 117) = 3.65, *p* < .05, partial ŋ^2 = .09, power = .79.

Simple main effect analyses for the significant interaction indicated when the aid was 25% confident automation expertise did not affect false alarm rates. However, when the aid was 75% confident participants who interacted with the expert aid committed more false alarms (*M* = .87, *SD* = .18) than participants who interacted with the novice aid (*M* = .74, *SD* = .26), *F*(1, 39) = 2.93, *p* = .09, albeit the effect did not reach statistical significance. A similar effect was found when the aid was 50% confident. As shown in Figure 3, participants who interacted with the expert aid were more prone to false alarms (*M* = .68, *SD* = .22) than the participants who interacted with the novice aid (*M* = .58, *SD* = .19) *F*(1, 39) = 7.81, *p* = .01 partial η^2 = .17.

Post hoc analysis for the significant main effect of system confidence indicated that participants were more likely to commit a false alarm when the was aid 75% (*M* = .81, *SD* = .23) confident than when the aid was 50% (*M* = .58, *SD* = .23) or 25% (*M* = .41, *SD* = .19) confident, or when they did not receive help from the aid (*M* = .57, *SD* = .22). The mean difference in false alarms between the no aid condition and the 75% confident and 25% confident condition was also statistically significant.

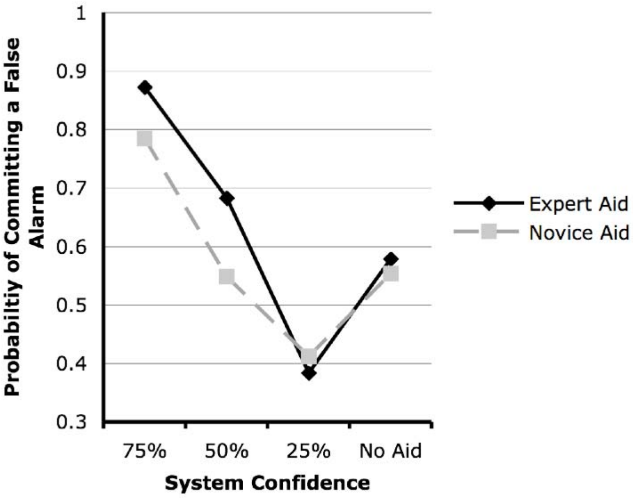


Figure 3: False alarm rates as a function of automation expertise and system confidence in the low image quality condition.

DISCUSSION

The goal of the current study was to determine the effects of automation expertise and system confidence on trust behaviors. As expected, system confidence influenced automation compliance. Specifically, participants calibrated their compliance rates to the system's level of confidence. These results support previous research that suggests system confidence can improve trust calibration (see McGuirl & Sarter, 2006). The interaction between system confidence and automation pedigree also confirmed our expectations. Individuals were more likely to accept advice from a highly confident expert aid than highly confident novice aid. These results suggest that, similar to interpersonal relationships, individuals weigh confidence information from novice and expert systems differently.

We also found an incurred cost of presenting system confidence information to operators. In our study, participants were more apt to generate false alarms as system confidence increased. Furthermore, expertise and confidence interact to influence false alarm rates. Participants were more likely to generate false alarms when the expert aid was highly and moderately confident in its diagnosis. This behavior resembles a form of automation bias in which operators use automated cues as a heuristic replacement for information processing. Automation bias can cause several unwanted consequences such as automation misuse and complacency.

Interestingly, in this study, automation expertise influenced compliance only when image quality was low. Lee and See (2004) indicated that automation trust can be influenced by surface level features of an automated aid via an analogical tuning method. Our results support this postulation and add to the literature by confirming that cues such as automation expertise emerge as a significant factor when operators are subdued with uncertainty.

In conclusion, this research adds to the theoretical literature on automation trust by demonstrating in a controlled study that trust behaviors, such as compliance, are influenced by the situational accuracy and dispositional characteristics of automated agents. These findings are particularly relevant given the rapid implementation of "expert" detection systems in complex task environments such as military command and control (C2) and homeland security. In these domains, operators rely on autonomous and semi-robotic agents to augment human performance. Certainly, briefing and training information concerning the expected performance and functioning of such systems can influence automation trust. Our study showed that these biases indeed affect trust behaviors. Furthermore, this study highlighted the effects of expert false alarms on trust. To date, this topic has not received extensive study. This is unfortunate because it is not only the quality of a system's algorithms that influence trust, but the manner in which these decisions are presented to the operator.

Future research should focus on best practices for conveying system confidence information to operators and methods for training operators to appropriately use system confidence information when making decisions. Future research should also focus on the effects of expert false alarms and misses on automation reliance and compliance. The heightened performance expectations associated with expert systems could influence the appraisal of automation errors, consequently decreasing dependence.

REFERENCES

Bliss, J. P. (2003). Collective mistrust of alarms. *International Journal of Applied Aviation Studies, 3*(1), 13-38.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology, 18,* 399-411.

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,* 697-781.

Lee, J. D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human machine systems. *Ergonomics, 35*, 1243-1270.

Madhavan, P., & Wiegmann, D. A. (2007b). Effects of information source, reliability, and pedigree on operator interaction with decision support systems. *Human Factors, 47*(2), 332-341.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48,* 656-665.

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*(3), 286-297.

Rice, S. (2009). Examining single and multiple-process theories of trust in automation. *Journal of General Psychology, 136*(3), 303-319.

Sheridan T., & Parasuraman, R. (2006). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*, 89-129.

Sniezek, J. A., & Von Swol, L. M. (2001). Trust, confidence and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes, 84*, 288-307.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human Computer Interaction, 1*, 49-75.