

EXPERTISE LEVELS OF HUMAN VERSUS AUTOMATED DECISION AIDS INFLUENCE RESPONSE BIASES IN A VISUAL SEARCH TASK

Poornima Madhavan¹ and Douglas A. Wiegmann²

¹Department of Social and Decision Sciences, Carnegie Mellon University

²Aviation Human Factors Division, Institute of Aviation,
University of Illinois at Urbana-Champaign

Studies have demonstrated that humans appear to apply norms of human-human interaction to interaction with automated decision aids. We examined the differences in perceptions of automation vs. humans when the expertise and reliability of these advisers varied. Participants ($n = 180$) performed a luggage-screening task with the assistance of human or automated advisers that differed in pedigree (expert vs. novice) and reliability (high vs. low), but had a similar neutral beta setting of 1.0. Shifts in sensitivity, criterion settings and accuracy were assessed. Participants who were presented with a low-reliable “expert” adviser shifted their bias away from the neutral bias of the adviser and more toward optimal beta compared to participants receiving unreliable advice from a ‘novice’. This effect increased across trials for participants using low-reliability automated advisers but not human advisers. The results have implications for the development of models of optimal utilization of decision aids.

INTRODUCTION

The introduction of automation into complex systems such as aircraft cockpits, nuclear power plants and air traffic control rooms has led to a redistribution of operational responsibility between human operators and computerized automated systems. Automated aids are increasingly being modeled as ‘partners’ that support or assist the human in performing functions that may be difficult for the operator to perform without the assistance of a ‘knowledgeable teammate’. Some researchers have argued that human-automation teams function similarly to human-human teams (Bowers, et al, 1996), and evidence suggests that people do enter into ‘relationships’ with computers and interactive machines in a manner similar to human partners (Nass, et al, 1993, 1996, 1999).

In general, automation is perceived as more credible than humans, and there appears to exist a ‘bias toward automation’ (Dzindolet, et al, 2001, 2002), which makes automation errors very salient to the observer. This leads to a rapid fall in trust when machines generate errors, leading to a breakdown in dependence on the machine While

existing research has established specific directions in the development of human-human vs. human-automation trust, there are several inconsistencies in the manner in which human vs. automated ‘advice’ has been operationalized. Studies have always portrayed the human adviser as ‘the previous participant’ and the automated aid as ‘a machine’ without specific information about the pedigree (expertise) of the adviser (e.g., Dijkstra, 1999). This raises to debate the validity and realism of comparisons between human and machine advisers.

Furthermore, existing research on the effects of source and pedigree on advice acceptance has exclusively focused on issues related to subjective trust, compliance and reliance (Madhavan & Wiegmann, 2005). The effects of different sources and pedigrees on users’ sensitivities, criterion settings and performance accuracy remains unexamined, which is critical to understanding the ultimate effect of diagnostic advice on the performance of the human-decision aid team. The purpose of the present study, therefore, was to examine the effects of information source, pedigree and reliability on indices of performance.

We used a factorial design in which participants performed a luggage screening task with the assistance of presumably different diagnostic sources (human vs. automated), and pedigrees (expert vs. novice), while observing the reliability (high vs. low) of advice in real time. The pedigree variable allowed for a higher degree of comparability between human and automated advisers. We examined the effects of the above sources and pedigrees on target detection sensitivities (d') and criterion settings (β), and their implications for the performance accuracy of the human-decision aid team.

We expected response criterion settings to be affected more than sensitivity in the present context since the nature of the advice was in the form of an indirect text cue (rather than a direct visual cue). Therefore, manipulation of the sources of advice was not expected to influence performance abilities per se.

On the contrary, the belief that an adviser was either human or automated, expert or novice, was expected to influence dependence strategies and consequently response tendencies or criterion settings. We expected that participants receiving advice from a low-reliability adviser would trust the advice less when the adviser was initially portrayed as an expert versus novice, due to a greater violation of initial expectations about expert reliability. Therefore, those receiving low-reliability 'expert' advice would anchor their responses bias less to the neutral bias of the advisor and would shift more toward an optimal β than those receiving low-reliability 'novice' advice. This pattern was hypothesized to be greater for automated vs. human advisers given a greater tendency of users to expect perfect performance from automation.

METHOD

Participants

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

Tasks and Procedures

Participants performed 200 trials of a computer simulated airline luggage-screening task wherein they detected the presence of a hidden knife in passenger baggage. The stimuli consisted of two-color x-ray images of luggage, cluttered with everyday objects. 20% ($n = 40$) of the images contained a digitally superimposed x-ray image of a knife. The knives were of two types, with four rotations of each. The task was to observe the stimulus image and decide whether to stop or pass the luggage. Participants received assistance from an adviser whom they were told was either an automated system or a human participant, novice or expert. In reality, the same computer program generated advice for all participants.

At the onset of each trial, a piece of luggage appeared on the monitor for three seconds, following which aided participants were presented the decision of the adviser in the form of a text message on the screen. Participants then input their diagnosis and received feedback as to the accuracy of their diagnosis. The probability of the adviser committing a hit/correct rejection or a miss/false alarm was either .70 and .30 (low reliability condition), or .90 and .10 (high reliability condition). The criterion setting (β) for both aids was "1" (neutral setting) resulting in equal probabilities of hits and correct rejections. Participants had no information about the reliability of their advisers. The source (human or automated), reliability (low or high) and pedigree (novice or expert) of the advice varied as below:

- (1) "Novice-human" groups: The adviser was portrayed as a novice student.
- (2) "Expert-human" groups: The adviser was an expert in airport security.
- (3) "Novice-automation" groups: The adviser was a novice computer program.
- (4) "Expert-automation" groups: The adviser was an expert computer program.

Prior to testing, participants were assigned to one of eight experimental groups (i.e., (1) novice-human-low reliability, (2) expert-human-low reliability, (3) novice-automation-low reliability, (4) expert-automation-low reliability, (5) novice-human-high reliability, (6) expert-human-high reliability, (7)

novice-automation-high reliability, and (8) expert-automation-high reliability). A control group performed the task unaided.

RESULTS

We divided the data into 5 blocks of 40 trials each in order to trace shifts in sensitivities and criterion settings during the course of the task. Several of the hypotheses we generated were directional. Nonetheless, we used two-tailed ANOVAs to analyze our data to maintain consistency across analyses. P-values less or equal to 0.05 were considered statistically significant. However, theoretically important effects involving alpha values of 0.10 or less were also further perused in some instances.

Sensitivity (d'). A 2 (source: human vs. automated adviser) X 2 (pedigree: expert vs. novice) X 2 (reliability: 90% vs. 70% reliable adviser) X 5 (trial block) ANOVA on sensitivities revealed a significant main effect only for reliability, $F(1, 152) = 143.14, p < .01$. Differences in performance accuracy between participants using 90% versus 70% reliable advisers were entirely due to differences in sensitivity as a consequence of differing adviser reliabilities. The lack of significant source and pedigree effects on sensitivities suggests that performance differences among aided groups that go beyond reliability differences were due to beta shifts

Criterion settings (beta) and accuracy. Given the ratio of noise (.80) to signal trials (.20), a beta of 4.0 would be the optimal setting in this situation. However, the adviser in this study was designed to be “unbiased” and therefore had a beta setting of 1.0.

The criterion settings of participants utilizing 90% reliable advice were not affected significantly by the pedigree and source. Therefore, we present the data for those receiving 70% reliable advice alone.

Participants using 70% reliable advisers demonstrated a significant upward shift away from the adviser’s neutral beta, $t(79) = 2.66, p < .01, d = .54$, suggesting that participants consciously adopted liberal criterion settings. Similarly, the average criterion setting of unaided participants ($M = 1.38, SD = .44$) was significantly higher than a

neutral bias of 1 across all trial blocks, $t(19) = 3.43, p < .01, d = 1.22$. As can be seen in Figure 1, participants receiving advice from ‘novice’ advisers did not deviate significantly from the beta setting of unaided participants, $t(58) = .44, p = .66, d = .12$. On the other hand, participants receiving ‘expert’ advice (both human and automated) demonstrated an upward shift in the direction of optimal beta, thereby demonstrating a greater trend toward conservative responding relative to unaided participants, $t(58) = 1.74, p = .06, d = .46$.

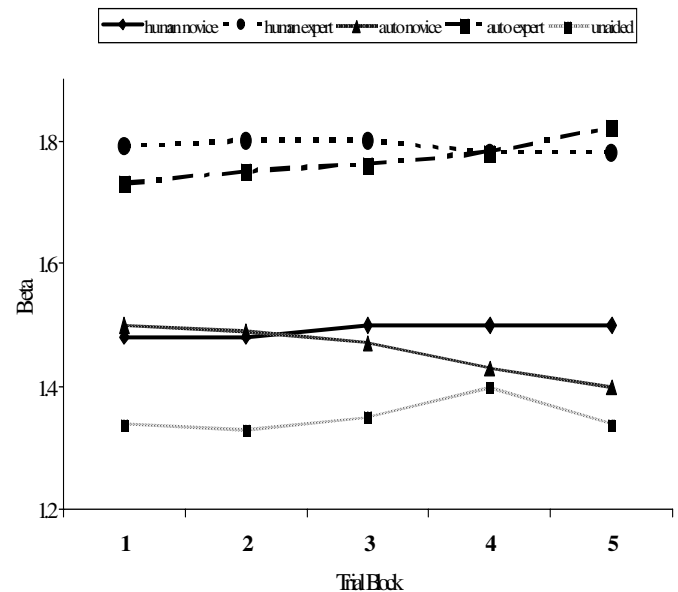


Figure 1. Criterion settings (note: optimal beta is 4.0; adviser’s beta is 1.0)

As illustrated in Figure 1, when advisers were portrayed as humans, criterion settings of participants receiving ‘expert’ advice ($M = 1.79, SD = .35$) were higher than those using receiving ‘novice’ advice, ($M = 1.49, SD = .57$), $t(38) = 1.88, p = .08, d = .61$, across all trial blocks. As a result, participants using ‘expert’ human advisers generated fewer hits ($M = .48, SD = .15$) but fewer false alarms ($M = .26, SD = .11$) than those using ‘novice’ human advisers, who demonstrated a relatively smaller shift from the adviser’s beta. Consequently the latter generated more hits ($M = .53, SD = .13$), $t(38) = 1.90, p = .07, d = .62$, and more false alarms ($M = .29, SD = .14, t(38) = 1.82, p = .09, d = .59$), than the former.

Similar to human advisers, criterion settings of participants using ‘expert’ automated advisers ($M = 1.75, SD = .33$) were higher than the criterion

settings of those using ‘novice’ automation ($M = 1.46$, $SD = .27$), $t(38) = 1.05$, $p = .06$, $d = .34$, across all blocks. Consequently, participants using ‘expert’ automated advisers generated fewer hits ($M = .50$, $SD = .17$) but fewer false alarms ($M = .25$, $SD = .11$) than those using ‘novice’ automation, who demonstrated a relatively smaller shift from the aid’s neutral beta (hit rate: $M = .55$, $SD = .12$, $t(38) = 1.12$, $p = .07$, $d = .36$; false alarm rate: $M = .29$, $SD = .09$, $t(38) = 1.16$, $p = .07$, $d = .38$).

In support of hypotheses, criterion settings of participants using automated advisers demonstrated slightly different patterns of change over the course of the task as a function of pedigree (illustrated in Figure 1). Participants using ‘expert’ automated advisers demonstrated a significant *upward* shift in criterion settings (or a shift toward optimal beta) (first block: $M = 1.73$, $SD = .33$; fifth block; $M = 1.82$, $SD = .40$, $t(19) = 1.99$, $p < .05$, $d = .91$). Participants receiving ‘novice’ automated advice demonstrated a *downward* shift in criterion settings (or a shift toward the adviser’s neutral beta) (first block: $M = 1.5$, $SD = .26$; fifth block: $M = 1.4$, $SD = .27$, $t(19) = 1.21$, $p = .09$, $d = .56$). Consequently, participants receiving advice from ‘expert’ automation generated a larger number of hits ($M = .59$, $SD = .17$) and false alarms ($M = .30$, $SD = .19$) in the first block than in the last block (hits: $M = .45$, $SD = .16$, $t(19) = 1.18$, $p < .05$, $d = .54$; false alarms: $M = .20$, $SD = .10$, $t(19) = 1.11$, $p = .07$, $d = .51$).

The opposite was true for participants using ‘novice’ automation, who generated fewer hits ($M = .50$, $SD = .13$) and fewer false alarms ($M = .27$, $SD = .10$) in the first block than in the last (hits: $M = .60$, $SD = .11$, $t(19) = 1.06$, $p = .08$, $d = .49$; false alarms: $M = .35$, $SD = .12$, $t(19) = 1.30$, $p < .05$, $d = .60$). In contrast, participants receiving advice from humans showed no significant changes in response patterns across blocks.

DISCUSSION

In this study, sensitivity differences merely led to global differences in performance between participants using 90% vs. 70% reliable advisers. Performance differences beyond broad reliability differences were primarily due to shifts in decision criteria. In general, aided participants demonstrated

a significant upward shift in criterion settings toward optimal beta, *away* from the adviser’s neutral beta. This suggests that all participants consciously attempted to develop their own response criteria moving in the direction of optimal beta, as a consequence of their degree of trust in their advisers.

Analyses did not reveal any global differences in the effect of human vs. automated advisers on criterion settings. However, there were significant differences in the pattern of criterion shifts as a function of pedigree. Participants receiving ‘expert’ advice were prone to “target absent” responses, demonstrated a greater shift from the adviser’s neutral beta and had a proportionately conservative response criterion setting; participants receiving ‘novice’ advice were prone to “target present” responses, demonstrated a smaller shift from neutral beta with a consequently more liberal criterion setting and, as a result, generated a large number of hits and false alarms than those using ‘expert’ advisers.

From the Elaboration Likelihood Model of persuasion (Petty & Cacioppo, 1986), it follows that participants receiving advice from ‘novices’ had lower initial expectations of the accuracy of their advisers. Therefore, they intentionally set their decision criteria at a more liberal level possibly to maximize their hits and minimize the probability of their missing a target, if their adviser missed a target.

Contrary to participants receiving advice from ‘novices’, participants receiving advice from ‘experts’ had higher initial expectations of their advisers’ abilities. Therefore, their decision criteria were set at a more conservative level in order to minimize the probability of false alarms while simultaneously maximizing the probability of their detecting a target when the adviser detected one. As hypothesized, since these initial high expectations for those using ‘expert’ advisers were violated by the low accuracy level of 70% reliable ‘expert’ advisers, participants presumably developed their own beta by shifting towards optimal. Conversely, the data suggests that participants receiving assistance from ‘novice’ advisers anchored more strongly to the response bias of their advisers and did not deviate as much from this neutral bias

because their trust was not violated as much as for those using ‘experts’.

The occurrence of diagnostic errors in this study was determined largely by shifts in decision bias and not by differences in sensitivity. This indicates that the pedigree and source of advice affect performance via psychological effects that lead to subjective shifts in decision biases on the part of the human operator. Results of this study have important implications for contexts developing models of operator optimal utilization of decision aids.

Conclusions

Research has shown that characterizing a decision support system as an expert system has effects that are fundamentally different from those of a source characterized as a human (Lerch, et al., 1997). The present study provides an initial answer to the question of whether the differential impact of source and pedigree would be salient if expertise were not just a unique property associated with automated systems but also with human advisers.

From the results of the present study, it follows that users’ utilization of decision aids is dependent not just on the recommendations of the advisers, but also on the assumed credibility of the source of information and the effect this has on users’ perceptions of the efficacy of advisers. Thus, it is to the exploration of the internal attributes of these systems that attention must be paid, in order to generate recommendations for the successful design of future decision aids and to ensure the seamless flow of communication between systems and humans.

REFERENCES

- Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors*, 40(4), 672-680.
- Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology*, 18(6), 399-411.
- 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.
- 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: The MIT Press.
- Madhavan, P., & Wiegmann, D.A. (2005). Effects of information source, pedigree and reliability on operators’ utilization of diagnostic advice, *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*, pp. 487-491. Santa Monica: CA.
- Nass, C., Fogg, B.J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, 45(6), 669-678.
- Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, (29(5), 1093-1110.
- Nass, C., Steuer, J., Tauber, E., & Reeder, H. (1993). Anthropomorphism, agency and ethopoeia: computers as social actors, *Proceedings of the International CHI Conference* (Amsterdam).
- Petty, R. E., & Cacioppo, J. T. (1986b). The elaboration likelihood model of persuasion. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Vol 19, pp. 123-205. New York: Academic Press.