

EFFECTS OF INFORMATION SOURCE, PEDIGREE, AND RELIABILITY ON OPERATORS' UTILIZATION OF DIAGNOSTIC ADVICE

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 their interaction with machines. Yet, there exist subtle differences in peoples' perceptions of automated aids compared to humans. We examined factors differentiating human-human and human-automation interaction, wherein 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). Dependence on advice was assessed. Participants agreed more with an automated 'novice' than a human 'novice' suggesting a bias toward automation. Automation biases broke down when automated aids portrayed as 'experts' generated errors, leading to a drop in compliance and reliance on automation relative to humans. The results have implications for the development of theoretical and computational models of optimal user dependence on 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' rather than as tools. (Klein, Woods, Bradshaw, Hoffman & Feltovich, 2004) These 'partners' support or assist in performing functions that may either be difficult or even impossible for humans to perform independently.

The Transportation Safety Administration (TSA) is currently exploring the efficacy of implementing such automated aids for assisting luggage screeners detect the presence of hazardous items in passenger baggage. Such diagnostic aids highlight suspicious objects in a piece of luggage as it passes through the x-ray machine, thereby providing valuable diagnostic assistance to the human screener and enhancing the overall quality of aviation security. Other examples of 'intelligent' decision aids that assist the human operator in performing complex and critical tasks are the Flight Management System (FMS) in the cockpit that is designed to provide pilots with critical advice on route planning, navigation and traffic patterns while detecting and diagnosing abnormalities in the

flight path (Sheridan, 2002), and computer-based aids used in radiology that assist the physician detect the presence of tumors and other anomalies in patient x-rays (Krupinski, Nodine & Kundel, 1993).

Decision aids such as those described above are designed to interact or behave in a manner similar to a human, imitating human language structures where applicable and often possessing unique knowledge and functional algorithms that may be inaccessible to the human teammate. Indeed, some researchers have argued that such human-automation teams function similarly to human-human teams (Bowers, Jentsch, Salas & Braun, 1998), and scientific evidence suggests that people do enter into 'relationships' with computers, robots, and interactive machines in a manner similar to other humans (Nass, Fogg & Moon, 1996). Much of the data supporting these assumptions have come from the 'Computers Are Social Actors' or CASA studies (Nass, Steuer, Tauber, & Reeder 1993; Nass & Moon, 2000) that have demonstrated that social rules guiding human-human interaction may apply equally to human-computer interaction, with users responding to machines as independent entities. Specifically, research has revealed that people apply politeness norms and gender stereotypes to computers, and reportedly get 'attracted' to computers with 'personalities' that match their own (cf. Nass & Lee,

2001). Furthermore, researchers contend that strong social bonds between humans and computers can be created when a computer is labeled a 'teammate' (cf. Nass, Moon & Carney, 1999).

Contrary to the CASA studies, however, are suggestions that the decision making processes of human-machine teams are often influenced strongly by operators' *trust* in an automated team member relative to a human partner (e.g. Dijkstra, 1999). These researchers have found that initial trust in automation tends to be higher than trust humans due to the existence of a bias toward automation or a "perfect automation schema" (Dzindolet, Pierce, Beck, Dawe & Anderson, 2001). However, this positive bias toward automation leads operators to be more sensitive to the errors made by automation than by a human, leading to a sharper drop in trust and dependence when machines generate errors (Dzindolet et al., 2001). This concept of automation trust, in general, has been the focus of a vast body of research over the last decade (e.g. Muir, 1994; Lee & Moray, 1992; Lerch, Prietula & Kulik, 1997; Lee & See, 2004). However, there are several inconsistencies in the current literature that need to be resolved.

Purpose of the Present Study

One of the primary drawbacks in existing research is that most comparisons between human and automated advisers have been made without associating any specific expertise level or pedigree with advisers (e.g., Dijkstra, 1999; Dzindolet et al., 2001). This raises to question the validity of comparisons between humans and machines.

Secondly, there are several inconsistencies in the measures of trust and dependence on advice used in existing literature. Some studies have used scales of trust, while others have interpreted confidence and perceived reliability of advice as synonymous with trust in advice (cf. Lerch et al., 1997). Likewise, agreement with advice has always been used as an index of dependence without further dichotomizing dependence into compliance and reliance (Meyer, 2004). These vast discrepancies in methodology have led to conflicting results in the existing research on human-human versus human-automation trust. Therefore, the comparison of human-human interaction to human-automation interaction requires a multi-pronged approach, as described in the following study.

In the present study, participants performed a luggage-screening task with the assistance of human or automated advisers that varied in reliability and pedigree (expertise). The pedigree variable allowed for a higher degree of comparability between human and automated advisers. Advice dependence was assessed by performance measures, namely agreement with the adviser's diagnoses, which was further dichotomized into compliance and reliance strategies.

Based on existing research, we hypothesized that:

- (1) Agreement (compliance/reliance) with 90% reliable advisers will be higher than with 70% reliable advisers due to the objective accuracy of the former.
- (2) Agreement with 'expert' advisers will be higher than with 'novice' advisers due to the greater dispositional credibility of the former.
- (3) When advisers are 70% reliable, agreement with automated advisers will be higher than with human advisers when advisers are portrayed as 'novices', due to the higher perceived credibility of automation relative to humans (similar to the effect for 90% reliable advisers). On the other hand, when portrayed as 'experts', we expected agreement with the automated 'expert' to drop below that of the human over the course of trial blocks. This is because the errors generated by automation (with 30% frequency) are more likely to be noticed by users due to the disruption of the "perfect automation schema" (Dzindolet, et al., 2001), leading to a rapid breakdown in agreement with 'expert' automated advisers relative to 'expert' humans.

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 that they believed to be 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 (beta) for both aids was “1” suggesting 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 participants’ agreement data into 5 blocks of 40 trials each in order to trace changes in agreement with advice during the course of the task. This was done by computing the conditional probability of participants’ tendencies to either say “yes, target present” or “no, target absent” as a

function of the adviser’s response. Compliance probabilities were calculated as the conditional probability of the participant saying “yes, target present” given the adviser said “yes, target present” or $p(Oy|Ay)$. Reliance probabilities were calculated as the conditional probability of the participant saying, “no, target absent” given the adviser said “no, target absent” or $p(On|An)$. Overall, agreement with 90% ($M = .51, SD = .11$) reliable advisers was always higher than with 70% reliable advisers ($M = .41, SD = .10$), with no effect of source and pedigree on the former. We therefore present compliance and reliance data only for participants using 70% reliable advice.

Compliance strategies. A 2 (source: human vs. automated adviser) X 2 (pedigree: expert vs. novice) X 5 (trial block) mixed ANOVA on *compliance* with 70% reliable advisers revealed no significant main effects. However, results revealed significant interactions between source and trial block, $F(4, 304) = 2.73, p < .05$, pedigree and trial block, $F(4, 304) = 3.69, p < .01$, and, source, pedigree and block, $F(4, 304) = 4.31, p < .01$.

Participants’ compliance strategies are illustrated in Figure 1. When 70% reliable advisers were portrayed as ‘novices’, compliance with the *automated* adviser ($M = .56, SD = .21$) was significantly *higher* than with the ‘novice’ *human* adviser ($M = .52, SD = .20$), $t(38) = 1.43, p = .067, d = .46$, throughout the task. However, participants using ‘expert’ advisers demonstrated an initial pattern of compliance that was contrary to that observed for ‘novices’. There were no significant differences in compliance between ‘expert’ *human* and *automated* advisers until the fourth trial block. In the fourth and fifth trial blocks, compliance with the ‘expert’ *automated* adviser ($M = .41, SD = .17$), dropped significantly *below* that of the ‘expert’ *human* adviser ($M = .58, SD = .21$), $t(38) = 2.70, p < .05, d = .88$.

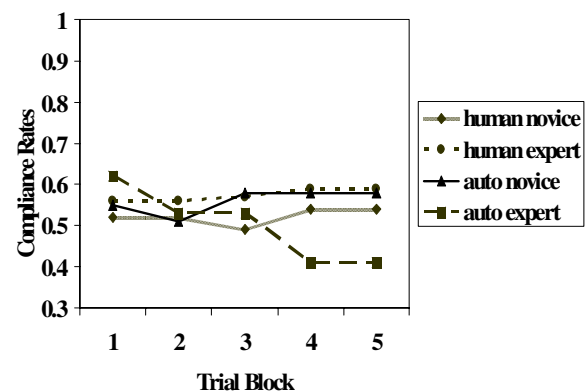


Figure 1. Compliance rates

Reliance strategies. A 2 (source: human vs. automated adviser) X 2 (pedigree: expert vs. novice) X 5 (trial block) mixed ANOVA on *reliance* rates revealed a significant main effect for trial block, $F(4, 304) = 4.33, p < .01$, an interaction between pedigree and trial block, $F(4, 304) = 1.63, p = .06$, as well as a three-way interaction between source, pedigree and block, $F(4, 304) = 1.88, p < .05$.

As illustrated in Figure 2, when advisers were portrayed as ‘novices’, reliance on the *automated* adviser ($M = .85, SD = .099$) was significantly *higher* than on the ‘novice’ *human* adviser ($M = .78, SD = .14$), $t(38) = 1.86, p = .07, d = .60$, throughout the task, which was similar to the pattern observed on compliance trials. There were no significant differences in reliance between ‘expert’ *human* and *automated* advisers until the last trial block. However, similar to compliance patterns, reliance on advice in the fifth trial block for participants using the ‘expert’ *automated* adviser ($M = .81, SD = .10$), was significantly *lower* than on ‘expert’ *human* adviser ($M = .84, SD = .12$), $t(38) = 1.66, p = .051, d = .54$.

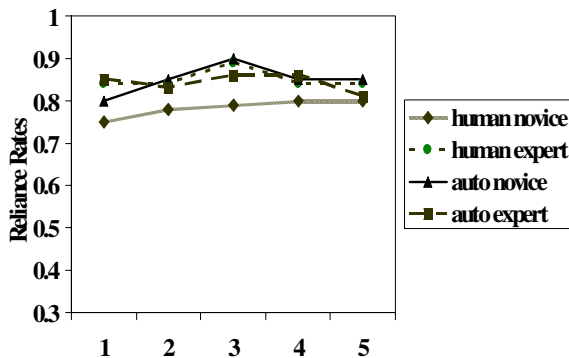


Figure 2. Reliance rates

DISCUSSION

In the present study, participants complied with and relied more on advisers that were 90% reliable than those that were 70% reliable. When advice was 90% reliable, there were no significant effects of source and pedigree on advice acceptance. This suggests that when DSSs are highly reliable, near-perfect adviser accuracy has a stronger psychological impact on advice utilization than dispositional information.

When advice was 70% reliable, participants demonstrated differences in advice acceptance as a function of both information source and pedigree. When advisers were portrayed as ‘novices’, agreement

with the automated adviser was significantly higher than with the human across all trial blocks. However, participants using ‘expert’ advisers demonstrated a slightly different pattern of dependence. Agreement with the automated ‘expert’ did not differ significantly from the human ‘expert’ at the beginning of the task. However, there was a drop in agreement with the automated ‘expert’ leading to a complete reversal in agreement patterns by the last trial block. At the end of the task, agreement with the automated ‘expert’ was significantly *lower* than with the human ‘expert’.

The above results suggest that different psychological factors influence operators’ development of advice utilization strategies when sources of information are either human or automated. When the source of diagnostic information is a human adviser, there appears to be a strong effect of dispositional credibility on advice acceptance. Merely portraying a human as an ‘expert’ leads users to agree more with advice regardless of the situational accuracy of the adviser. Conversely, automated aids are judged more by situational factors than dispositional traits (cf. Lerch et al., 1997). In keeping with hypotheses, participants receiving 70% reliable automated advice demonstrated a pattern of advice dependence that reflects a breakdown in the ‘perfect automation schema’ over the course of the task. This is due to an interaction of high expectations with continued experience of low-reliability automation. When automated aids were portrayed as ‘experts’, initial expectations were likely very high. However, when participants repeatedly observed their 70% reliable automated adviser generating errors on 30% of occasions, it led to a dramatic negative trend in agreement with advice. Interestingly, operators seem more forgiving of errors generated by an ‘expert’ human adviser compared to an ‘expert’ automated adviser.

Conclusions

Research has shown that characterizing a DSS as an expert system has effects that are fundamentally different from those of a source characterized as a human (Lerch, et al., 1997). The results of 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, it follows that users' utilization of decision support systems is dependent not just on the recommendations of the advisers, but also on the assumed credibility of decision aids 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 support systems 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., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, 13(3), 147-164.
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 4, 1541-1672.
- Krupinski, E. A., Nodine, C. F., & Kundel, H. L. (1993). Perceptual enhancement of tumor targets in chest x-ray images. *Perception and Psychophysics*, 53 (5), 519-526.
- Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human machine systems. *Ergonomics*, 22 (6), 671-691.
- Lee, J. D., & See, S. A. (2004). Trust in automation: Designing for appropriate reliance, *Human Factors*.
- 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.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905-1922.
- Nass, C., Fogg, B.J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, 45(6), 669-678.
- Nass, C., & Lee, K.M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency attraction. *Journal of Experimental Psychology: Applied*, 7, 171-181.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103.
- 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).
- Sheridan (2002). Human performance in relation to automation. In T. B. Sheridan (Ed.), *Human and Automation: System Design and Research Issues*. (pp. 69-89). Santa Monica, CA: John Wiley.