

Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems

Poornima Madhavan and Douglas A. Wiegmann, University of Illinois at Urbana-Champaign, Champaign, Illinois

Objective: Two experiments are described that examined operators' perceptions of decision aids. **Background:** Research has suggested certain biases against automation that influence human interaction with automation. We differentiated preconceived biases from post hoc biases and examined their effects on advice acceptance. **Method:** In Study 1 we examined operators' trust in and perceived reliability of humans versus automation of varying pedigree (expert vs. novice), based on written descriptions of these advisers prior to operators' interacting with these advisers. In Study 2 we examined participants' post hoc trust in, perceived reliability of, and dependence on these advisers after their objective experience of advisers' reliability (90% vs. 70%) in a luggage-screening task. **Results:** In Study 1 measures of perceived reliability indicated that automation was perceived as more reliable than humans across pedigrees. Measures of trust indicated that automated "novices" were trusted more than human "novices"; human "experts" were trusted more than automated "experts." In Study 2, perceived reliability varied as a function of pedigree, whereas subjective trust was always higher for automation than for humans. Advice acceptance from novice automation was always higher than from novice humans. However, when advisers were 70% reliable, errors generated by expert automation led to a drop in compliance/reliance on expert automation relative to expert humans. **Conclusion:** Preconceived expectations of automation influence the use of these aids in actual tasks. **Application:** The results provide a reference point for deriving indices of "optimal" user interaction with decision aids and for developing frameworks of trust in decision support systems.

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. The role of the human operator, therefore, has metamorphosed from that of a primary controller to that of an active teammate sharing control with automation. Specifically, automated decision 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 be difficult or even impossible for humans to perform independently.

Decision support systems (DSSs) such as these

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. The earliest attempt at distinguishing operator trust in and use of different types of DSS goes back to the research of Turing (1950), who found that, when the source of a piece of diagnostic information was masked, human judges were largely unable to distinguish between a human adviser and a machine adviser imitating a human.

More recently, several researchers have found that human-automation teams and human-human teams function similarly (Bowers, Jentsch, Salas, & Braun, 1998; Bowers, Oser, Salas, & Cannon-Bowers, 1996). Evidence suggests that people enter into "relationships" with computers, robots,

and interactive machines which are similar to their relationships with other humans (Nass, Fogg, & Moon, 1996; Reeves & Nass, 1996). Much of the data supporting these assumptions has come from the Computers Are Social Actors (CASA) studies (Nass & Moon, 2000; Nass, Moon, Fogg, Reeves, & Dryer, 1995), which have demonstrated that social rules guiding human-human interaction may apply equally to human-computer interaction, with users responding to machines as independent entities rather than as a manifestation of their human creators (Sundar & Nass, 2000).

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 in humans because of 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). Lewandowsky, Mundy, and Tan (2000) found that delegation to human collaborators was slightly different from delegation to automation. In human-automation partnerships, operators perceive the ultimate responsibility to lie with the operator, whereas operators in a human-human partnership may perceive the ultimate responsibility as being shared (Lewandowsky et al., 2000).

Lee and See (2004) pointed out that the primary reason trust in humans differs from trust in machines is that the latter lacks "intentionality" or traits such as loyalty, benevolence, and values that are critical to the development of trust in a human partner. This leads to differences in the process of trust development over the course of a relationship. Human-human trust begins with a basis in performance or reliability, progresses to the level of dependability, and finally evolves to faith (Rempel, Holmes, & Zanna, 1985); trust in automation follows the opposite pattern of development, with faith playing a major role in the initial stages, followed by dependability, and then by predictability (Muir & Moray, 1996).

As is evident from the preceding discussion, the concept of trust in humans versus machines has

been the focus of a vast body of research over the last decade. However, there are some marked discrepancies in the current literature. Most studies on human-automation trust have typically modeled the human adviser as "the previous participant" and the automated aid as "a machine," without providing participants any background information about the pedigree or expertise levels of these advisers (e.g., Dijkstra, 1999). Such minimal information does not provide sufficient grounds for drawing effective comparisons between humans and machines, and this questions the validity of the experimental scenarios created. In a real task context, operators invariably have access to background information about the expertise levels of their advisers that would influence their advice acceptance.

Only one study has attempted to incorporate pedigree information into the experimental scenario. Lerch, Prietula, and Kulik (1997) had participants interact with either an "expert" or a "novice" human adviser pitted against an automated aid. However, they did not use a completely crossed design and failed to present the automated aid as "expert versus novice." Sheridan (2002) and Parasuraman and Riley (1997) have opined that there is a tendency for operators to be somewhat distrustful of new warnings, alarms, or DSSs until such aids have proven themselves. Nonetheless, studies have yet to be conducted in which participants are explicitly informed that the automated aid is "new" or unproven in terms of its reliability. Whether trust in such novice automation is different from trust in other types of advisers needs to be empirically examined.

Existing research on trust measurement has seldom attempted to address critical differences between operators' preconceived biases of a DSS before observing the adviser's situational reliability and the development of trust after experiencing the accuracy of the DSS during the course of a task. One possible way to examine the translation of initial biases into post hoc opinions would be to have participants provide their a priori expectations before a task and then measure their opinions after the conclusion of the task as well. However, if participants were to declare their expectations of an adviser before beginning the task, it could serve as a "cognitive anchor" (see Madhavan & Wiegmann, 2005) that significantly affects the manner in which they perceive the utility of the adviser during the course of the task.

In most real-world tasks, initial biases and notions about human and automated advisers are bound to be largely implicit in nature—that is, users are rarely required to verbally document their opinions of advisers prior to their interaction with these advisers. Therefore, it is important to first establish empirically whether such implicit biases exist prior to operator-DSS interaction, which was the purpose of Study 1. Second, it is of equal importance to examine how biases ultimately affect trust calibration and the use of these DSSs in an actual task context, which was done in Study 2. Different participants were used in the two studies so as to avoid psychological issues related to cognitive anchoring.

An additional inconsistency in the existing research is that most studies have treated “agreement” with advice as being synonymous with dependence. However, Meyer’s (2004) current dichotomy of compliance-reliance promises to be a more effective standpoint from which to quantify advice use processes. According to Meyer (2004), *compliance* refers to the probability of agreeing with advice when a DSS generates a diagnosis of “target present”; *reliance* refers to the probability of agreeing with a diagnosis of “target absent.”

It is important to represent agreement as compliance versus reliance because the compliance-reliance trade-off determines the types of errors generated by operators during a task, thereby providing the opportunity for the direct mapping of advice use with performance efficiency. The second purpose of Study 2, therefore, was to explore the manner in which users comply with and rely on advice and to examine its ultimate implications for the performance efficacy of the operator-DSS team.

STUDY 1: ASSESSING A PRIORI BIASES

The purpose of the first study was to assess users’ preconceived opinions of human versus automated advisers, coupled with differences in pedigree/expertise, before they interacted with advisers in an actual task.

Method

Participants. Forty undergraduate and graduate students (18 men, 22 women; mean age = 22.5 years) completed the experiment. Participants were paid \$8 for 1 hr of participation.

Procedure. Participants were instructed that they were soon to perform a complex luggage in-

spection task with the assistance of four different advisers. Participants read written descriptions of each adviser, as shown in the Appendix (but without the labels). Participants then rated their subjective trust in and perceived reliability of each adviser on adjoining questionnaires. The first questionnaire, adapted from Jian, Bisantz, and Drury’s (2000) System Trust Scale, addressed participants’ trust in the adviser on a scale of 1 (*strongly disagree*) to 10 (*strongly agree*). The second questionnaire measured participants’ perceived reliability of the adviser, again on a 10-point scale. These scales are internally consistent, reliable, and valid (Safar & Turner, 2005). All participants received all descriptions, and the scenarios were counterbalanced across participants.

Hypotheses. Based on findings on human perception of automated and human advisers (e.g., Dzindolet, Pierce, Beck, & Dawe, 2002; Dzindolet et al., 2001; Lerch et al., 1997) and the effects of framing on automation trust (e.g., Dzindolet et al., 2002; Lacson, Wiegmann, & Madhavan, 2003), we hypothesized trust in automation to be higher than in humans when portrayed as novices, as automation is perceived as more rational than humans (e.g., Dijkstra, 1999). When both are portrayed as experts, we expected human advisers to be trusted more than automated aids, as portraying a human as an expert will lead to the association of higher dispositional credibility (i.e., degree of trustworthiness based on personal traits) with the human (see Lerch et al., 1997).

We predicted automated advisers to be perceived as more reliable than human advisers at both pedigree levels. This was based on our theoretical premise that “perceived reliability” (i.e., expected performance accuracy) is an index of performance and is less likely to be influenced by disposition.

Results

We analyzed the data using 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) within-subjects two-tailed ANOVAs. These were followed by post hoc comparisons, with alpha values less than .05 being reported as significant. Additionally, we used Cohen’s *d* as a measure of effect sizes.

Subjective trust ratings. A 2 (source: human vs. automated adviser) \times 2 (pedigree: novice vs. expert adviser) within-subjects ANOVA on consolidated trust scores revealed significant main effects

for source, $F(1, 156) = 2.9, p = .0009$, and pedigree, $F(1, 156) = 108.87, p = .0001$, as well as an interaction between source and pedigree, $F(1, 156) = 10.42, p = .008$. In support of the hypotheses, expert advisers were perceived as more trustworthy than novices. When advisers were portrayed as novices, trust in the automated aid ($M = 6.18, SD = 1.42$) was higher than in the human adviser ($M = 5.15, SD = 1.56$), $t(78) = 3.06, p = .0091, d = 0.69$. However, when advisers were portrayed as experts, trust was higher in the human ($M = 8.00, SD = 1.08$) than in the automated adviser ($M = 7.00, SD = 1.13$), $t(78) = 1.28, p = .01, d = 0.30$.

Perceived reliability ratings. A 2 (source: human vs. automated adviser) \times 2 (pedigree: novice vs. expert adviser) within-subjects ANOVA on consolidated reliability scores revealed significant main effects for source, $F(1, 156) = 5.4, p = .007$, and pedigree, $F(1, 156) = 119.41, p < .0002$, but no interaction between source and pedigree, $F(1, 156) = 0.12, p = .31$. Contrary to the pattern for trust ratings, automated advisers ($M = 6.95, SD = 1.63$) were perceived as significantly more reliable than human advisers ($M = 5.95, SD = 1.39$), $t(78) = 2.71, p = .043, d = 0.62$, at both levels of pedigree.

Discussion

Subjective biases and preconceived notions are often largely responsible for operators' choices to trust and use a DSS. Research suggests that subtle differences exist in human perceptions of automated aids as compared with human advisers (see Dijkstra, 1999; Dzindolet et al., 2001; Lerch et al., 1997). According to Dzindolet et al. (2001), users have preconceived notions of automated aids being perfect or near perfect ("perfect automation schema"); humans, on the other hand, are judged to be less perfect but more consistent in terms of traits and behavioral patterns (Dijkstra, 1999). The primary purpose of Study 1 was to isolate such preconceived notions regarding human versus automated advisers.

In keeping with the hypotheses, participants demonstrated a higher degree of trust in automation aids than in humans when the advisers were novices. However, when they were portrayed as experts, there was evidence of a reversal in subjective trust estimates, which represents an interaction between preconceived biases about information source and pedigree. Research has revealed that humans judge other humans based on traits, whereas automation is judged more by its performance

in a particular context (Lerch et al., 1997). In the present study, participants merely read descriptions of their advisers without being provided any specific information about (or experience of) the advisers' performance accuracy. Therefore, participants' subjective assessments of trust were largely influenced by pedigree or apparent credibility of advisers.

When advisers were portrayed as novices, the credibility of the human was low and generated a low level of trust relative to the automated adviser. On the other hand, when advisers were portrayed as experts, it likely raised the dispositional credibility of the human above that of the automated adviser, eventually leading to a reversal in the perceived trustworthiness of advice. Because humans are judged more by dispositional factors than are automated aids (see Lerch et al., 1997) and participants were provided information on the dispositional features of their advisers alone, with no information on their task performance, this difference in background information provides an explanation for the observed skew in favor of human advisers in the present study.

Participants demonstrated a different pattern of preference in their estimates of perceived reliability of advice, which was always higher for automation than for humans, both for novice and expert advisers. Reliability addresses an entity's ability to perform a particular task, as opposed to trust, which taps into dispositional features. Consequently, automation, which is judged by its situational performance, was perceived as more reliable than human advisers, regardless of pedigree. The reasons for these differences between the effects of situational reliability and dispositional credibility, however, are conjecture and need to be examined empirically, which forms the premise for the luggage-screening experiment described in Study 2.

The results of Study 1 indicate that subjective biases and preconceived notions of human versus automated advisers exert a strong influence on operators' perceptions of advice from either source. The question that arises is whether these subjective biases translate into operators' decisions to accept or reject information from human and automated advisers, or whether they are tempered or mediated by the actual accuracy of advice. Furthermore, the results suggested the need for a consolidated experimental paradigm that (a) provides participants dispositional information about their advisers and (b) allows users to visually observe the

performance of their advisers during the task. Such an experimental design is described in Study 2.

STUDY 2: ASSESSING USE OF ADVICE AND POST HOC TRUST

In the literature on human interaction with decision aids, there exists an entire range of tasks, from simple generic target-detection tasks (e.g., looking for numbers among letters; Madhavan, Wiegmann, & Lacson, 2006), to moderately complex paper-and-pencil decision-making tasks (Dijkstra, 1999; Lerch et al., 1997), to tasks involving highly complex automation (Lee & Moray, 1994; Parasuraman & Riley, 1997). These tasks represent a continuum of contexts (from the very simple to the very complex) in which human cognition and decision making can be assisted by implementing DSSs of various levels of sophistication.

The task we present in Study 2 is a binary detection task. The DSSs performed the functions of Stage 2 (diagnostic) automation, in which they provided the participant with a consolidated diagnosis of the situation (presence or absence of a target). Participants then used this information to generate a decision to stop or pass a piece of passenger luggage in an airport security task.

In this study we examined whether (a) the initial subjective biases (demonstrated by participants in Study 1) can logically relate to the objective use of advice by another set of participants and (b) the a priori opinions of participants in Study 1 differ significantly from post hoc opinions of the DSSs generated by participants in Study 2.

Method

Participants. A total of 180 (72 men, 108 women; mean age = 20.5 years) undergraduate and graduate students were paid \$8 for 1 hr of participation.

Procedure. Participants completed 200 trials of a computer simulation task that required them to detect the presence of a hidden weapon embedded in various types of airline passenger luggage depicted on a screen. The stimuli consisted of two-color x-ray images of luggage, densely cluttered with a variety of everyday objects (e.g., clothes, hair dryers, pill bottles). Participants received assistance from an adviser that represented the diagnoses of either an automated system or a human participant, in the form of a text message on the screen. In reality, all participants received assis-

tance from the same computer program. They believed the source was different, based on one of the descriptions of advisers (shown in Appendix A) that they were provided prior to their beginning the task. Participants input their own decision after receiving the adviser's diagnosis.

The probability of the adviser generating a hit/correct rejection or a miss/false alarm was either .70 and .30 (low-reliability condition) or .90 and .10 (high-reliability condition), respectively. The criterion setting (beta) for both aids was 1 – that is, the aids had an equal probability of generating hits and correct rejections. No participants were given any information about the true reliability of the adviser. The information source (human or automated), pedigree (novice or expert), and reliability (low or high) of the adviser varied as per experimental group. In addition to the eight experimental groups, an additional group ($n = 20$) performed the task unaided. At the end of each trial participants received feedback as to whether they had generated a hit, miss, false alarm, or correct rejection. After the 200 trials, participants estimated their trust in and perceived reliability of advisers on postexperimental questionnaires.

Hypotheses. We expected situational factors to mediate the effects of dispositional features in this study. When adviser reliability was high (90%), we expected agreement with automated aids to be higher than with human advisers, as the near-perfect accuracy rate of advice will maintain users' perfect automation schema. When advisers were 70% reliable, we expected agreement with automation to be higher than with humans when advisers were portrayed as novices. When they were portrayed as experts, we predicted agreement with the automated expert to drop below that with the human because errors generated by automation (with 30% frequency) are more likely to be noticed by users because of the disruption of the perfect automation schema (Dzindolet et al., 2001), leading to a rapid breakdown in agreement.

Results

Given the between-groups design of this study, we analyzed the data using mixed two-tailed ANOVAs. Similar to Study 1, these were followed by post hoc comparisons, with alpha values less than .05 being reported as significant. Again, we used Cohen's d as an index of effect size.

Advice acceptance. Data for computing agreement with the adviser (for aided groups) were

grouped into five blocks of 40 trials each, so as to trace any changes during the course of the task. Compliance was calculated as the conditional probability of the participant saying “target present” given that the adviser said “target present,” or $p(\text{Oy}|\text{Ay})$ (Oy = operator says “yes”; Ay = adviser says “yes”). Reliance was calculated as the conditional probability of the participant saying “target absent” given that the adviser said “target absent,” or $p(\text{On}|\text{An})$ (On = operator says “no”; An = adviser says “no”).

A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 2 (reliability: 90% vs. 70% reliable adviser) \times 2 (response type: compliance vs. reliance) \times 5 (trial block) mixed ANOVA on agreement with advice revealed significant main effects for reliability, $F(1, 152) = 35.52, p = .004$, response type, $F(1, 152) = 152.73, p = .0001$, and trial block, $F(4, 608) = 6.79, p = .0051$. Results also revealed the following interactions: reliability and response type, $F(1, 152) = 5.59, p = .006$; reliability and trial block, $F(4, 608) = 6.79, p = .0072$; pedigree and trial block, $F(4, 608) = 4.09, p = .0081$; pedigree, reliability, and trial block, $F(4, 608) = 3.90, p = .009$; and a five-way interaction among source, pedigree, reliability, response type, and trial block, $F(4, 608) = 3.18, p = .0064$.

Overall, agreement with 90% reliable advisers ($M = .51, SD = .11$) was consistently higher than with 70% reliable advisers ($M = .41, SD = .10$). Therefore, we analyzed the data for compliance

and reliance within reliability levels. Analysis of agreement rates of participants receiving 90% reliable advice did not reveal significant effects of source or pedigree. Therefore, we present data only for participants receiving 70% reliable advice.

For compliance strategies, a 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 5 (trial block) mixed ANOVA with 70% reliable advisers revealed no significant main effects. However, results revealed significant interactions between source and trial block, $F(4, 304) = 2.73, p = .043$, pedigree and trial block, $F(4, 304) = 3.69, p = .003$, and, source, pedigree, and block, $F(4, 304) = 4.31, p = .0072$. Participants' compliance strategies are illustrated in Figure 1. Participants using expert advisers demonstrated an initial pattern of compliance that contradicted hypotheses.

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 below that with the expert human adviser ($M = .58, SD = .21$), $t(38) = 2.70, p = .035, d = 0.88$. On the other hand, when advisers were portrayed as novices, compliance with the automated adviser ($M = .56, SD = .21$) was higher than that with the human adviser ($M = .52, SD = .20$), $t(38) = 1.43, p = .067, d = 0.46$, throughout the task, although this difference did not reach statistical significance.

For reliance strategies, a 2 (source: human vs.

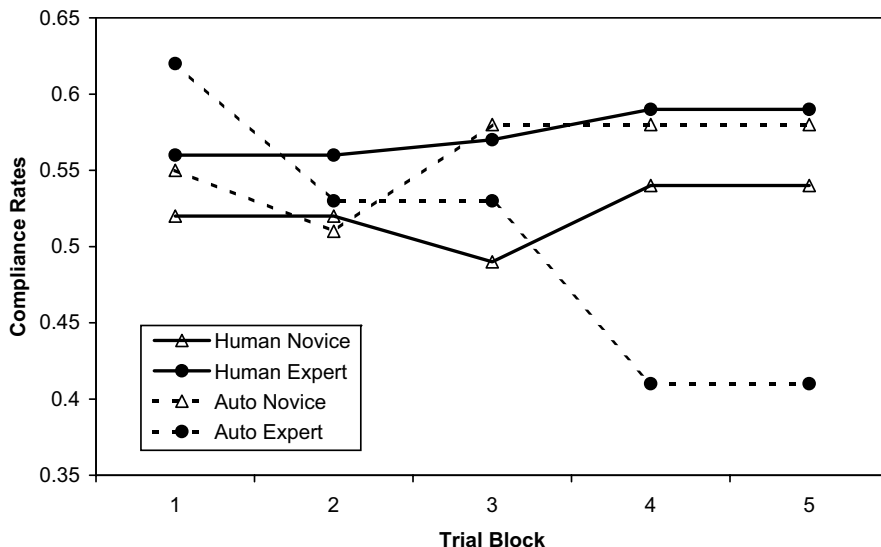


Figure 1. Compliance rates of participants receiving 70% reliable advice.

automated adviser) \times 2 (pedigree: expert vs. novice) \times 5 (trial block) mixed ANOVA revealed a significant main effect for trial block, $F(4, 304) = 4.33, p = .0033$, a significant interaction between pedigree and trial block, $F(4, 304) = 1.63, p = .025$, and a three-way interaction among source, pedigree, and block, $F(4, 304) = 1.88, p = .035$. Similar to the results for compliance strategies, and as depicted in Figure 2, there were no significant differences in reliance between expert human and automated advisers until the last trial block. However, reliance on advice in the fifth trial block for participants using the expert automated adviser ($M = .81, SD = .10$) was significantly lower than for those using the expert human adviser ($M = .84, SD = .12$), $t(38) = 1.66, p = .051, d = 0.54$.

Again, when advisers were portrayed as novices, reliance on the automated adviser ($M = .85, SD = .099$) was generally higher than that on the human adviser ($M = .78, SD = .14$) throughout the task. However, this difference did not reach statistical significance, $t(38) = 1.86, p = .07, d = 0.60$.

For sensitivity (d'), all aided participants were significantly more sensitive (90% reliable group: $M = 1.58, SD = 0.57$; 70% reliable group: $M = 0.68, SD = 0.35$) than unaided participants ($M = 0.38, SD = 0.38$), 70% reliable group, $t(98) = 3.43, p = .001, d = 0.69$; 90% reliable group, $t(98) = 8.96, p = .0035, d = 1.81$. Participants receiving 90% reliable advice performed significantly better (hit rate: $M = .70, SD = .18$; correct rejection rate: $M = .83, SD =$

.073) than unaided participants (hit rate: $M = .34, SD = .16$), $t(98) = 8.05, p = .0056, d = 1.63$ (correct rejection rate: $M = .78, SD = .14$), $t(98) = 2.5, p = .045, d = 0.51$. However, participants receiving advice from 70% reliable advisers (hit rate: $M = .51, SD = .14$; correct rejection rate: $M = .73, SD = .11$) performed better than unaided participants only on target trials, $t(158) = 7.28, p = .0071, d = 1.16$, but not on nontarget trials, $t(98) = 1.57, p = .12, d = 0.32$.

A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 2 (reliability: 90% vs. 70% reliable adviser) \times 5 (trial block) ANOVA on sensitivities of aided participants revealed a significant main effect only for reliability, $F(1, 152) = 143.14, p = .0001$. As expected, participants receiving advice from 90% reliable advisers were more sensitive than those using 70% reliable advisers, $t(158) = 12.43, p = .006, d = 1.98$. As a result, participants receiving 90% reliable advice performed the task at a significantly higher level of accuracy than participants receiving 70% reliable advice; hit rate: $t(98) = 4.65, p = .0092, d = 0.94$; correct rejection rate: $t(98) = 6.82, p = .0071, d = 1.09$. The lack of significant source and pedigree effects on sensitivities suggests that performance differences among aided groups that go beyond broad reliability differences were primarily attributable to shifts in decision criteria or beta.

Criterion settings (beta) and accuracy. The criterion setting of unaided participants ($M = 1.38, SD = 0.44$) did not differ significantly from that of

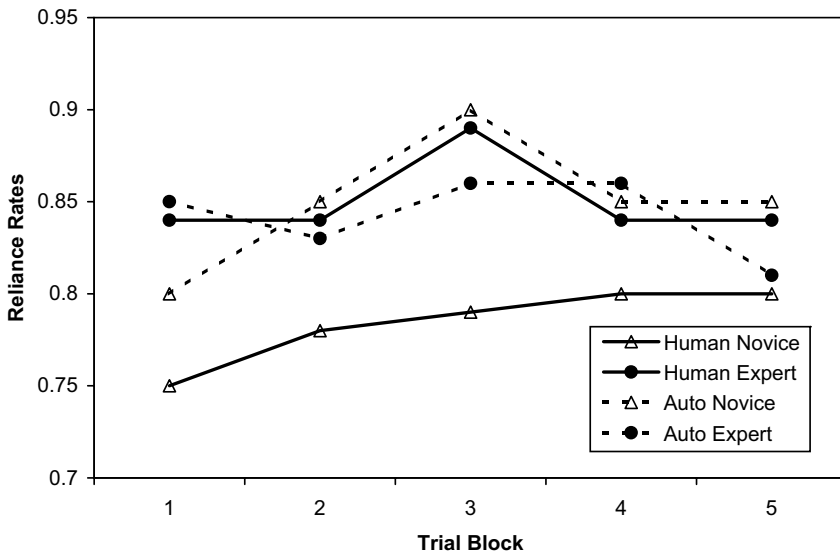


Figure 2. Reliance rates of participants receiving 70% reliable advice.

aided participants ($M = 1.52$, $SD = 1.00$), $t(98) = 0.36$, $p = .72$, $d = 0.07$. A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 2 (reliability: 90% vs. 70% reliable adviser) \times 5 (trial block) ANOVA on criterion settings revealed a significant main effect for pedigree, $F(1, 152) = 2.28$, $p = .021$, and a four-way interaction among source, pedigree, reliability, and trial block, $F(4, 608) = 1.19$, $p = .037$. The criterion settings of participants utilizing 90% reliable advice were not affected significantly by the pedigree or source of advice. Therefore, we present the data for those receiving 70% reliable advice alone.

Given the ratio of noise (.80) to signal (.20) trials, the optimal beta setting for participants was 4. A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 5 (trial block) ANOVA on criterion settings of participants utilizing 70% reliable advisers revealed a significant main effect for pedigree, $F(1, 76) = 1.41$, $p = .035$, and a significant interaction among source, pedigree, and trial block, $F(4, 304) = 1.20$, $p = .042$. As indicated by the main effect for pedigree, criterion settings of participants receiving expert advice ($M = 1.77$, $SD = 0.34$) were higher and more conservative than those of participants receiving novice advice ($M = 1.48$, $SD = 0.42$), $t(38) = 1.77$, $p = .035$, $d = 0.61$, across all trial blocks for both human and automated advisers.

When advisers were human, a conservative criterion setting led participants receiving expert ad-

vice to generate more target-absent responses and, consequently, fewer hits ($M = .48$, $SD = .15$) and fewer false alarms ($M = .26$, $SD = .11$) across trial blocks than did those using novice human advisers (hits: $M = .53$, $SD = .13$), $t(38) = 1.90$, $p = .034$, $d = 0.62$ (false alarms: $M = .29$, $SD = .14$), $t(38) = 1.82$, $p = .033$, $d = 0.59$, who demonstrated a relatively smaller shift from the adviser's beta. Likewise, with automated advisers, a liberal criterion setting led participants using novice advisers to generate more target-present responses and, consequently, more hits ($M = .54$, $SD = .13$), $t(38) = 1.99$, $p = .023$, $d = 0.62$, and slightly more false alarms ($M = .29$, $SD = .14$), $t(38) = 1.82$, $p = .09$, $d = 0.59$, than did participants receiving expert advice, although the latter difference did not approach statistical significance.

As illustrated in Figure 3, participants using human advisers demonstrated no changes in their pattern of criterion settings across trial blocks at both pedigree levels. On the contrary, criterion settings of participants using automated advisers demonstrated different patterns of change over the course of the task as a function of pedigree.

Participants using expert automated advisers demonstrated a significant upward shift in criterion settings (away from the aid's neutral beta, toward optimal beta) over the course of the experiment (first trial block: $M = 1.73$, $SD = 0.33$; fifth trial block: $M = 1.82$, $SD = 0.40$), $t(19) = 1.99$, $p = .041$, $d = 0.91$. On the other hand, participants receiving novice automated advice demonstrated a

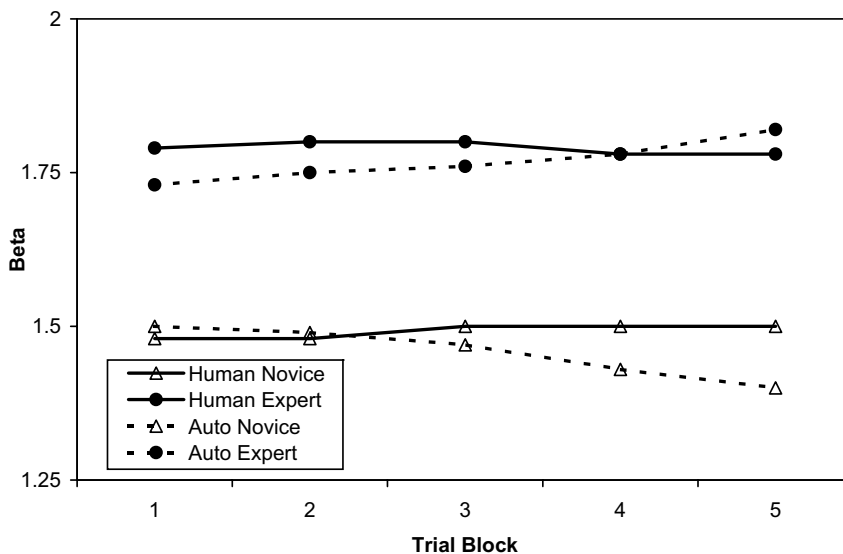


Figure 3. Criterion settings of participants receiving 70% reliable advice.

significant downward shift in criterion settings (toward the adviser's neutral beta, away from optimal beta) over the five trial blocks (first trial block: $M = 1.5$, $SD = .26$; fifth trial block: $M = 1.4$, $SD = 0.27$), $t(19) = 1.21$, $p = .045$, $d = 0.56$. Consequently, the differences in beta for the expert and novice automated groups generally got larger over blocks, whereas the differences in beta for those receiving advice from either expert or novice human advisers stayed the same.

Trust ratings. A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 2 (reliability: high vs. low) mixed ANOVA on trust scores revealed significant main effects for source, $F(1, 152) = 3.16$, $p = .0001$, pedigree, $F(1, 152) = 3.33$, $p = .004$, and reliability, $F(1, 152) = 43.23$, $p = .0064$. In keeping with hypotheses, trust in 90% reliable advisers ($M = 7.12$, $SD = 1.28$) was significantly higher than that in 70% reliable advisers ($M = 5.71$, $SD = 1.43$), $t(158) = 6.5$, $p = .003$, $d = 1.03$. Trust in automated advisers ($M = 6.61$, $SD = 1.41$) was always higher than that in human advisers ($M = 6.22$, $SD = 1.62$), $t(158) = 1.58$, $p = .038$, $d = 0.25$, and expert advisers were trusted ($M = 6.60$, $SD = 1.50$) significantly more than novice advisers ($M = 6.20$, $SD = 1.53$), $t(158) = 1.62$, $p = .0063$, $d = 0.26$. Contrary to the pattern observed in Study 1, there were no significant interactions between source and pedigree in the present study.

Perceived reliability ratings. A 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) \times 2 (reliability: high vs. low) \times 2 (response type: hit rate, correct rejection rate) mixed ANOVA on participants' estimates of hit rates and correct rejection rates revealed a significant main effect only for reliability, $F(1, 152) = 82.76$, $p = .002$, and nonsignificant interactions between pedigree and reliability, $F(1, 152) = 3.38$, $p = .068$, and among source, pedigree, and reliability, $F(1, 152) = 3.77$, $p = .054$. Similar to trust ratings, 90% reliable advisers ($M = 82.5\%$, $SD = 11.76\%$) were perceived as more accurate than those that were 70% reliable ($M = 62.55\%$, $SD = 15.45\%$), $t(158) = 9.18$, $p = .0042$, $d = 1.46$.

Given the similarity in perceived hit rates and correct rejection rates, we grouped these estimates into a consolidated perceived reliability rating. Similar to the data for compliance and reliance, we present the data for perceived reliability grouped within reliability levels.

For 90% reliable advisers, a 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs.

novice) ANOVA on perceived reliability estimates of participants revealed a significant main effect for pedigree alone, $F(1, 76) = 1.32$, $p = .046$. Expert advisers ($M = 84\%$, $SD = 10.64\%$) were perceived as more reliable than novice advisers ($M = 78.3\%$, $SD = 15.76\%$), $t(38) = 1.39$, $p = .033$, $d = 0.45$, for both human and automated advisers. All participants utilizing 90% reliable advice underestimated the actual reliability of their advisers. However, one-sample t tests comparing participants' estimates with actual reliability revealed that participants using expert advisers were better calibrated in their assessments (mean difference = 4.75), $t(19) = 2.54$, $p = .042$, $d = 1.17$, than those using novice advisers (mean difference = 9.85), $t(19) = 2.65$, $p = .036$, $d = 1.22$.

For 70% reliable advisers, a 2 (source: human vs. automated adviser) \times 2 (pedigree: expert vs. novice) ANOVA on perceived reliability estimates of participants revealed no significant main effects but did show a significant interaction between source and pedigree, $F(1, 76) = 1.08$, $p = .033$. For human advisers, perceived reliability of the expert ($M = 67.3\%$, $SD = 15.25\%$) was significantly higher than that of the novice ($M = 59.9\%$, $SD = 19.35\%$), $t(38) = 1.54$, $p = .025$, $d = 0.50$. However, for automated advisers, perceived reliability of the novice ($M = 67.6\%$, $SD = 16.25\%$) was higher than that of the expert ($M = 60.07\%$, $SD = 19.1\%$), $t(38) = 0.82$, $p = .42$, $d = 0.27$, albeit nonsignificantly. Almost all participants were well calibrated in their perceptions of the adviser reliability, except those using novice human advisers, who significantly underestimated reliability (mean difference = 10.1), $t(19) = 2.34$, $p = .017$, $d = 1.07$.

Discussion

The results of Study 1 revealed that users hold preconceived biases or notions about human and automated advisers even prior to interacting with these advisers. The purpose of Study 2, therefore, was to examine the effects of these different sources of diagnostic advice on the efficacy of decision making in an actual task environment.

Behavioral dependence on advisers. When advice was 90% reliable, the high situational reliability of advice compensated for the effects of source and pedigree, suggesting that near-perfect adviser accuracy appears to have a stronger psychological impact on advice use than does dispositional information about the source of advice. However, not all real-world DSSs are likely to

function at such high levels of reliability, so we examined how users respond to DSSs of a lower reliability level.

When advice was 70% reliable, results suggested that merely portraying the human adviser as an expert led users to agree more with the advice, regardless of accuracy. However, participants receiving 70% reliable advice from automated advisers demonstrated a pattern of dependence that reflects a breakdown in the perfect automation schema. When automated aids were portrayed as experts, initial expectations were likely very high (as revealed in Study 1). However, when participants observed their adviser generating errors on 30% of occasions, it likely led to a rapid breakdown in initial expectations, thereby generating a negative trend in dependence.

Decision criterion settings. When advice was 70% reliable, there were differences in the degree of shift away from or toward neutral beta as a function of adviser pedigree. Applying the results of Study 1 to the choice of dependence strategies, 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 the adviser missed a target. Unlike participants receiving advice from novices, those 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, because the initial high expectations of participants using expert advisers were violated by the low accuracy level of 70% reliable expert advisers, they presumably developed their own beta by shifting away from the adviser's neutral bias (toward optimal). Conversely, the data suggest 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 it was for those using experts.

Subjective trust and perceived reliability. As indicated by the results of Study 1, initial trust in and expectations about novices are significantly lower than those for experts. According to Madhavan

et al. (2006), the perceived reliability of advice is affected strongly when operators observe advisers generating errors that appear easy to the operator (the "easy errors hypothesis"), regardless of the actual accuracy of advisers.

In the present study, participants using novice advisers were likely more sensitive to "easy" adviser errors (e.g., instances when the adviser missed a knife and the participant easily saw the knife) because of their prematurely low expectations of novice advisers. Such expectancy-driven judgments possibly led to a greater degree of miscalibration of reliability, relative to those receiving expert advice. Conversely, high initial expectations of participants receiving expert advice (also observed in Study 1) were likely maintained by the 90% near-perfect performance of expert advisers, leading to better calibration of reliability.

This potential interaction of source and pedigree on perceived reliability (when advice was 70% reliable) contradicts the pattern observed in Study 1, wherein subjective trust was more strongly affected by the interaction than was perceived reliability. This suggests that preconceived biases about humans and machines (measured in Study 1) influence operators' perceptions of human versus automated advisers in a manner different from biases that develop as a consequence of operators' interaction with advisers (measured in Study 2).

The opportunity to observe the situational reliability of the adviser in Study 2 possibly led to participants basing their subjective opinions more on situational factors than on dispositional factors, ultimately leading to significant effects on perceived reliability and not on subjective trust. These results support our premise that subjective trust and perceived reliability tap into different cognitive processes of users. The implications of these results for the development of a model of trust in humans and machines, and potential suggestions to the designers of DSSs, are discussed next.

GENERAL DISCUSSION

The results of the present studies support the model of trust in human advisers versus automated aids developed recently by Madhavan and Wiegmann (2007). According to this model, trust in automated systems develops in a manner akin to trust among humans; however, there are critical differences in the manner in which people react to automated advice versus human advice, which are

particularly salient when advisers begin generating errors. This model suggests that calibration of user trust in a DSS can be improved by attempting to bridge the gap between human perceptions of humans versus machines by incorporating “humanlike” characteristics into the design of automated systems. Such anthropomorphizing will eventually lead to comparable levels of trust among human and automated DSSs with varying pedigree and reliability levels.

However, the present studies point to the primary weakness in this reasoning because operators are typically prone to preconceived notions and biases about humans versus automation. These a priori biases continue to influence advice acceptance when users observe the situational accuracy of these advisers, even when automated aids behave in a manner identical to human advisers. Therefore, associating humanlike characteristics with machines may not, in most circumstances, necessarily help equate human perceptions of automation to perceptions of other humans.

Perhaps providing operators with a rationale as to why automation might differ from a human might assist the calibration of trust more effectively than trying to make automation appear similar to a human. Providing operators the opportunity to understand the functions and limitations of a DSS will help operators develop a clearer model of situations when a machine’s algorithms might fail, providing the operator opportunities to cognitively compensate for machine errors. This debate between the psychological effects of anthropomorphizing of machines versus the benefits of increasing the salience of human and automation differences is an issue that requires further exploration.

Conclusions

The present experimental paradigms provide data on the manner of trust development in systems and humans. However, there are certain limitations. The extent to which real-world operators will demonstrate the observed biases and response tendencies is likely to be a function of several factors, including the operator’s training and expertise, cultural factors that determine the user’s degree of “comfort” with the DSS, and the consequences associated with correct and incorrect decisions (Dzindolet et al., 2002).

Study 2 used a relatively contrived situation in which there were no specific consequences for wrong decisions. Such a situation is not compara-

ble to that faced by a physician who cannot afford to misdiagnose an illness by incorrectly utilizing information from an automated aid or by a luggage screener who could face fatal consequences by failing to diagnose a weapon. The base rate of target stimuli in Study 2 was higher than would be the case in a real luggage-screening task, and participants received feedback after each trial, which is more frequent than feedback mechanisms in the real world.

Despite these limitations, the current studies provide valuable information for the development of a conceptual framework of DSS trust and use in complex environments. A potential suggestion to the designers of DSSs is to design systems that elicit appropriate compliance or reliance strategies so as to reduce the occurrence of a costly miss or a false alarm. A more challenging suggestion is to arrive at the optimal reliability level that would provide positive assistance while minimizing the negative complacency effects associated with “overly reliable” automation.

As computer-based decision aids are increasingly being incorporated in organizations, more critical decisions are being influenced by computer systems. From the results of both studies, it follows that internal attributes of a DSS, when combined with extraneous information on the adviser’s apparent pedigree, evidently has a stronger effect on DSS trust and dependence than does the objective reliability of advice in itself. Thus, it is to the exploration and manipulation of the internal attributes of these systems that attention must be paid, in order to generate concrete recommendations for the successful design of future DSSs and to ensure the seamless flow of communication between systems and humans.

APPENDIX

The Novice Human Adviser

“Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from a novice student named BILL JOHNSON. BILL JOHNSON is currently an undergraduate student and has no prior experience in real world luggage screening tasks. He is currently majoring in criminology at a small technical college in the Midwest and is interested in specializing in antiterrorism and airport security. However, BILL is still a novice when it comes to modern terrorist tactics. He possesses

limited knowledge of the types of modern weapons and explosives commonly smuggled aboard aircraft. He also has a partial understanding of the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. BILL has recently applied for an internship at the Transportation Safety Administration (TSA) to help oversee security operations.”

The Expert Human Adviser

“Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from an expert in airport security named DR. BILL JOHNSON. DR. BILL JOHNSON was originally trained as a luggage screener, serving 10 years in some of the busiest airports in the United States. He went on to earn his Ph.D. in criminology from the Massachusetts Institute of Technology (MIT), specializing in antiterrorism and airport security. DR. BILL JOHNSON is an expert when it comes to modern terrorists’ tactics. He possesses extensive knowledge of the types of modern weapons and explosives commonly smuggled aboard aircraft. He also has a keen understanding of the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. DR. BILL JOHNSON has recently been appointed by the Transportation Security Administration (TSA) to oversee security operations at Chicago’s O’Hare International Airport, which is one of the largest airports in the world.”

The Novice Automated Adviser

“Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from a novice computer system called DETECTOR, which is an automated diagnostic aid that has been designed to identify hidden contraband in airline passenger luggage. DETECTOR is based upon the technology traditionally used at major airport security checkpoints over the past 10 years. DETECTOR was designed and developed at a small technical college in the Midwest, which contains a recently established department in antiterrorism and airport security. It currently possesses a limited database of the types of modern weapons and explosives commonly smuggled aboard aircraft. Its algorithms are relatively unsophisticated in their attempts to capture the tactics and strategies used by terrorists to conceal weapons and explosives in-

side luggage. The Transportation Security Administration (TSA) is considering whether to conduct a limited field test of DETECTOR at a small airport in the hope of making it employable at larger airports to enhance security operations in the future.”

The Expert Automated Adviser

“Imagine you are going to perform a difficult luggage inspection task at a busy airport. You have the option of receiving assistance from an expert computer system called SUPER-DETECTOR, which is an automated diagnostic aid that has been programmed to identify hidden contraband in airline passenger luggage. SUPER-DETECTOR is based upon, yet far exceeds, the technology traditionally used at major airport security checkpoints over the past 10 years. SUPER-DETECTOR was designed and developed at the Massachusetts Institute of Technology (MIT), which contains a highly specialized department in antiterrorism and airport security. It possesses an extensive database of the types of modern weapons and explosives commonly smuggled aboard aircraft. Its algorithms are highly sophisticated and effectively capture the tactics and strategies used by terrorists to conceal weapons and explosives inside luggage. SUPER-DETECTOR has recently been employed by the Transportation Security Administration (TSA) to enhance security operations at Chicago’s O’Hare International Airport, which is one of the largest airports in the world.”

REFERENCES

- Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40*, 672–680.
- Bowers, C. A., Oser, R. A., Salas, E., & Cannon-Bowers, J. A. (1996). Team performance in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 243–263). Mahwah, NJ: Erlbaum.
- Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology, 18*, 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*, 147–164.
- Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*, 53–71.
- 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.
- Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2003, August). *Effects of instructional framing on automation trust and reliance*. Paper presented at the 111th Annual Meeting of the American Psychological Association, Toronto, Canada.

- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Lee, J. D., & See, S. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 50-80.
- Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). The Turing effect: The nature of trust in expert system advice. In P. J. Feltovich & K. M. Ford (Eds.), *Expertise in context: Human and machine* (pp. 417-448). Cambridge, MA: MIT Press.
- Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6, 104-123.
- Madhavan, P., & Wiegmann, D. A. (2005). Cognitive anchoring on self-generated decisions reduces operator reliance on automated diagnostic aids. *Human Factors*, 47, 332-341.
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8, 277-301.
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48, 241-256.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 4, 196-204.
- Muir, B. M., & Moray, N. (1996). Trust in automation: II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39, 429-460.
- Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45, 669-678.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56, 81-103.
- Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43, 223-239.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Stanford, CA: Cambridge University Press.
- Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49, 95-112.
- Safar, J. A., & Turner, C. W. (2005). Validation of a two-factor structure for system trust. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 497-501). Santa Monica, CA: Human Factors and Ergonomics Society.
- Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. New York: Wiley/Human Factors and Ergonomics Society.
- Sundar, S. S., & Nass, C. (2000). Source orientation in human computer interaction: Programmer, networker or independent social actor? *Communications Research*, 27, 683-703.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.

Poornima Madhavan is an assistant professor of human factors in the Department of Psychology at Old Dominion University, where she also holds an appointment at the Virginia Modeling, Analysis and Simulation Center. She received her Ph.D. in engineering psychology (human factors) from the University of Illinois at Urbana-Champaign in 2005.

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

Date received: November 4, 2005

Date accepted: December 10, 2006