Theoretical Issues in Ergonomics Science
Vol. 8, No. 4, July–August 2007, 277–301

Taylor & Francis
Taylor & Francis Group

Similarities and differences between human–human and human–automation trust: an integrative review

P. MADHAVAN*† and D. A. WIEGMANN‡

†Carnegie Mellon University, Porter Hall 208-J, Pittsburgh, PA 15213, USA
‡University of Illinois at Urbana-Champaign, Champaign, IL, USA

The trust placed in diagnostic aids by the human operator is a critical psychological factor that influences operator reliance on automation. Studies examining the nature of human interaction with automation have revealed that users have a propensity to apply norms of human–human inter-personal interaction to their interaction with 'intelligent machines'. Nevertheless, there exist subtle differences in the manner in which humans perceive and react to automated aids compared to human team-mates. In the present paper, the concept of trust in human–automation dyads is compared and contrasted with that of human–human dyads. A theoretical framework that synthesizes and describes the process of trust development in humans vs automated aids is proposed and implications for the design of decision aids are provided. Potential implications of this research include the improved design of decision support systems by incorporating features into automated aids that elicit operator responses mirroring responses in human–human inter-personal interaction. Such interventions will likely facilitate better quantification and prediction of human responses to automation, while improving the quality of human interaction with non-human team-mates.

Keywords: Decision aids; Automation; Trust; Reliance

1. 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 an active team-mate sharing control with automation. Specifically, automated decision aids are increasingly being modelled as 'partners' rather than as tools (Klein *et al*. 2004). These 'partners' support or assist the human in performing functions that may either be difficult or even impossible for the operator to perform without the assistance of a 'knowledgeable team-mate'.

The Transportation Safety Administration (TSA) is currently exploring the efficacy of implementing such automated aids for assisting luggage screeners in

*Corresponding author. Email: madhavan@andrew.cmu.edu

Theoretical Issues in Ergonomics Science
ISSN 1463–922X print/ISSN 1464–536X online
http://www.tandf.co.uk/journals
DOI: 10.1080/14639220500337708

detecting 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 in detecting the presence of tumours and other anomalies in patient x-rays (Krupinski *et al*. 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 team-mate. Indeed, some researchers have argued that such human–automation teams function similarly to human–human teams (Bowers *et al*. 1996, 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 *et al*. 1996, Reeves and Nass 1996). This is supported by the 'Computers Are Social Actors' or CASA studies (Nass *et al*. 1993, 1994, 1995, Nass and 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 rather than as a manifestation of their human creators (Sundar and Nass 2000). 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 (Nass and Lee 2001). Furthermore, researchers contend that strong social bonds between humans and computers can be created when a computer is labelled a 'team-mate' (Nass *et al*. 1999, Cassell and Bickmore 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 (Dijkstra 1999). The concept of automation trust, in general, has been the focus of a vast body of research over the last decade (Muir 1987, 1994, Lee and Moray 1992, 1994, Blomqvist 1997, Lerch *et al*. 1997, Lee and See 2004). However, existing literary reviews have addressed automation trust as a largely global construct (Blomqvist 1997, Lewandowsky *et al*. 2000, Lee and See 2004) and research portraying human–automation trust as a mirror of human–human inter-personal trust in situations demanding 'team co-ordination' has yet to be reviewed. Therefore, the present paper is focused on studies that specifically address the development of trust in dyadic decision teams that constitute a primary decision maker and either a human or an automated 'advisor'.

We begin with a brief definition of automation and decision support systems followed by a general discussion of trust. We then provide a detailed analysis of three characteristics of advice, namely, source of information, pedigree and reliability that affect operator trust in and utilization of advice. This is followed by two integrative frameworks that compare trust in human–automation teams with that of human–human teams. We conclude with a discussion of other factors (besides the three mentioned above) that are likely to influence trust in advisors

and the potential implications of this article for bridging the gap between human trust in humans vs trust in automation.

2. Automation and decision support systems

Automation refers to the full or partial replacement of a function previously carried out by a human operator (Parasuraman *et al.* 2000). This implies that a task can be selectively automated, with the degree or level of automated decision aiding varying according to the nature of the task being performed. A decision support system (DSS) is a system that supports technological and managerial decision-making by assisting in the organization of knowledge (Sage 1987). DSSs and automated diagnostic aids often assist human operators in several critical decision-making tasks, where little raw data about system states is available to the human (i.e. system functions are either opaque or unclear to the human operator). DSSs offer the advantages of increased efficiency, better data monitoring and analysing capabilities and higher processing speed than human decision makers and, although not infallible, offer the potential to enhance diagnostic decision-making by circumventing the limited information processing capabilities of human operators (Mosier and Skitka 1996). In short, DSSs help supplement or elucidate the information already available to the user by highlighting relevant areas, providing suggestions, recommending courses of action or in some cases even performing the action for the human operator. The precise degree of decision aiding can vary along a continuum, with the lowest level being characterized by complete manual control and the highest level being characterized by fully automated control. Several levels between these two extremes have been proposed (Sheridan and Verplank 1978, Riley 1989, Endsley and Kaber 1999).

Sheridan and Verplank (1978) derived a 10-point scale of diagnostic aiding, with higher levels representing increased autonomy of machine over human action. For example, at low level 2, the machine provides several suggestions to the human but the system does not influence the final decision in any further way. At level 4, the computer suggests one alternative, but the human has the power to veto the system's suggestion in favour of any other option. At higher levels, the human has limited time or power to veto the system's choice before the action is carried out by the system. While automated decision aiding at any of the mentioned 10 levels is invaluable to the human operator, the degree to which humans reportedly trust in and rely on these decision aids may vary with the degree of automated support being provided to the human.

Ideally, the combination of human decision maker and automated decision aid should result in a high-performing team that succeeds in circumventing the errors normally made by the human decision maker alone (Mosier and Skitka 1996). However, in reality, the facets of the human–machine decision-making system are as complex as the environments in which they function (Mosier and Skitka 1996). If decision aiding systems are to be created such that they truly enhance the quality of human–machine joint performance, it is essential that the design of such automated support systems be based on a thorough analysis of human cognition and decision-making processes. Therefore, at a time when the role of automation in complex systems is being given increasing importance, numerous researchers are re-emphasizing the concerns expressed by Sheridan and Ferrell (1974) decades ago,

which is the fact that operator *trust* in automation is a fundamental element governing human–automation team performance. Thus, trust in automation continues to be an issue of concern to human factors researchers in the present day.

3. The nature of trust (general)

Trust refers to the expectation of, or confidence in, another and is based on the probability that one party attaches to co-operative or favourable behaviour by other parties (Barber 1983, Muir 1987, Hwang and Buergers 1997). According to Couch and Jones (1997), trust has been defined in diverse ways: as a generalized expectancy (Rotter 1967), as an enduring attitude or trait (Deutsch 1958, Giffin 1967) and as a transitory situational variable (Kee and Knox 1970, Driscoll 1978). Rempel *et al.* (1985) identified three coherent dimensions of trust that influence people's acceptance of information provided by an external source: predictability, dependability and faith. According to this model, the most concrete component of trust is based on the *predictability* or consistency of an individual's actions. While this component of trust does not concern the dispositional attributes of an individual, it depends largely on the stability of performance over a period of time. In the context of human interaction with decision aids, operator assessment of the predictability of the latter may affect their tendency to *agree* with it. The second component of trust, *dependability*, is based on the internal dispositional characteristics of a person. This attribute is reflected in the level of *confidence* one has in an advisor. Finally, *faith* is based on beliefs about the future behaviour or accuracy of an information source. This attribute might be reflected in a person's willingness to use a particular aid while performing a new task in the future.

Sheridan (2002) makes a distinction between the different meanings of the term 'trust' in the context of human–automation interaction. Specifically, he distinguishes between trust as an *effect* or outcome of certain automation characteristics (e.g. reliability) and trust as a *cause* of operators' behaviour when utilizing automation. According to Sheridan (2002: 77), 'human operator trust in automation is now a major topic of interest because it significantly affects whether and how automation is used'.

Lee and See (2004) have identified three general bases of trust specifically in the context of automation: performance, process and purpose. *Performance* refers to the current and past operation of automation and primarily encompasses dimensions such as reliability, predictability and ability. *Process* refers to the degree to which the algorithms of automation are appropriate for a situation and describes how the automation operates. The third basis of trust, *purpose*, refers to the extent to which automation functions correspond to the designer's intent and describes why the automation was designed. While Lee and See's first two bases of trust, namely performance and purpose correspond to Rempel *et al.*'s (1985) components of predictability and dependability, the third dimension (purpose) roughly corresponds to the component of faith and benevolence in the Rempel *et al.* model. Wickens *et al.* (2000) have discussed the development of trust in a system given its specific level of predictability (i.e. reliability). Appropriate levels of trust in a DSS require accurate calibration between the individual's perceived reliability of an aid and it's actually reliability. When an operator under-estimates the true reliability of an aid, this can

lead to under-trust. However, when an operator over-estimates the actual reliability of an aid, over-trust occurs.

In relation to the above operator responses to automation reliability, Luhmann (1979, 1988, see also Lewis and Wiegert 1985) introduces the notion of *system trust* that refers to an operator's belief that a system is operating in a predictable manner and in keeping with expectations. Jian *et al*.'s (2000) empirically driven concept of trust in machines vs trust in humans suggests that subtle differences exist in people's assessments of trust vs distrust in human–human relationships compared to human–machine relationships. Specifically, people tend to be less extreme in their ratings of human–human distrust than trust, as an assessment of distrust in humans is perceived as more negative than low or negative trust. This is not true for ratings of human–machine trust vs distrust, suggesting that people are less hesitant to label a machine in terms of distrust compared to a human partner. Since the above research does indicate that trust in humans has inherently different properties than trust in machines, the specific factors affecting trust in different sources of decision support is a matter of considerable interest to researchers in the present day.

4. Factors affecting the development of trust

Given the ubiquitous effect of trust on operator acceptance of information provided by advisors, numerous studies have been conducted over the past several years to identify factors that influence human trust in the advice of others, including that of automated aids. While trust is a valid construct describing human–automation interaction, several fundamental differences exist between the manner of trust development in a human–automation team and a human–human team. In this section, we discuss various features of advisors or DSSs that influence the degree of trust in diagnostic advice and propose three organizing frameworks that address the manner in which these variables interact to influence the efficacy of human–DSS interaction.

As can be seen in figure 1, the accurate perception of the reliability of a DSS is a necessary pre-condition for appropriate trust calibration. Most research in trust development has focused on the human propensity to under-estimate the true reliability of diagnostic aids. However, the development of trust is likely to be strongly influenced by several characteristics of the DSS in itself. Calibration of trust is further influenced by the interaction of DSS features with environmental factors; dynamic environmental states are likely to affect the maximal level of reliability achievable by a DSS (cf. the Lens Model approach; Bisantz *et al*. 2000) and consequently the degree of operator trust in the DSS. For the sake of brevity, we chose to focus the present paper on characteristics of the DSS alone. As illustrated in figure 1, three DSS features in particular, viz (a) the source of diagnostic information (human advisor or automated aid), (b) source reliability and (c) credibility (degree of perceived expertise), draw interesting parallels between human trust in automation vs other humans in a team decision-making environment. Each of these factors will be addressed in the ensuing sections.

4.1. *Source of diagnostic information (human advisor vs automated aid)*

Many psychologists refer to trust as a personal trait (Deutsch 1958, Rotter 1967, 1971, 1980). Social psychologists in particular have argued that trust (in general)

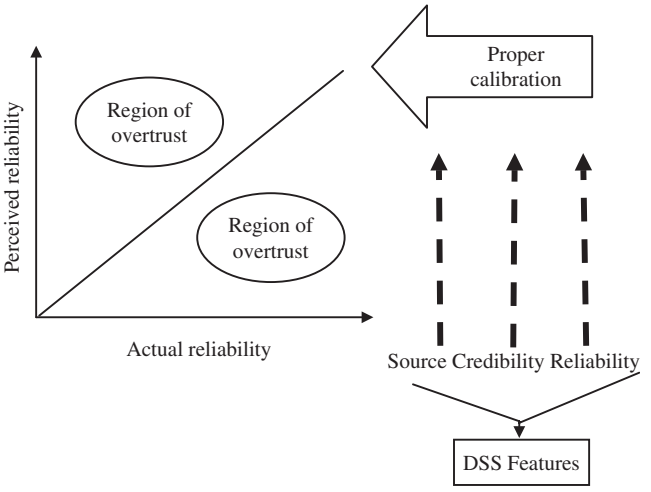

Figure 1. Factors affecting the calibration of user trust in DSSs. Modified from Wickens *et al*. (2000).

is social in nature (Blau 1964) and is partially the result of an attribution process (Reeder and Brewer 1979, Rempel *et al*. 1985). An *attribution* refers to an inference about why an event occurred in terms of a person's dispositions, personality or beliefs. Some early studies in social psychology have demonstrated that interaction with computers can trigger attribution processes. For example, Earley (1988) found that computer-generated data regarding performance was perceived as more trustworthy than supervisor-supplied (i.e. human) feedback.

Along similar lines, a series of recent studies has demonstrated that the social psychological principles governing human–human interaction provide a powerful theoretical framework for the study of human–machine interaction (cf. the CASA studies). For instance, humans are frequently prone to committing the fundamental attribution error while making attributions of responsibility when dealing with machines, in a manner similar to that experienced while interacting with humans. Research has demonstrated an increasing tendency for users to blame technology for mistakes and errors; humans are abdicating more and more responsibility for negative outcomes to machines (Sampson 1986, Morgan 1992, Friedman 1995), while exhibiting reluctance to provide credit for positive outcomes to their non-human partners, thereby demonstrating the fundamental attribution error in human–automation interaction. However, evidence to the contrary suggests that people tend to conceive of human–human relationships differently from human–automation relationships because an assessment of distrust in another human seems more negative than in a non-human entity (Jian *et al*. 2000).

The earliest attempt at examining the trust between human recipients and automated advisors goes back to the research of Turing (1950), who proposed a test to examine people's answers to the question 'Can machines think?' The experiment was aimed at determining whether a human judge could distinguish between a human advisor and a computer imitating a human advisor (with the interaction occurring through a typed medium, thus masking the source). Results revealed that the automated advisor succeeded in correctly 'fooling' the human participant into believing that the advisor was another human and not a machine, 95% of

the time. This early finding that suggested the inability of humans to distinguish between another human and an automated advisor encouraged a series of research attempts focused on examining human trust in and reliance upon automated team members in various simulated real-world contexts. In general, the 'Turing effect' introduced the question of whether human trust in information is different when the advice is attributed to either an automated aid or to another human.

Modern research has begun to examine such issues raised by Turing (1950). For example, Dijkstra *et al*. (1998) examined the manner in which human decision-makers perceived and utilized information provided by a human or automated advisor while performing a decision-making task. The task required participants to make judgements about the appropriateness of solutions generated by either human advisors or expert systems to problems pertaining to law, medicine and secular topics. Participants were told that the 'advisor' was either a human or a computer program depending on condition, while in reality the source of advice was the same for all participants. Results revealed that participants perceived the expert system to be more objective and rational than the human advisor, suggesting that merely informing people that advice comes from an expert system is enough to influence their evaluation of information. This might explain why users sometimes neglect the incompetence of an erroneous expert system, implying that the expert system use can decrease attention paid towards problems and lead to users becoming complacent or passive observers of automation.

A follow-up study by Dijkstra (1999) using a similar experimental paradigm in a legal context revealed that participants often agreed with the incorrect advice of an expert system in spite of alternative and more accurate written information presented to them through human advisors. Almost 80% of participants' answers were in line with the incorrect advice of the expert system and more than half of the participants always agreed with the system. In a previous experiment (Dijkstra 1995) in which participants judged the same criminal law cases as in the later experiment by Dijkstra (1999) *without* receiving information from an expert system or a human advisor, participants managed to attain almost 75% accuracy in judgements unaided. These previous unaided participants clearly out-performed those who used the advice of the expert system in the later experiment. Furthermore, participants who always agreed with information from the expert system evaluated the advice more positively and thought the system to be more useful than participants who agreed less with the system.

A similar series of studies on the perceived utility of human and automated aids in a visual detection task (Dzindolet *et al*. 2002), revealed that humans perceived the utility of an automated decision aid as higher than that of a human aid. College participants performed a target-detection task while being assisted by the decisions reached from a human or an automated system. Before they began performing the actual experiment, participants were asked to rate their own and their aid's expected reliability during the course of the experiment. Results revealed a significant bias toward automation in that the automated aid was perceived as more reliable than the human aid. However, this higher perceived source reliability or credibility did not reflect in objective automation reliance strategies, as participants in the experiments showed a strong tendency toward self-reliance.

One possible reason for the apparent inflated trust in automated systems (Dijkstra *et al*. 1998, Dijkstra 1999) is closely linked to the users' levels of confidence in their own ability to perform the task unaided. Waern and Ramberg (1996)

examined the difference between trust in advisors and self-confidence in a series of experiments in which participants solved problems with the help of a computer and a human advisor sequentially. Results revealed that subjective estimates of trust in the expert system were lower than in human advisors. Ratings of self-confidence revealed that trust in the computer was lower than their self-confidence when participants gave correct answers or believed that the accuracy of their own unaided decisions was high. Conversely, trust in the expert system was higher than self-confidence when participants realized that their unaided decisions were incorrect, confirming that self-confidence in one's own independent abilities is an important predictor of trust in an expert automated advisor. These findings on the relationship between trust and self-confidence were supported by Lee and Moray (1992, 1994), who mathematically described trust based on self-confidence and perceived reliability of automated aids. They empirically demonstrated that when humans have low confidence in their own abilities combined with a high level of perceived reliability of the automated aid, automated aids are more likely to be used.

While the above studies support the notion that people perceive automated systems as more credible sources of information than humans, research has also revealed that people are more sensitive to the errors made by automation, leading to a very rapid weakening of its credibility and a swift decline in trust (Lee and Moray 1992, 1994, Dzindolet *et al*. 2001, Wiegmann *et al*. 2001). According to Dzindolet *et al*. (2001), this rapid drop in trust occurs because humans expect automated aids to perform at near perfect rates, leading them to pay too much attention to errors made by automation.

Cognitive researchers have found that information contrary to a schema (in this case, errors made by automation) is especially likely to be noticed and remembered (Stangor and McMillan 1992). This situation consequently exerts a large influence on task allocation strategies of the human operator. Since humans expect automated aids to out-perform them, this expectation may increase the likelihood for humans to notice and remember the errors made by automation. For example, in the Dzindolet *et al*. (2002) study, participants paired with automated aids recounted in detail instances in which they were highly confident that the aid had made an error in target detection compared to those paired with a human aid.

Although individual differences exist in the expectations people have of automated aid reliability (Lee and Moray 1992, 1994, Singh *et al*. 1993), people on average show a 'positivity bias' (Bruner and Tagiuri 1954, Cacioppo and Bernston 1994) or a tendency to assign positive evaluations to unfamiliar objects rather than negative or neutral evaluations (Cacioppo *et al*. 1997). This positivity bias could also be a reason for human operators having an unrealistically high expectation of automated aid reliability. Consequently, this leads operators to focus very strongly on the errors made by automated aids, as these errors represent a violation of expectations held by the user. This eventually leads to a rapid decline in the perceived reliability of the aid that is a consequence of the aid making errors in the diagnostic process.

On the other hand, human advisors may be perceived as more 'familiar' and consequently lead to decision-makers having more realistic expectations of human than machine advisors. In other words, people do not expect their human partners to be perfect. Therefore, human errors are not easily remembered and perceivers are likely to be more forgiving of incorrect information provided by an imperfect

human advisor rather than by an imperfect automated 'partner'. Perhaps biases in favour of or against automated aids vs humans are a consequence of people making different attributions to human and machine behaviour. Consequently, observers perceive the dispositional 'credibility' of a machine as different from that of a human.

4.2. *Source credibility (pedigree)*

Literature on inter-personal communication has identified *source credibility* as an important factor that influences the strength of information provided by an advisor (Sternthal *et al*. 1978, Birnbaum and Stegner 1979). Indeed, there is a large body of literature on the topic of persuasion indicating that the extent to which a perceiver accepts or rejects information depends to a large extent on the perceiver's opinion of the *source* of information. The use of well-known movie stars to promote a product or service is an everyday application of this principle.

'Credibility' has been defined as 'the extent to which an entity can be relied on to do what it says it will do' (Herbig and Milewicz 1993). The effect of credibility on user perception of information sources is supported by the Elaboration Likelihood Model (ELM, Petty and Cacioppo 1986a, b), a social psychological theory of persuasion. According to the ELM, there are two paths that individuals can use when processing persuasive communications: the central route and the peripheral route. The central route is used by individuals to evaluate a message when they are highly motivated and confident in their ability to critically analyse the message. However, when a person is not motivated or capable or becomes distracted, the processing of information goes through the peripheral route. When this happens, the recipient's judgement of a message is not based on a thorough examination of its contents. Instead, judgements are based upon or influenced by surface characteristics of the source, such as presumed expertise or credibility. Thus, they are more likely to agree with advice if the source is thought to be credible, even if the actual advice is incorrect.

The concept of credibility has several dimensions, out of which four emerge as most relevant, viz honesty, expertise, predictability and reputation (Deustch 1958, Kee and Knox 1970, Barber 1983, Dasgupta 1988, Lewicki and Bunker 1995, Mc Knight and Chervany 2002, Corriotore *et al*. 2003). Fogg and Tseng (1999) describe the construct of credibility in terms of 'expertise' and 'trustworthiness', with trustworthiness frequently being interpreted as synonymous with 'honesty'. 'Familiarity' of the trustee to the trustor reportedly has a strong effect on the perceived credibility of a source of information and has been demonstrated to be a powerful mediator of the human–computer relationship (Moon and Nass 1998). Ganesan (1994) identified 'reputation' as a primary characteristic of credibility assessments, while other researchers suggest that 'predictability' of a trustee's action is an important factor that affects the trustor's assessments of the credibility of the information source (Kee and Knox 1970, Rotter 1971, Barney and Hansen 1994, Fogg *et al*. 2001a, b).

Relatively few studies, however, have compared the effects of source credibility on users' trust of human vs automated advisors. One such study was conducted by Lerch *et al*. (1997) using an adaptation of the 'Turing test' described earlier. In their experiment, advice on a series of financial problems was attributed to three different sources depending on experimental condition: an expert automated system,

an expert human and a novice human, thus attempting to manipulate *source pedigree* (degree of expertise). For each of 10 financial problems participants were presented with a description of the decision to be made and a solution (i.e. advice) proposed by the source. After reading the source's proposed advice, participants rated the extent to which they agreed with the advice (assessing predictability) and their confidence in the source of the advice (assessing dependability). Results revealed a discrepancy between people's tendency to rely on the advisor and their confidence in their final decisions. Agreement with the advisor was generally higher, albeit non-significantly, for participants using the automated expert system than for participants receiving advice from either human advisor. On the other hand, confidence estimates were significantly higher when information was provided by an expert human advisor than a novice human or an expert system.

In addition to the apparent discrepancy between agreement rates and confidence estimates, participants in the above experiment also generated different reasons for the performance of the human and expert system advisors. While the human expert's performance was explained in terms of knowledge, effort and experience, the expert system's performance was attributed *only* to knowledge. Moreover, participants using the expert system rated 'effort' the lowest among the three sources in determining performance. In other words, participants utilized knowledge-linked information (i.e. effort, expertise) while assessing the human aids, but used performance-linked information (i.e. relative accuracy) while assessing the performance of the expert system. This finding suggests that participants viewed 'effort' as an internal attribution that can only be applied to describe human performance.

On the whole, the findings by Lerch *et al*. (1997) revealed that (1) participants had lower confidence in advice from an expert system than from a human expert, (2) participants rated 'effort' as contributing less to the performance of an expert system than a human expert and (3) while human advisors conveyed information on *dependability* to the user leading to higher confidence ratings, the expert system conveyed information on *predictability* of behaviour leading to greater agreement rates. These findings support the hypotheses of Rempel *et al*. (1985) that confidence ratings are generated by an attributional process regarding the source, while agreement rates are based on the source's dependability in specific situations.

The findings of Lerch *et al*. (1997) appear to contradict those reported by Dzindolet *et al*. (2002), Dijkstra *et al*. (1998) and Dijkstra (1999); discussed in the earlier section on 'source of information', in which operators tended to trust an automated aid more than a human advisor. However, in these other previous studies, human operators were *not* characterized as 'experts'. For example, in the Dzindolet *et al*. (2001, 2002) studies, the human advisors were characterized as other college students who had previously completed the study. Hence, they were presumably novices rather than experts, as utilized by Lerch *et al*. Perhaps, as suggested by the findings of Lerch *et al*., the bias to trust automation more than humans is reduced, eliminated or even reversed when perceived expertise of the human and automated advisors is equated. However, it should be noted that previous studies other than that by Lerch *et al*. did not characterize the automated aids as 'expert systems' either and participants were often informed that the automated aids were not perfectly reliable. Hence, the pedigree explanation of automation bias in these previous studies may not fully account for the discrepancies across experiments.

In summary, research on human interaction with DSSs has revealed that decisions to trust in and rely on an aid are largely a function of the perceived credibility of these systems. Studies that have examined the effect of source credibility on user trust and reliance have revealed that human operators have a tendency to be influenced by surface characteristics of the source of information rather than the actual diagnostic value of a piece of information. This eventually leads operators to interact with the aid based on its presumed rather than actual expertise. While research comparing human and automated advisors has revealed that automation, in general, is trusted more than humans (Dijkstra 1999, Dzindolet *et al*. 2002), this relative preference for automation is attenuated when a human advisor's reliability is framed in terms of greater relative expertise. Therefore, trust in a DSS is a function of multiple psychological factors that include users' perceptions of the source of information as well as the actual and perceived credibility of the source. While users' subjective perceptions of information source and credibility evidently have significant effects on their choice to utilize DSSs, these subjective effects are moderated by the actual reliability or accuracy of performance of the DSS itself.

4.3. *Source reliability*

The appropriateness of DSS reliance associated with trust depends to a large extent on how trust matches the true reliability of the diagnostic aid. Lee and See (2004) identify calibration, resolution and specificity of trust as three parameters that describe mismatches between subjective trust and the actual capabilities of an automated aid. Calibration refers to the correspondence between a person's trust and the capabilities of automation (Muir 1987, Lee and Moray 1994). Poor calibration is characterized as either over-trust where trust exceeds system capabilities or distrust where trust falls short of automation capabilities. Resolution refers to how precisely a judgement of trust differentiates levels of automation reliability (Cohen *et al*. 1998), while specificity refers to the degree to which trust is associated with a particular component or aspect of the trustee such as a DSS (Lee and See 2004).

As mentioned, source reliability is undoubtedly a critical factor affecting trust. However, the actual reliability of DSSs vary across the experiments discussed above. Some studies have used highly reliable systems, while others have used relatively imperfect DSSs. As mentioned above, the current conceptualization of automation trust is based on the assumption that users generally adapt their trust levels to accommodate different levels of DSS reliability (although such changes in trust may not always be perfectly calibrated with changes in diagnostic aid reliability). Only recently, however, have researchers attempted to directly test this assumption by comparing the effects that different levels of diagnostic aid reliability have on users' trust levels (see Wickens and Hollands (2000) for a review). Indeed, most studies have examined how trust develops when interacting with automation of a single reliability level or how a solitary automation failure can affect users' trust of a system that has been completely reliable prior to the failure (Lee and Moray 1994). Results of the few studies that have systematically varied automation reliability levels are mixed, with some suggesting that operators are sensitive to different levels of reliability (Parasuraman *et al*. 1993), while others suggest that operators are insensitive to reliability differences (Dzindolet *et al*. 2001).

A growing body of evidence, however, is showing that operators are sensitive to at least moderate differences in aid reliabilities, as well as the types of errors made by an automated aid. For example, Wiegmann *et al*. (2001) examined the effects that different levels of and changes in automation reliability have on users' trust of automated diagnostic aids. Participants were presented with a series of testing trials in which they diagnosed the validity of a system failure using only information provided to them by an automated DSS. The initial reliability of the aid was 60%, 80% or 100%. However, for participants who were initially provided the 60%-reliable aid, the accuracy of the aid increased to 80% half way through testing, whereas for participants who were initially provided the 100%-reliable aid, the aid's reliability was reduced to 80%. Aid accuracy remained at 80% throughout testing for participants in the 80%-reliability group. Both subjective measures (i.e. perceived reliability of the aid and subjective confidence ratings) and objective measures of performance (concurrence with the aid's diagnosis and decision times) indicated that users were sensitive to different levels of aid reliabilities, as well as to subsequent changes in initial aid reliabilities. However, objective performance measures were related to, but not perfectly calibrated with subjective measures of confidence and reliability estimates.

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. Such observations appear to be inconsistent with the findings discussed earlier that people on an average show a 'positivity bias' (Bruner and Tagiuri 1954, Cacioppo and Bernston 1994) or a tendency to assign positive evaluations to unfamiliar objects rather than negative or neutral evaluations (Cacioppo *et al*. 1997), as well as operators' proclivity to assume that automation is reliable until it proves otherwise. Nonetheless, not much research has yet examined the effect of automated aids of 'new' or unproven reliability on trust and reliance.

One recent study by Madhavan and Wiegmann (in review) has attempted to address this issue in a task where participants performed an airline luggage-screening task with the assistance of different sources of information (human vs automated advisor) that varied in credibility (expert vs novice) and reliability (high, 90% vs low, 70%). Results of the study revealed that, when the overall reliability of advisors was low (70%), trust in a 'novice' automated aid was higher than in a 'novice' human advisor, suggesting that low expectations of expertise lead users to prefer a machine to a human advisor. This supports the notion that operators assume an automated aid to be more reliable than a human until the former proves otherwise. However, when 70% reliable advisors were portrayed as 'experts', there were more rapid breakdowns in trust in the 'expert' automated aid compared to the 'expert' human advisor, suggesting that high initial expectations of expertise accompanied by low levels of actual performance (or reliability) lead to more damaging effects on automation trust than on human trust. When advisors were highly reliable (90%), high situational accuracy compensated for the effects of information source and credibility, ultimately equating the nature of human–human trust to human–automation trust.

5. Model of sequential trust development in human vs automated advisors

Based on the above review, we propose an integrated framework that compares the process of sequential trust development when the advisor is either an automated

system or a human with differing credibility and reliability (see figure 2). As indicated by previous research (Lerch *et al*. 1997) the development of trust is typically a function of both the dispositional features of the source of information and its behaviour in specific situations, as well as recipient biases and response tendencies. The precise dispositional features of the source may vary depending on whether the source is a machine or another human. For instance, a machine is perceived as having unique properties such as invariance and perseverance (Wiener and Curry 1980, Connors *et al*. 1994) or the ability to perform consistently across situations. On the other hand, humans are perceived as relatively more adaptable

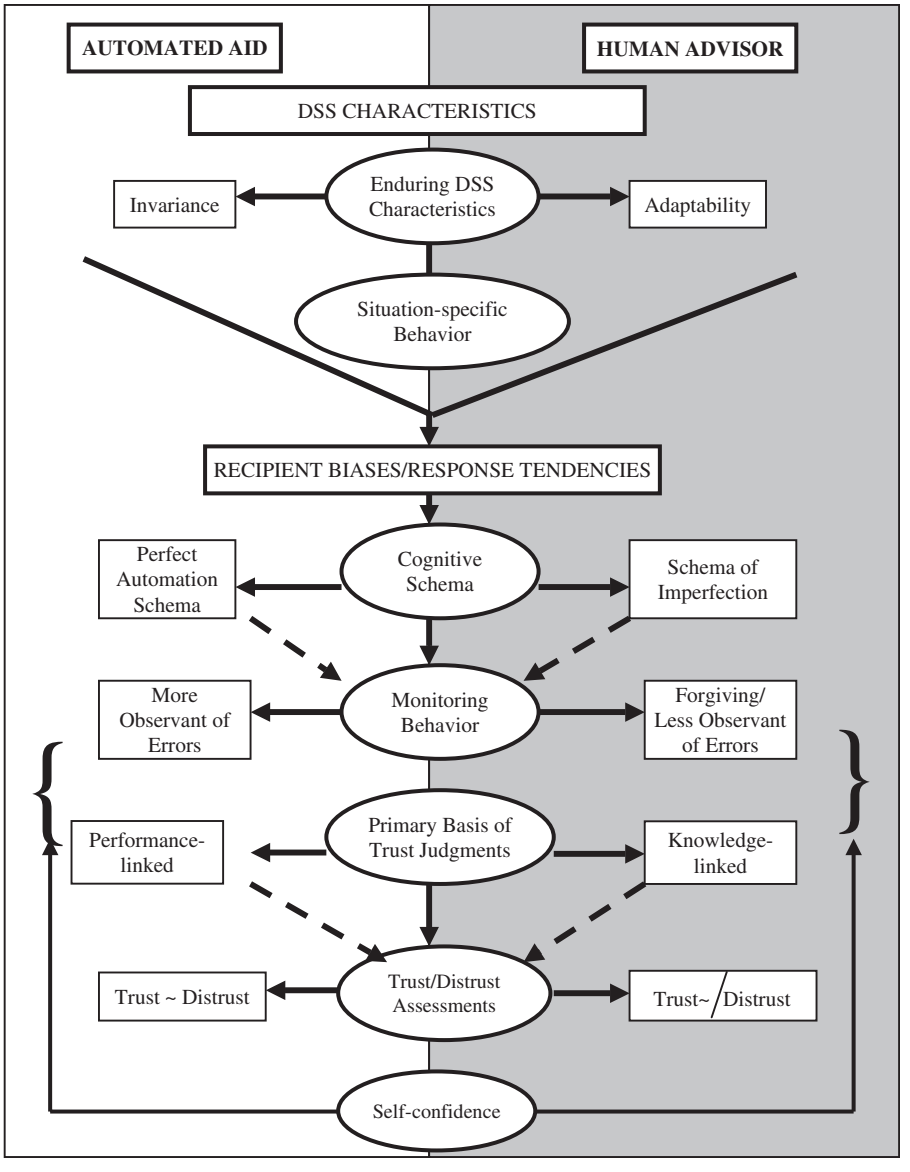

Figure 2. Model of sequential trust development in human vs automated DSSs.

(Connors 1990) and capable of changing their behavioural patterns according to the demands of specific situations. As depicted in figure 2, the knowledge of such unique dispositional characteristics of the DSS combined with its observed behaviour is filtered through the user's perceptions and biases leading to the development of a specific level of trust in the aid.

Assessments of aid behaviour are filtered through the operator's cognitive schemas that are either expectations of 'perfection' or high credibility assessments in the case of automation (Dzindolet *et al*. 2002) or 'imperfection' or low credibility assessments in the case of humans. Such filtering of observed aid behaviour induces operators to adopt a particular aid monitoring strategy, i.e. whether to monitor the aid's behaviour more closely and, consequently, become more sensitive to errors in the case of automation; or to be more forgiving and consequently less observant of errors, as in the case of a human advisor. This monitoring strategy combines with the primary bases of human trust judgements, which are either performance-linked reflecting strong situational influences in the case of automation or knowledge-linked reflecting the influence of dispositional characteristics (e.g. effort, expertise) in the case of a human advisor (Lerch *et al*. 1997).

Operator self-confidence in unaided performance has a strong influence on both aid monitoring behaviour and trust judgements. This ultimately leads to the actual assessment of subjective trust or distrust in the aid, which differs in the case of human–human relationships compared to human–machine relationships. Research in human–automation interaction has revealed that people are less extreme in their assessments of human–human distrust than trust, while this is not the case for assessments of human–machine trust and distrust (Jian *et al*. 2000). In other words, people are relatively more reluctant to claim that they distrust another human as opposed to an automated aid (Jian *et al*. 2000). Therefore, while the overall situation specific behaviour or reliability of the aid may be similar in the case of both human and automated DSSs, multiple cognitive biases and response tendencies of the user are often instrumental in producing verbal assessments of trust and distrust that differ distinctly for human and automated DSSs.

5.1. *Existing integrative frameworks of DSS trust and reliance*

The above review indicates that operator trust in DSSs is often influenced by the source of information (i.e. automated aid or another human), system characteristics such as reliability and visible 'behaviour' and the perceived credibility or expertise of the aid. These factors have been effectively integrated into several existing frameworks of automation trust and reliance (Lee and Moray 1992, 1994, Mosier and Skitka 1996, Parasuraman and Riley 1997, Dzindolet *et al*. 2001, Wiegmann *et al*. 2001, Wiegmann 2002), which provide valuable insight into the precise manner in which trust interacts with various psychological and environmental factors to lead to the choice of a particular automation utilization strategy. These frameworks illustrate that the relationship between automation trust and reliance is mediated by several cognitive factors such as the perceived reliability of automation, operator self-confidence and decision-making biases of the human operator (Dzindolet *et al*. 2001, Wiegmann *et al*. 2001, Madhavan and Wiegmann in press).

Riley (1989) suggested that an operator's decision to rely on automation might depend not only on the operator's level of trust in the system, but rather on a more

complex relationship among trust, self-confidence and a number of other individual difference factors. These other factors could include the operator's level of workload, level of risk associated with the situation, task complexity, etc. Automation 'reliance' is based on the probability that an operator will use the automation in future and is influenced by the operator's level of trust in automation. Trust, in turn, is influenced by the actual reliability of the automation and a 'duration' factor, which is meant to account for increasing stability of the operator's opinion of the automation as the operator gains experience with it. Muir's (1987) theory of trust between humans and automated aids provides support for the above-proposed relationship among automation reliability, trust in automation and reliance.

One recent model of trust and reliance pertaining specifically to DSSs provides a conceptual framework for understanding the relationship between automation trust and reliance. This framework, which is based on the model described by Dzindolet *et al*. (2001), illustrates that the actual reliability of an automated diagnostic aid is filtered through the operator's 'perfect automation schema', which typically distorts an operator's perceived reliability (i.e. trust) of an imperfect aid. This perceived reliability of the aid is then compared to the operator's perceived reliability of his or her own performance without the aid (i.e. manual performance or unaided diagnosis).

Perceptions of manual reliability or self-competence, in turn, are determined by the actual reliability of manual performance as filtered through the operator's personal biases such as self-confidence and general misconceptions about one's own competence (Dunning *et al*. 2003, Ehrlinger and Dunning 2003). These two factors (perceived reliability of self and aid) combine to influence the operator's impression of the aid's utility for a given task and the subsequent selection of an 'appropriate' utilization (i.e. agreement) strategy. Thus, there is substantial evidence that both the perceived reliability of an automated aid and operator self-confidence play a major role in automation utilization (Lee and Moray 1992, Wickens and Hollands 2000, Dzindolet *et al*. 2001, Wiegmann *et al*. 2001). This supports Riley's (1989) suggestion that three cognitive components of the human operator—reliability, trust and confidence—act as mediating variables in determining the interaction of human operators with automation.

Figure 3 illustrates a modified version of Dzindolet *et al*.'s (2001) framework of diagnostic aid reliance that reflects the differential effects of information source (human or automated aid), credibility and reliability (and/or 'visible aid behaviour') on operator trust in a DSS. While existing models have addressed the reliability of the aid (actual and perceived) and self-biases as integral factors in the selection of a utilization strategy, the present model emphasizes the distinction in the process of trust development when the source of diagnostic information (human or automation) is either *known* or *unknown* to the user. As can be seen in figure 3, the actual reliability of the aid combines with extraneous information regarding source credibility as well as user observations of the aid's behaviour in specific situations to lead to the development of a certain level of trust in the diagnostic aid.

When the source of information is *known* to the operator (i.e. human advisor or automated aid), the combined perception of aid reliability, credibility and behaviour is filtered through the operator's schema regarding the expected behaviour of the source. These schemas could either be expectations of 'near perfect' performance by an automated aid (Dzindolet *et al*. 2002) or 'less than perfect' performance by a human team-mate. Information filtered through such cognitive schemas gives

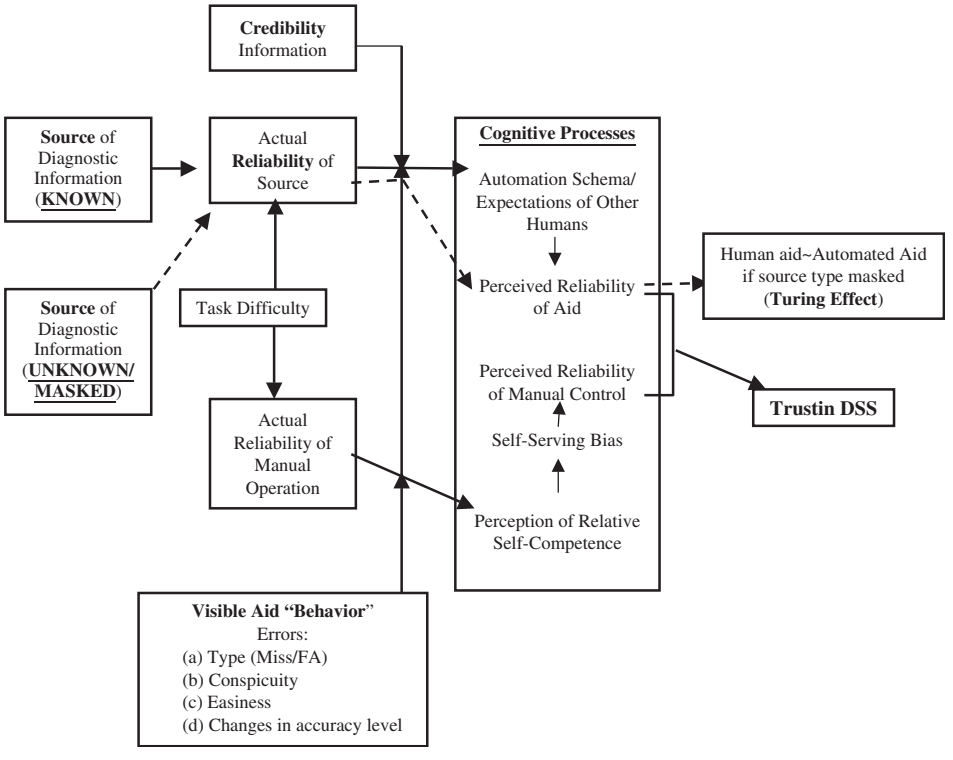

Figure 3. Differential effects of unknown and known sources of diagnostic information on the development of trust in a DSS. Modified from Dzindolet *et al.* (2001).

rise to the operator's perceived reliability of the aid that combines with the operator's perceptions of self-competence (relative to the diagnostic aid), ultimately influencing the operator's level of trust in the DSS. While the original Dzindolet *et al.* model emphasized self-biases and the actual reliability of manual operation as the primary factors affecting the operator's perceptions of self-competence, we propose the aid's 'visible behaviour', primarily constituting the conspicuity, easiness and type of errors being generated by a DSS, as well as changes in aid accuracy during a task, as important factors that influence operator's weighting of self-competence relative to a diagnostic aid.

The second scenario occurs when the source of diagnostic information is *unknown* or when user interaction with the aid occurs through mediums where the actual source of information is masked. Examples of such interactions where users are unaware of the source of information are computer-based troubleshooting tasks where the user types in requests or questions into a 'chat window' and receives responses from a human or automated 'advisor' whose identity is unknown. In such situations the combination of source reliability, credibility and aid behaviour will directly influence the user's perceived reliability of the aid *without* being filtered through cognitive schemas or specific expectations of behaviour by the DSS. In such cases, the perceived reliability of an automated aid is likely to be more or less equivalent to that of a human advisor, as the information about the source is schema-ambivalent, thereby demonstrating the 'Turing effect'.

6. Secondary factors mediating the effects of information source, credibility and reliability on trust development

While some of the primary factors affecting trust development have been discussed in detail above, several researchers have enumerated various additional subjective and objective factors that appear to mediate the effects of source, credibility and reliability on DSS trust and dependence. Recently, Maltz and Shinar (2003) examined the interaction between automation reliability and *task difficulty* on performance and automation reliance in a visual search task. In general, the findings indicated that participants depended on the aid (or cue) when it was more reliable and when the task was difficult. The results also indicated that performance was affected more by the false alarm rate of the cueing system than its hit rate. In other words, performance was affected more by false alarms than misses.

This is in keeping with other research findings that aids generating a large number of false alarms (i.e. 'cry wolf', Breznitz 1983) create under-trust in automation (Parasuraman and Riley 1997, Gupta *et al*. 2001) and consequently affect user compliance with automated aids. Relatively little research appears to have directly examined the relative consequences of *false alarms vs misses* in influencing human trust and reliance on DSSs. However, a few recent studies such as discussed above suggest that false alarms may indeed be more degrading of trust than misses (Cotte' *et al*. 2001, Gupta *et al*. 2001, Maltz and Shinar 2003).

In the context of automated alarm systems, a false alarm directs the operator's attention from other important tasks, ultimately resulting in the operator wasting time and effort in dealing with the alarm. Therefore, operators are likely to lose trust in such a system that demands this extra effort (Thomas *et al*. 2003). A miss, on the other hand, does not require extra attention to be focused on a false event, thereby maintaining the initial level of trust in the system. However, contradicting evidence suggests that under some circumstances misses are potentially more damaging and degrading of trust than false alarms (Masalonis and Parasuraman 2003). An example of such a situation is the catastrophic mid-air collision of two aircraft as a consequence of cockpit automation failing to detect and signal the presence of conflicting flight paths. Therefore, the misses vs false alarms trade-off and its effects on trust calibration is likely mediated by both the *immediacy* and *criticality* of error consequences.

Research also suggests that automated aids may be trusted differently depending upon the apparent *easiness or simplicity of automation errors*. Specifically, Madhavan *et al*. (2003) conducted a study where participants performed a target detection task wherein some trials were easy and others were difficult, with the help of an automated diagnostic aid. While the overall 70% reliability of the automated aid was equal for all participants, automation errors were distributed such that the aid generated errors either exclusively on easy or on difficult trials for each of two groups. Results of the study revealed that participants using the aid that missed targets on 'easy' trials mistrusted the aid, misperceived its reliability and disagreed with the aid more frequently than did participants using an aid that generated errors only on difficult trials. A follow-up study (Madhaven *et al*. 2004) revealed that the effect of 'easy' automation errors on trust and dependence was salient even when the errors were only occasional and comprised false alarms. The results of these studies support the 'easy-errors hypothesis' that automation errors on tasks easily performed by operators undermines operator trust in and reliance on

automated aids, even if the aid is, on average, more accurate than the unaided human operator.

As is evident from the above study, trust and DSS utilization are significantly influenced by the individual *operator's ability* to perform the task without diagnostic assistance. For example, when operators trust a diagnostic aid that is more reliable than manual (i.e. unaided) performance, they are more likely to rely on the aid than on their own diagnoses. Similarly, when operators distrust a diagnostic aid that is less reliable than manual performance, they are more likely to ignore the aid and rely on themselves to diagnose a situation (i.e. self-reliance). In both cases, appropriate reliance occurs (Dzindolet *et al.* 2003). However, over-reliance or misuse can occur when an operator over-trusts an aid that is less reliable than unaided performance. In addition, when operators under-trust an aid that is more reliable than manual performance, under-reliance or disuse of the aid can occur (Parasuraman and Riley 1997).

Assessments of manual ability are frequently affected by participants' *self-confidence* in their own abilities to perform the various tasks without diagnostic assistance. Self-confidence, in turn, is reportedly influenced by the type of *feedback* provided by a DSS. A study by Fogg and Nass (1997) revealed that humans are easily susceptible to 'flattery' or praise from computers because humans have a natural tendency to accept positive feedback without scrutiny. On the one hand, such DSS-generated praise might increase user assessments of self-efficacy and task persistence; on the other hand, it might adversely inflate operator self-confidence and lead to inappropriate reliance on diagnostic aids.

Finally, the calibration of trust is further influenced by the interaction of DSS features such as source, credibility and reliability with *environmental factors*. A necessary step in improving decision performance in complex decision-making tasks with computerized aids or training programs is to understand the sources of performance limitations of DSSs (Bisantz *et al.* 2000). In addition to the factors already discussed above, some additional factors that could limit the performance standards of DSSs could be the fundamental content knowledge needed to perform complex tasks, dynamic environmental states, the degree of adaptability of the system and the nature of interface design. Overall, the issue of operator trust in a DSS is a multi-faceted one that is dependent on a variety of factors that encompass features of the DSS, the human operator and the environment.

7. Implications for DSS design and trust calibration: bridging the gap between human–human trust and human–machine trust

The models of user perception of automated aids vs human advisors described above clearly suggest that, while trust in expert systems develops in a manner akin to trust among humans, there are some critical differences in the manner in which people react to automated advice vs human advice. Specifically, the two frameworks of trust presented above imply that the process of trust development in humans and automation is comparable. Yet, there are differences in the manner in which this trust is ultimately expressed. This suggests that calibration of user trust in DSSs can be

improved by attempting to bridge the gap between human perception of humans vs their perception of machines, thereby leading to comparable levels of trust among different types of DSSs with varying credibility and reliability levels. This is corroborated by Lee and See (2004) who contend that the primary goal of human–automation research is to make automation highly, but not excessively, 'trustable'.

Given that people apply similar response tendencies and filtering strategies to calibrate their trust in humans and automated aids while frequently applying social rules of human–human interaction to machines (a phenomenon known as ethopoeia (Nass and Moon 2000)), automation users would benefit if machines were designed to incorporate characteristics of humanness that would, in turn, elicit social responses from the human user. For example, computer-generated feedback and voice alert mechanisms that mimic human language structures and accents might have a strong potential to elicit user responses that mirror responses typically generated in social inter-personal contexts. Such anthropomorphizing of automation will likely have the following potential consequences:

- 'Humane responses' or characteristics of machines will lead users to automatically apply reciprocal behavioural actions and reactions (Nass and Moon 2000) in their routine interaction with automation.
- The tendency toward human–human social responses while dealing with automated aids will neutralize or reduce the biases uniquely associated with automation such as the schema of *credibility* and over attention to automation errors or its situation-specific *reliability* (illustrated in figure 2).
- Reduction of specific automation biases will lead to greater correspondence between the processes of trust development in automated DSSs vs human partners (an extension of the 'Turing effect' illustrated in figure 3).
- Uniformity in the process of human trust development in automated aids vs human partners will increase the feasibility of deriving a relatively predictable model of human trusting behaviour in relation to automation, which can eventually be incorporated into the design of future DSSs.

Although there are many categories of social rules that are likely to be mindlessly triggered by machines, no formal typology exists (Nass and Moon 2000). Social attitudes and behaviours associated with human–human inter-personal exchange are often culture- and situation-dependent. Global social norms that are used frequently (e.g. politeness, good manners) might be easier to elicit when users interact with 'humane' DSSs. Conversely, such stereotypical reactions could likely get confused with more intricate patterns of social behaviour that are dependent on culture, ethnicity and gender, ultimately causing the quality of human–DSS interaction to breakdown. For instance, Winograd and Flores (1987) observe that certain characteristics of computers such as limited vocabulary and occasionally inexplicable behaviour may remind the user of a 'foreigner' or a person from a different culture, thereby eliciting varied reactions from users that are either likely to be misinterpreted by the designer or difficult to quantify. Therefore, while the application of humane characteristics to automation is a commendable goal in the design of future DSSs, care must be taken to avoid over-anthropomorphizing of automation in a manner that begins to encroach on the deficiencies and biases inherent in human–human social interaction.

8. Conclusion

The *trust* placed in diagnostic aids by the human operator is one of the most critical psychological factors that influences operator acceptance or rejection of advice from DSSs (Dzindolet *et al*. 2002). On the one hand, an unusually low level of trust in a DSS will lead to aid disuse and inappropriate self-reliance; on the other hand, a very high level of trust in an automated system might lead to the development of automation-induced complacency or over-trust (Muir 1994).

Research comparing human–human trust with human–automation trust has revealed that, while humans have a natural propensity to react socially to machines, there are, nevertheless, subtle differences in the manner in which humans perceive automated aids vs human advisors. During the last few decades, various attempts have been made to empirically examine and quantify the precise role played by human trust in the development of automation utilization strategies and reliance compared to similar trust development in human team-mates or partners. One such method has been to model the automated system as an 'advisor' akin to a human 'advisor' in a decision-making context and draw inferences about performance, trust and reliance based on existing social psychological theories of human–human trust and advice acceptance.

Such research has resulted in the delineating of several psychological factors that typically bias human operators in favour of their automated or human team-mates as per the situation. The conceptual frameworks of trust introduced in the present paper suggest the benefits of attempting to bridge the gap between human understanding of automated agents vs human team-mates. Such attempts will allow for a seamless flow of communication between humans and their non-human counterparts in team environments, that is undoubtedly a necessary pre-condition for the smooth functioning of complex systems that have become increasingly sophisticated in recent times.

References

Barber, B., 1983, *The Logic and Limits of Trust* (New Brunswick, NJ: Rutgers University Press).

Barney, J.B. and Hansen, M.B., 1994, Trustworthiness as a source of competitive advantage. *Strategic Management Journal*, **15**, pp. 175–190.

Birnbaum, M.H. and Stegner, S.E., 1979, Source credibility in social judgment: bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, **37**, pp. 48–74.

Bisantz, A.M., Kirlik, A., Gay, P., Phipps, D., Walker, N. and Fisk, A.D., 2000, Modeling and analysis of a dynamic judgment task using a lens model approach. *IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **30**, pp. 605–616.

Blau, P.M., 1964, *Exchange and Power in Social Life* (New York: Wiley).

Blomqvist, K., 1997, The many faces of trust. *Scandinavian Journal of Management*, **13**, pp. 271–286.

Bowers, C.A., Jentsch, F., Salas, E. and Braun, C.C., 1998, Analyzing communication sequences for team training needs assessment. *Human Factors*, **40**, pp. 672–680.

Bowers, C.A., Oser, R.A., Salas, E. and Cannon-Bowers, J.A., 1996, Team performance in automated systems. In *Automation and Human performance: Theory and Application*, R. Parasuraman and M. Mouloua (Eds), pp. 243–263 (Mahwah, NJ: Lawrence Erlbaum Associates).

Breznitz, S., 1983, Cry-wolf: The Psychology of False Alarms (Mahwah, NJ: Erlbaum).

Bruner, I.S. and Tagiuri, R., 1954, The perception of people. In *Handbook of Social Psychology*, G. Lindzey (Ed.), Vol. 2, pp. 634–654 (Reading, MA: Addison-Wesley).

Cacioppo, J.T. and Bernston, G.G., 1994, Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, **115**, pp. 401–423.

Cacioppo, J.T., Gardner, W.L. and Bernston, G.G., 1997, Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Personality and Social Psychology Review*, **1**, pp. 3–25.

Cassell, J. and Bickmore, T., 2000, External manifestations of trustworthiness in the interface. *Communications of the ACM*, **43**, pp. 50–56.

Cohen, M.S., Parasuraman, R. and Freeman, J., 1998, Trust in decision aids: a model and its training implications. In *Proceedings of the Command and Control Research and Technology Symposium*.

Connors, M.M., 1990, Crew system dynamics: combining humans and automation. SAE Technical Paper 891530. In *19th Intersociety Conference on Environmental Systems*, San Diego, CA.

Connors, M.M., Harrison, A.A. and Summit, J., 1994, Crew systems: integrating human and technical subsystems for the exploration of space. *Behavioral Science*, **39**, pp. 183–212.

Corriotore, C.L., Kracher, B. and Wiedenbeck, S., 2003, On-line trust: concepts, evolving themes, a model. *International Journal of Human–Computer Studies*, **58**, pp. 737–758.

Cotte, N., Meyer, J. and Coughlin, J.F., 2001, Older and younger drivers' reliance on collision warning systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (Santa Monica, CA: Human Factors Society), pp. 227–230

Couch, L.L. and Jones, W.H., 1997, Measuring levels of trust. *Journal of Research in Personality*, **31**, pp. 319–336.

Dasgupta, P., 1988, Trust as a commodity. In *Trust: Making and Breaking Cooperative Relations*, D. Gambetta (Ed.), pp. 49–72 (New York: Basil Blackwell).

Deutsch, M., 1958, Trust and suspicion. *Journal of Conflict Resolution*, **2**, pp. 265–279.

Dijkstra, J.J., 1995, The influence of an expert system on the user's view: how to fool a lawyer. *New Review of Applied Expert Systems*, **1**, pp. 123–138.

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

Dijkstra, J.J., Liebrand, W.B.G. and Timminga, E., 1998, Persuasiveness of expert systems. *Behaviour and Information Technology*, **17**, pp. 155–163.

Driscoll, J.W., 1978, Trust and participation in organizational decision making as predictors of satisfaction. *Academy of Management Journal*, **21**, pp. 44–56.

Dunning, D., Johnson, K., Ehrlinger, J. and Kruger, J., 2003, Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, **1**, pp. 83–87.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. and Beck, H.P., 2003, The role of trust in automation reliance. *International Journal of Human–Computer Studies*, **58**, pp. 697–718.

Dzindolet, M.T., Pierce, L.G., Beck, H.P. and Dawe, L.A., 2002, The perceived utility of human and automated aids in a visual detection task. *Human Factors*, **44**, pp. 79–94.

Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A. and Anderson, B.W., 2001, Predicting misuse and disuse of combat identification systems. *Military Psychology*, **13**, pp. 147–164.

Earley, P.C., 1988, Computer-generated performance feedback in magazine subscription industry. *Organizational Behavior and Human Decision Processes*, **41**, pp. 50–64.

Ehrlinger, J. and Dunning, D., 2003, How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, **84**, pp. 5–17.

Endsley, M.R. and Kaber, D.B., 1999, Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, **42**, pp. 462–492.

Fogg, B.J. and Nass, C., 1997, Silicon sycophants: the effects of computers that flatter. *International Journal of Human–Computer Studies*, **46**, pp. 551–561.

Fogg, B.J. and Tseng, H., 1999, The elements of computer credibility. In *Proceedings of the CHI'99* (New York: ACM Press), pp. 295–296.

Fogg., B.J., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J. and Brown, B., 2001a, Web credibility research: a method for online experiments and early study results. In *Proceedings of the Conference on Human Factors in Computing Systems CHI 2001* (New York: ACM Press), pp. 295–296.

Fogg., B.J., Marshall., J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. and Treinen, M., 2001b, What makes websites credible? A report on a large quantitative study. In *Proceedings of the Conference on Human Factors in Computing Systems CHI 2001* (New York: ACM Press), pp. 61–68.

Friedman, B., 1995, 'It's the computer's fault'—reasoning about computers as moral agents. *Proceedings of the CHI Conference*, Denver, CO.

Ganesan, S., 1994, Determinants of long-term orientation in buyer-seller relationships. *Journal of Marketing*, **58**, pp. 1–19.

Giffin, K., 1967, The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin*, **68**, pp. 104–120.

Gupta, N., Bisantz, A.M. and Singh, T., 2001, Investigation of factors affecting driver performance using adverse condition warning systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (Santa Monica, CA: Human Factors Society), pp. 1699–1703.

Herbig, P. and Milewicz, J., 1993, The relationship of reputation and credibility to brand success. *Journal of Consumer Marketing*, **10**, pp. 18–24.

Hwang, P. and Buergers, W.P., 1997, Properties of trust: an analytical view. *Organizational Behavior and Human Decision Processes*, **69**, pp. 67–73.

Jian, J.Y., Bisantz, A.M. and Drury, C.G., 2000, Foundations for an empirically determined scale of trust in automated systems, *International Journal of Cognitive Ergonomics*, **4**, pp. 53–71.

Kee, H.W. and Knox, R.E., 1970, Conceptual and methodological considerations in the study of trust and suspicion. *Conflict Resolution*, **14**, pp. 357–366.

Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R. and Feltovich, P.J., 2004, Ten challenges for making automation a 'team player' in joint human–agent activity. *IEEE Intelligent Systems*, **4**, pp. 1541–1672.

Krupinski, E.A., Nodine, C.F. and Kundel, H.L., 1993, Perceptual enhancement of tumor targets in chest x-ray images. *Perception and Psychophysics*, **53**, pp. 519–526.

Lee, J.D. and Moray, N., 1992, Trust, control strategies and allocation of function in human machine systems. *Ergonomics*, **22**, pp. 671–691.

Lee, J.D. and Moray, N., 1994, Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human–Computer Studies*, **40**, pp. 153–184.

Lee, J.D. and See, K.A., 2004, Trust in automation: designing for appropriate reliance. *Human Factors*, **46**(1), pp. 50–80.

Lerch, F.J., Prietula, M.J. and Kulik, C.T., 1997, The Turing effect: the nature of trust in expert system advice. In *Expertise in Context: Human and Machine*, P.J. Feltovich and K.M. Ford (Eds), pp. 417–448 (Cambridge, MA: The MIT Press).

Lewandowsky, S., Mundy, M. and Tan, G.P.A., 2000, The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied*, **6**, pp. 104–123.

Lewicki, R.J. and Bunker, B.B., 1995, Trust in relationships: a model of development and decline. In *Conflict, Cooperation, and Justice: Essays Inspired by the Work of Morton Deutsch*, B.B. Bunker and J.Z. Rubin (Eds), pp. 133–173 (San Francisco, CA: Jossey-Bass).

Lewis, D.J. and Weigert, A., 1985, Trust as a social reality. *Social Forces*, **63**, pp. 967–985.

Luhmann, N., 1979, *Trust and Power* (Chichester: Wiley).

Luhmann, N., 1988, Familiarity, confidence, trust: problems and alternatives. In *Trust—Making and Breaking Relationships*, D. Gambetta (Ed.), pp. 94–109 (Oxford: Basil Blackwell).

Madhaven, P. and Wiegmann. D.A., 2005, Cognitive anchoring on self-generated decisions reduces trust in automated diagnostic aids. *Human Factors*, **47**(2), pp. 332–341.

Madhaven, P. and 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* (Santa Monica, CA: Human Factors & Ergonomics Society).

Madhaven, P., Wiegmann, D.A. and Lacson, F.C., 2003, Automation failures on tasks easily performed by operators undermines trust in automated aids. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society* (Santa Monica, CA: Human Factors & Ergonomics Society), pp. 335–339.

Madhaven, P., Wiegmann, D.A. and Lacson, F.C., 2004, Occasional automation failures on easy tasks undermines trust in automation. *Paper presented at the 112th Annual Meeting of the American Psychological Association* (Honolulu, HI).

Maltz, M. and Shinar, D., 2003, New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, **45**, pp. 281–295.

Masalonis, A.J. and Parasuraman, R. 2003, Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (Santa Monica: Human Factors & Ergonomics Society), pp. 533–537.

McKnight, D.H. and Chervany, N.L., 2002, What trust means in e-commerce customer relationships: an interdisciplinary conceptual typology. *International Journal of Electronic Commerce*, **6**, pp. 35–59.

Moon, Y. and Nass, C., 1998, Are computers scapegoats? Attributions of responsibility in human–computer interaction. *International Journal of Human–Computer Studies*, **49**, pp. 79–94.

Morgan, T., 1992, Competence and responsibility in intelligent systems. *Artificial Intelligence Review*, **6**, pp. 217–226.

Mosier, K.L. and Skitka, L.J., 1996, Human decision makers and automated decision aids: made for each other? In *Automation and Human performance: Theory and Applications*, R. Parasuraman and M. Mouloua (Eds), pp. 201–220 (Mahwah, NJ: Lawrence Erlbaum Associates).

Muir, B.M., 1987, Trust between humans and machines, and the design of decision aids. *International Journal of Man–Machine Studies*, **27**, pp. 527–539.

Muir, B.M., 1994, Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, **37**, pp. 1905–1922.

Nass, C. and 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**, pp. 171–181.

Nass, C. and Moon, Y., 2000, Machines and mindlessness: social responses to computers. *Journal of Social Issues*, **56**, pp. 81–103.

Nass, C., Fogg., B.J. and Moon, Y., 1996, Can computers be teammates? *International Journal of Human–Computer Studies*, **45**, pp. 669–678.

Nass, C., Moon, Y. and Carney, P., 1999, Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, **29**, pp. 1093–1110.

Nass, C., Moon, Y., Fogg, B.J., Reeves, B. and Dryer, D.C., 1995, Can computer personalities be human personalities? *International Journal of Human–Computer Studies*, **43**, pp. 223–239.

Nass, C.L., Steuer, J. and Tauber, E., 1994, Computers are social actors. In *Proceedings of the CHI Conference* (Boston, MA: CHI).

Nass, C., Steuer, J., Tauber, E. and Reeder, H., 1993, Anthropomorphism, agency and ethopoeia: computers as social actors. In *Proceedings of the International CHI Conference* (Amsterdam: CHI).

Parasuraman, R. and Riley, V., 1997, Humans and automation: use, misuse, disuse, abuse. *Human Factors*, **39**, pp. 230–253.

Parasuraman, R., Molloy, R. and Singh, L.L., 1993, Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, **3**, pp. 1–23.

Parasuraman, R., Sheridan, T.B. and Wickens, C.D., 2000, A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, **30**, pp. 286–297.

Petty, R.E. and Cacioppo, J.T., 1986a, *Communication and Persuasion, Central and Peripheral Routes to Attitude Change* (New York: Springer-Verlag).

Petty, R.E. and Cacioppo, J.T., 1986b, The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology*, L. Berkowitz (Ed.), Vol. 19, pp. 123–205 (New York: Academic Press).

Reeder, G.D. and Brewer, M.B., 1979, A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, **86**, pp. 61–79.

Reeves, B. and Nass, C., 1996, *The Media Equation: How People Treat Computers, Television and the New Media Like Real People and Places* (Stanford, CA: Center for the Study of Language and Information).

Rempel, J.K., Holmes, J.G. and Zanna, M.P., 1985, Trust in close relationships. *Journal of Personality and Social Psychology*, **49**, pp. 95–112.

Riley, V., 1989, A general model of mixed-initiative human–machine systems. In *Proceedings of the 33rd Annual Human Factors Society Conference* (Santa Monica, CA: Human Factors & Ergonomics Society), pp. 124–128.

Rotter, J.B., 1967, A new scale for the measurement of interpersonal trust. *Journal of Personality*, **35**, pp. 651–665.

Rotter, J.B., 1971, Generalized expectancies for interpersonal trust. *American Psychologist*, **26**, pp. 443–452.

Rotter, J.B., 1980, Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, **35**, pp. 1–7.

Sage, A.P., 1987, Decision support systems. In *Handbook of Human Factors*, G. Salvendy (Ed.), pp. 1409–1448 (New York: Wiley).

Sampson, J.P., 1986, Computer technology and counseling psychology: regression toward the machine? *Counseling Psychologist*, **14**, pp. 567–583.

Sheridan, T.B., 2002, *Humans and Automation: System Design and Research Issues* (Santa Monica, CA: Wiley Interscience).

Sheridan, T.B. and Ferrell, W., 1974, *Man–machine Systems: Information, Control, and Decision Models of Human Performance* (Cambridge, MA: MIT Press).

Sheridan, T.B. and Verplank, W.L., 1978, *Human and Computer Control of Undersea Teleoperators* (Cambridge, MA: MIT Man-Machine Systems Laboratory).

Singh, I.L., Molloy, R. and Parasuraman, R., 1993, Individual differences in monitoring failures of automation. *Journal of General Psychology*, **42**, pp. 403–407.

Stangor, C. and McMillan, D., 1992, Memory for expectancy-congruent and expectancy-incongruent information: a review of the social and social developmental literatures. *Psychological Bulletin*, **111**, pp. 42–61.

Sternthal, B., Dholakia, R. and Leavitt, C., 1978, The persuasive effect of source credibility: tests of cognitive response. *Journal of Consumer Research*, **4**, pp. 252–260.

Sundar, S.S. and Nass, C., 2000, Source orientation in human–computer interaction: programmer, networker, or independent social actor? *Communication Research*, **27**, pp. 683–703.

Thomas, L.C., Wickens, C.D. and Rantanen, E.M., 2003, Imperfect automation in aviation traffic alerts: a review of conflict detection algorithms and their implications for human factors research. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society* (Santa Monica, CA: Human Factors & Ergonomics Society).

Turing, A.M., 1950, Computing machinery and intelligence, *Mind*, **59**, pp. 433–460.

Waern, Y. and Ramberg, R., 1996, People's perception of human and computer advice. *Computers in Human Behavior*, **12**, pp. 17–27.

Wickens, C.D. and Hollands, J.G., 2000, *Engineering Psychology and Human Performance*, 3rd edn. (Upper Saddle River, NJ: Prentice Hall).

WICKENS, C.D., GEMPLER, K. and MORPHEW, M.E., 2000, Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors Journal*.
WIEGMANN, D.A., 2002, Agreeing with automated diagnostic aids: a study of users' concurrence strategies. *Human Factors*, **44**, pp. 44–50.
WIEGMANN, D.A., RICH, A. and ZHANG, H., 2001, Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, **2**, pp. 352–367.
WIENER, E.L. and CURRY, R.E., 1980, Flight-deck automation: promises and problems. *Ergonomics*, **23**, pp. 995–1101.
WINOGRAD, T. and FLORES, C., 1987, *Understanding Computers and Cognition: A New Foundation for Design* (Reading, MA: Addison-Wesley).

About the authors

Poornima Madhaven is a post-doctoral fellow in the Dynamic Decision Making Laboratory within the Department of Social and Decision Sciences at Carnegie Mellon University. She received her PhD 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 within the Institute of Aviation at the University of Illinois at Urbana-Champaign. He holds appointments in the Department of Psychology and the Beckman Institute for Advanced Science and Technology. He received his PhD in Experimental Psychology from Texas Christian University in 1992. He is currently a visiting research scientist in the Division of Cardiovascular Surgery at the Mayo Clinic College of Medicine.