

A NEW LOOK AT THE DYNAMICS OF HUMAN-AUTOMATION TRUST: IS TRUST IN HUMANS COMPARABLE TO TRUST IN MACHINES?

Poornima Madhavan and Douglas A. Wiegmann
Aviation Human Factors Division, Institute of Aviation
University of Illinois at Urbana-Champaign

The trust placed in automated diagnostic aids by the human operator is one of the most critical psychological factors that influences operator reliance on decision support systems. Studies examining the nature of human interaction with automation have revealed that users have a propensity to apply norms of human-human interpersonal 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 teammates. The present review is focused on comparing the process of trust development in human-automation teams with that of human-human partnerships, specifically in the context of dyads that constitute a primary decision maker and either a human 'advisor' or an intelligent automated decision support system. A conceptual framework that synthesizes and contrasts the process of trust development in humans versus automation is proposed. Potential implications of this research include the improved design of decision support systems by incorporating features into automated aids that elicit operator responses that mirror responses in human-human interpersonal interaction.

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 has metamorphosed from that of a primary controller to an active teammate sharing control with automation. Automated decision aids are increasingly being modeled as 'partners' that support or assist the human in performing functions that may be either difficult or even impossible for the operator to perform without the assistance of a 'knowledgeable teammate'.

Intelligent decision aids are designed to interact or behave in a manner similar to humans, imitating human language structures where applicable and possessing unique knowledge and functional algorithms that may be inaccessible to the human teammate. Some researchers have argued that such human-automation teams function similarly to human-human teams (Bowers, Oser, Salas, & Cannon-Bowers, 1996), and evidence suggests that people do

enter into 'relationships' with computers, robots, and interactive machines in a manner similar to other humans (Nass, Fogg & Moon, 1996; Reeves & Nass, 1996). However, evidence to the contrary suggests 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), thereby raising the question of whether the genesis and development of human-machine trust is comparable to that of human-human trust.

RESEARCH QUESTIONS

Literature has addressed automation trust as a largely global construct (e.g. Blomqvist, 1997, Lewandowsky, Mundy & Tan, 2000, Lee & See, in press). In the present paper, we utilize existing research findings to portray human-automation trust as a mirror of human-human interpersonal trust in dyadic teams that constitute a primary decision maker and either a human 'advisor' or an intelligent automated aid. We attempt to address two main research issues: (1) whether human-automation trust

develops in a manner akin to human-human trust, and (2) whether trust in humans and automation will be indistinguishable if users are unaware of whether the source of diagnostic assistance is a machine or another human. We propose two unique theoretical frameworks that synthesize the process of trust development in human vs. automated decision aids and provide implications for the design of decision support systems.

Question 1: How Comparable is Human-automation Trust to Human-human Trust?

Model 1: Comparing sequential trust development in humans vs. automation. Figure 1 represents an integrated framework that compares the process of trust development when the decision aid is either an automated system or a human advisor. The development of trust is typically a function of both the dispositional features of the source of information and its behavior in specific situations, as well as recipient biases and response tendencies (cf. Lerch, Prietula, & Kulik, 1997). The dispositional features of the source may vary depending on whether the source is a machine or a human. For instance, a machine is perceived as having properties such as ‘invariance’, or the ability to perform consistently across situations. On the other hand, humans are perceived as relatively more adaptable and capable of changing their behavioral patterns across situations. As depicted in figure 1, the knowledge of such unique dispositional features of the aid combined with its observed behavior is filtered through the user’s biases leading to the development of a specific level of trust in the aid.

Assessments of aid behavior are filtered through the operator’s cognitive schemas that are either expectations of ‘perfection’ or high credibility assessments in the case of automation (cf. Dzindolet, Pierce, Beck, & Dawe, 2002), or expectations of ‘imperfection’ or low credibility assessments in the case of humans. Such filtering induces operators to adopt a particular aid monitoring strategy, i.e., whether to monitor the aid’s behavior more closely and become more sensitive to errors in the case of automation; or to be more forgiving and less observant of errors in the case of a human advisor. This monitoring strategy combines with the primary bases of trust judgments, which are 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 (cf. Lerch et al. 1997).

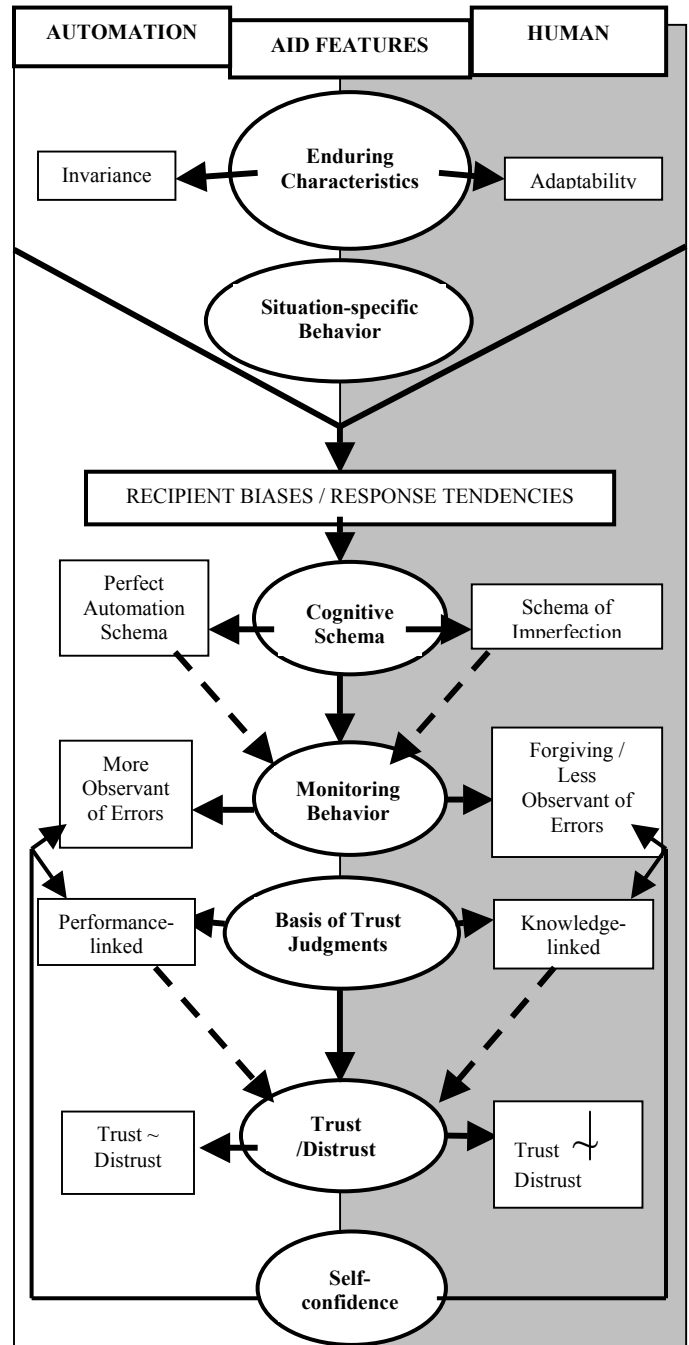


Figure 1. Model of sequential trust development in humans vs. automation

Operator self-confidence has a strong influence on both aid monitoring behavior and trust judgments. 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. People tend to be 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, Bisantz, & Drury, 2000). Users are relatively more reluctant to claim that they distrust another human as opposed to an automated aid. Therefore, when the reliability or ‘behavior’ of the aid is similar for both human and automated aids, the *process of trust development* is comparable in human-human and human-automation teams. However, multiple cognitive biases of the user often produce *verbal assessments of trust* and distrust that differ significantly for human and automated aids thereby giving human-human and human-automation trust the semblance of distinctiveness.

Question 2: Does the Awareness of Information Source Influence Trust Development?

Model 2: Trust development when source of diagnostic advice is either known or unknown to the recipient. Since verbal assessments of human-automation trust differ from that of human-human trust (model 1), the next question is whether this distinction is neutralized when users are unaware of whether the source of decision support is an automated aid or another human. In order to answer this, we present an extension of the framework of automation utilization developed by Dzindolet et al. (2002) to emphasize the distinction in trust development when the source of diagnostic information (human or automation) is either *known* or *unknown* to the user. As illustrated in figure 2, the actual reliability of the aid combines with extraneous information regarding source credibility and user observations of the aid’s behavior to lead to the development 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 visible behavior is filtered through the operator’s schema regarding the expected behavior of the source. These schemas could either be expectations of ‘near perfect’ performance by an automated aid or ‘less than perfect’ performance by a human teammate. Information filtered through cognitive schemas gives rise to the operator’s perceived reliability of the aid that combines with the operator’s perceptions of self-competence influencing the operator’s trust in the aid.

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 behavior’, namely the conspicuity, easiness and type of errors being generated by an aid as factors that influence operator’s weighting of self-competence relative to a diagnostic aid (cf. Madhavan, Wiegmann & Lacson, 2003).

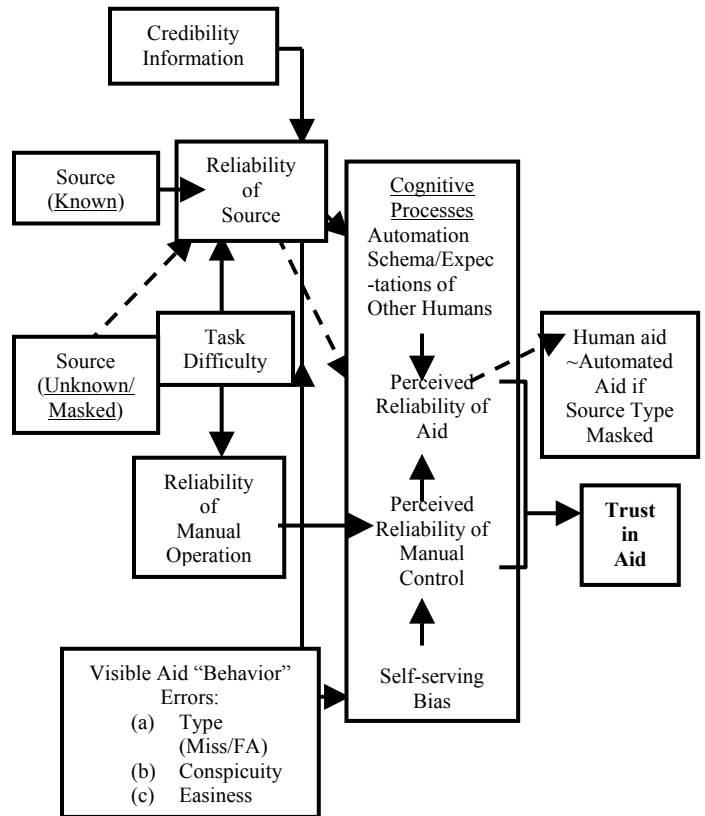


Figure 2. Trust development when source of diagnostic advice is either known or unknown.

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 (e.g. typed mediums or computer-based tasks) the combination of source reliability, credibility and aid behavior will directly influence the user’s perceived reliability of the aid without being filtered through cognitive schemas or expectations of behavior by the aid. 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. Therefore, the dynamics of human-

automation trust and human-machine trust are indistinguishable when the actual source of diagnostic information is masked or unknown to the user, suggesting that the process of trust development in a decision aid is primarily a function of the cognitive and psychological biases and response tendencies of the user (also demonstrated in model 1).

DISCUSSION AND IMPLICATIONS FOR DECISION AID DESIGN

Research on user perception of automated advisors has revealed 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 versus human advice. In general, expert systems are perceived as more credible than human advisors (e.g. Dijkstra 1999). While this initially leads to a bias toward automation in that people rely on automated information as a heuristic replacement for elaborate information processing, it eventually leads to a bias against automation, as people are more observant of automation errors than human errors (Dzindolet et al. 2002). Therefore, trust in automated aids is likely to breakdown more rapidly than trust in human teammates due to the existence of initial biases in favor of automation.

The subtle similarities and differences between human-human and human-automation interaction have several implications for decision aid design. According to Lee and See (2004), the primary goal of human-automation research is to make automation highly, but not excessively, 'trustable'. Given that people frequently apply social rules of human-human interaction to machines (a phenomenon known as *ethopoeia*, cf. Nass & 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 interpersonal contexts. Such anthropomorphizing of automation is likely to:

- Lead users to apply reciprocal behavioral actions (cf. Nass & Moon, 2000) in their routine interaction with automation.

- The tendency toward human-human social responses to automation will neutralize the biases associated with automation such as the schema of perfection and over-attention to errors (illustrated in model 1).
- Reduction of automation biases will lead to greater correspondence between the processes of trust development in automated aids versus human partners (illustrated in model 2).
- Uniformity in the process of human trust development in automation and humans will increase the feasibility of deriving a predictable model of human trust in relation to automation, which can be incorporated into the design of future decision aids.

Although there are many social behaviors that can be triggered by machines, the primary limitation of the above theory is that no formal typology of behaviors that can be elicited by effective anthropomorphism exists. Social behaviors associated with human-human interpersonal exchange are often culture and situation-dependent.

While global social norms that are used frequently (e.g. politeness) might be easier to elicit when users interact with 'humane' automated aids, such stereotypical reactions are likely to get confused with more intricate patterns of social behavior that are dependent on culture and gender, causing the breakdown of human-automation team performance. For instance, Winograd and Flores (1987) observe that certain characteristics of computers such as limited vocabulary and occasionally inexplicable behavior 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 human characteristics to automation is a commendable goal, care must be taken to avoid over-anthropomorphizing of automation in a manner that begins to encroach on the deficiencies inherent in human-human interactions.

Conclusions

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 versus 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 teammates 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 favor of their automated or human teammates as per the situation. The conceptual frameworks of human–decision aid trust introduced in the present paper suggest the benefits of attempting to bridge the gap between human understanding of automated agents versus human teammates. 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 precondition for the smooth functioning of complex systems that have become increasingly sophisticated in recent times.

REFERENCES

- Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, 13 (3), 271-286.
- 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: Lawrence Erlbaum Associates.
- 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.
- Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology*, 18(6), 399-411.
- 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(1), 53-71.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46 (1), 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: The MIT Press.
- Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6, 104-123.
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2003). Automation failures on tasks easily performed by operators undermines trust in automated aids. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: CA.
- Nass, C.L., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45, 669–678.
- Nass, C. L., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56 (1), 81-103.
- Reeves, B., & Nass, C. (1996). The media equation: how people treat computers, television, and the new media like real people and places. *Center for the Study of Language and Information*, Cambridge University Press, Stanford, CA.
- Winograd, T. and Flores, C. (1987). *Understanding Computers and Cognition: A New Foundation for Design*. (Reading, MA: Addison-Wesley).