Available online at www.sciencedirect.com
ScienceDirect

International Journal of Industrial Ergonomics 38 (2008) 608–625

International Journal of
INDUSTRIAL ERGONOMICS
www.elsevier.com/locate/ergon

The impact of cognitive feedback on judgment performance and trust with decision aids

Younho Seong[a,*], Ann M. Bisantz[b]

[a] *Department of Industrial and Systems Engineering, North Carolina A&T State University, Greensboro, NC 27411, USA*
[b] *Department of Industrial Engineering, State University of New York at Buffalo, Amherst, NY 14260, USA*

Received 8 February 2007; received in revised form 13 November 2007; accepted 25 January 2008
Available online 12 March 2008

Abstract

Computerized decision aids are designed to support human operators' decision-making activities in a variety of contexts including medicine, military command and control, and aviation industry. One common characteristic of these systems is of the role of the decision-aid, integrating a variety of measured information to produce a simple form of more meaningful information that can be used to support human operators' judgment about environmental states of interest. When these aids malfunction, the decision makers may ignore the aid due to the lack of trust in the aid. This study examines the effect of automated decision aids of varying quality in producing environmental estimates, and investigates the effect of meta-information in supporting judgment performance of human operators' with the decision aids and calibrating human trust in such aids. A Lens Model based feedback is used to provide meta-information about the decision-aid. An aircraft identification task is performed under varying conditions of aid quality and the presence of meta-information. Results show that performance, as well as assessments of trust in the aid, are affected by the decision aid's quality. More importantly, participants given with the meta-information performed significantly better than those without it. Results indicate that human operators can compensate for a poor performing aid when meta-information is available. Further, operators' trust was better calibrated corresponding to the decision aid's quality. A practical implication of this study is that meta-information can be useful to human operators in increasing their understanding and appropriate utilization of automated decision aids.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Trust in automation; Judgment; Cognitive feedback; Automated decision aids; Meta-information

1. Introduction

In a typical human decision-making task, a human makes a judgment about an environmental state based on data that represent a variety of environmental aspects with the use of technological advances such as sensors or artificial intelligence. This task is challenging because of the uncertainty it involves. To ease the task, some types of automation are developed to support judgments by collecting and processing data to transform it into more "value-added" information.

Automation is an information technology that performs functions or tasks based on information collected about environments with the use of mechanical or computing powers. For example, an automatic pump in a process control plant opens or closes automatically based on the meter readings from other components to regulate the inflow or outflow rate of water to another component. In situations involving more complicated automation, humans receive information to render judgments about environmental states because they do not have direct access to the environment due to a variety of reasons, such as danger, or physical remoteness. The role of automated systems in these kinds of environments is to process information measured from any intermediaries, (i.e., sensors), using computing powers to produce more "value-added," comprehensible information for humans. Therefore, it is obvious that the impact of automated systems on human judgment performance can be dramatic in cases when there are malfunctions in automated systems.

*Corresponding author. Tel.: +1 336 334 7780; fax: +1 336 334 7729.
E-mail addresses: yseong@ncat.edu (Y. Seong), bisantz@eng.buffalo.edu (A.M. Bisantz).

0169-8141/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ergon.2008.01.007

However, humans still have many opportunities to approve or veto automated systems' outputs. They may decide not to trust or rely an automated system based on their belief that the system may be inconsistent or unreliable. Humans may also decide not to trust automated systems even when these systems perform very well, when their trust is misaligned with the system's reliability. It is important to understand how humans' trust in automated systems is affected by automated systems' characteristics, (e.g., reliability), and to identify ways to calibrate human trust according to a system's characteristics.

There is a rich history theoretically as well as empirically in the field of sociology on definitions and role of trust with its importance (Barber, 1983; Rempel et al., 1985). Recently, there are also a few studies of trust in human machine interaction (Muir and Moray, 1996; Lee and Moray, 1992). Previous studies on trust in automated systems revealed that characteristics of automated systems affected human trust in the systems, which in turn impacted utilization of the systems. These studies provided a useful conceptual map existed between automated systems and human operators. However, none of these studies attempted to calibrate human trust so that automated systems could be better utilized. A better utilization of automated systems may lead to better performance in human judgment.

One method to support human judgment performance is the use of feedback. Cognitive feedback refers to providing information pertinent to the relationships between the components in the judgment framework, (e.g., environment, information or cues, and human), and helps humans increase their understanding of automated systems' behavior. Some characteristics of automated systems such as transparency and understandability appear to suggest that information regarding automated systems may be useful in increasing the understanding of the inner workings of the automated systems, which in turn may lead to a better utilization of these systems. Further, this "inside" information can be used to calibrate human trust in these systems. Cognitive feedback is used in this research to achieve two purposes: support of human judgment performance and provision of an opportunity to calibrate trust according to the characteristics of automated systems.

2. Theoretical backgrounds

2.1. Automated decision aids and human judgment performance with aids

Automated decision aids (ADA) refer to the automated systems designed to support human operators in problem solving activities or in the selection of decisions that require an embedded knowledge base (Parasuraman et al., 2000). As decision-making environments become more complex and data intensive, the use of automated decision aids is likely to become even more commonplace and more critical in decision making. The benefits of automated decision aids and expert systems in terms of increased efficiency for data monitoring and analysis capabilities are fairly obvious. However, the behavioral influences of the use of decision-aid must be assessed to properly determine the benefits, limitations, and costs associated with decision-aid development and implementation. Computerized, automated decision aids are designed to support human operators' judgment performance about uncertain environments. In these settings, operational sensors measure information that can be used to identify environmental status and provide direct inputs to human operators and the automated systems. The role of automated decision aids in this process is to integrate the environmental information and to generate environmental estimates using computational data integration algorithms, as in data fusion methods (Waltz and Llinas, 1990).

There are many different kinds of automated decision aids that have been designed to support humans in understanding the environment, diagnosing the situations, or engaging in problem solving activities (Rasmussen et al., 1994). For instance, in a system where human operators do not have any access to the actual state of the environment, information regarding the state of the environment is usually collected via various methods such as sensors or other personnel. It is the automated decision aid's job to integrate the information and provide an estimate of the state of the environment. That is, automated decision aids are defined as providing an estimate of the state of the environment based on the information captured from the environment, in order to support human operators' judgment and decision-making (Morrison et al., 1998). A typical example would be a human operator identifying an unknown object as friendly or hostile, based on the data collected via sensors, along with an identification estimate provided by the automated decision-aid. Together with the informational cues which are measured or sensed from the state of the environment, the estimate can provide valuable information about the environmental state for the human operator making the final judgment.

Research has shown that people are very limited in their ability to process information in uncertain environment, especially so with judgment and decision-making (e.g., Tversky and Kahneman, 1974). Researchers from various disciplines have demonstrated that decision aids can assist the human judges making more accurate and consistent judgments (Morrison et al., 1998). Enhanced judgment performance can be attributed to increment in information processing capacity, or to compensating shortcomings of human judgments, such as biases and anchors. However, other studies have found that the benefits were not uniform because of malfunctions or failures. For example, Will (1991) showed performance decrement of expert's judgments when experts were given falsified information which was consistent with Riley (1996).

One factor that can affect operators' decision-making performance is the complexity of environment. It is obvious that the more complex the environment is, the more difficult the judgment tasks become. Similarly, the complexity of automated decision aids can have impact on operators' understanding of such systems. For example, automated decision aids for the highly structured environment need not to be complex expert systems that require a variety of knowledge built-in (Rasmussen et al., 1994). Simple deterministic type of automated decision aids can perform the task efficiently. Glover et al. (1997) investigated whether the decision-aid could have an effect on users' acquisition of knowledge and their reliance on a decision-aid. They performed an experiment with a computerized determination of tax liabilities, including computation of taxes on net capital gains. The decision-aid used in this study, similar to the actual worksheets used by the government represents a structured task in which participants could be supported by answering initial questions regarding the status of the tax reporter. The decision-aid is more or less a decision template which users have to collect necessary information and enter into the template. Then, the decision-aid provided a suggested solution for the users. Therefore, the information collection and analysis tasks remained with the users to figure out what information to look for and how to draw conclusion from the collected data. Their results were consistent with prior research, indicating that the use of decision-aid enhanced performance.

On the other hand, highly unstructured problem domain where events are intertwined and multiple causes can create multiple stages of events, need not only expert systems with specific knowledge database, but also human experts. Morrison et al. (1998) developed a decision-aid to support human operators in understanding environmental situations, making judgments and planning or employing countermeasures in a complex environment. The information provided in this decision support system was required to make judgments about the environmental state and to evaluate the degree of risk or threat so they could prioritize different activities. For example, the decision support system provided an assessment as to whether or not an unknown object that the human operator intended to identify was a threat along with supporting evidence, counter evidence, and assumptions it was based on. They performed an experiment evaluating the effectiveness of this decision support system and found that the operators reported the system useful for making judgments, and the operators were able to recognize the critical tracks earlier than without the decision support system. Additionally, they found that the operators were able to take more appropriate defensive actions in a timely manner against imminent threats.

When automation is involved in the decision-making process, human decision makers may be negatively impacted in terms of the acquisition of knowledge. Glover et al. (1997) hypothesized that users with decision aids would acquire less knowledge compared to those without decision aids. The decision-aid used by Glover et al. (1997) was a simple and highly structured decision-aid that provided a suggested solution for a problem in a simple calculation of tax liability. They found that the knowledge score of participants, measured by asking them to write about the domain concept, was significantly less for participants with the decision-aid than those without the decision-aid. This result was in line with arguments by Wiener and Curry (1980), who claimed that one of the negative impacts of automation was operators' skill degradation.

2.1.1. Automated decision aids in the current work

In this research, an automated decision-aid typically refers to the types of automation ranging from information acquisition to information analysis according to Parasuraman et al. (2000), which they collectively called information automation. In such systems, data or information collected from the environment is analyzed and fused together to produce further advanced or enhanced information to be useful for human operators making final judgments. Also, it could be categorized as the decision selection type of automation because of the estimates generated by the automated decision aids. For example, the automated decision-aid notifies the human operators of the decision outcomes when it decides to do, or upon human operators' request on demand. Note that the human operators are not involved in the decision-making process. Because of this reason, the automated decision aids in this study is limited to the information automation in which the decision selection is allocated to the human operators.

Specifically, automated decision aids in the current work refer to systems supporting human operators' decision-making activities by integrating sensed information from the environment which the operators do not have any means of access to, and providing sensible estimates of the state of the environments based on a computational algorithm. Therefore, the automated decision-aid collects necessary data or information representing the environment, combines them into an estimate in a more comprehensible form of the environmental state. Therefore, automated decision aids support human operators computation and information integration task, which is otherwise a cognitively challenging task. Each set of information is probabilistically related to the environment which means that the information is not completely diagnostic of the environment, and that the information fusion mechanism plays a critical role in generating valid environmental estimates. This is important because it not only affects the quality of the estimates, but also has an impact on operators' trust in automated decision aids. Consequently, it may further impact the utilization of the automated decision-aid. This conceptualization and flow of information among the environment, decision-aid, and human judge is shown in Fig. 1.

As Endsley and Kaber (1999) described, level 5 automation, decision support, indicates that the computer generates a list of decision options that the human can select from or the operator may generate his or her own options. This level represents of many expert systems or decision support systems that provide options guidance which the human operator may use or ignore in performing a task. The automated decision aids in this current work can be categorized as the level 5 in Endsley and Kaber (1999) taxonomy of the level of automation, and can be any levels between level 2 and level 5 in Sheridan's level of automation. In fact, the level of automation that Endsley and Kaber (1999) described as producing superior overall performance is the same level as the level of automation in this study.

2.2. Trust and its calibration

Studies on human trust in automated systems have been dramatically increasing during the past decade. Since several studies provided an extensive review on trust in sociological perspective (Lee and Moray, 1992; Muir, 1994), this paper will focus on the issue that has not been discussed elsewhere. First is the question of how humans build their trust in automated systems. In other words, what dimensions of trust in automated systems humans examine to decide to or not to trust. Sheridan (1988) provided a list of dimensions or attributes of trust specifically for human machine environment. These are reliability, robustness, validity, transparency, understandability, usefulness, and utility. While one dimension, reliability, has been extensively used by many landmark studies (Lee and Moray, 1992; Muir, 1994) to operationally and successfully manipulate human operator's trust in automated systems, the rest of Sheridan's dimensions were never discussed or used in an attempt to calibrate operator's trust. This is precisely the value of this study by visualizing other dimensions to provide an opportunity for operator to calibrate trust. Therefore, brief examinations of these unused dimensions are necessary.

First is the reliability factor. Briefly, this refers to a system of reliable, predictable, and consistent functioning (Sheridan, 1988). Most of the prior definitions of trust addressed this attribute as the first step in developing trust, based on the premise that a person who behaves in a *consistent* manner will be trusted easily. In Lee and Moray's (1994) experiments, definition of reliability is broader than Sheridan's. While the latter indicates consistent functioning, the former includes the degree to which automatic controlled values are close to the target values as well, which includes the notion of validity which is Sheridan's second dimension. Validity refers to the degree to which it produces correct output. It seems intuitive that the automated system that produces more valid outputs to the human operators will be trusted more.

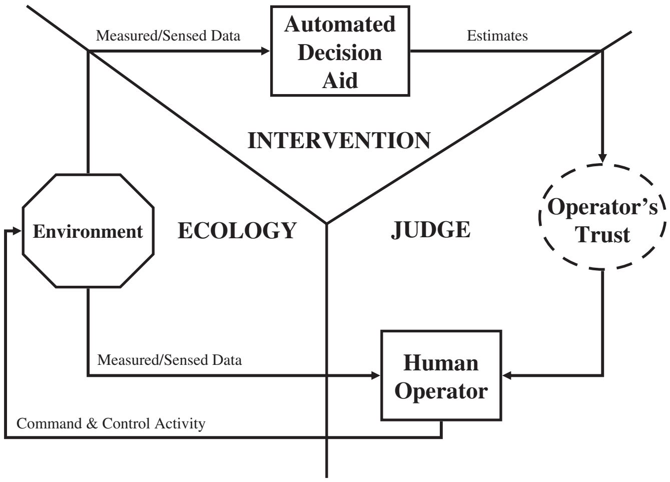


Fig. 1. Conceptualization and flow of information among the environment, decision-aid, and human judge.

Third is robustness. Robustness supports expectations of future performance based on capabilities and knowledge not strictly associated with specific circumstances that have occurred before. If the automated system was designed to handle this situation whether it is (un)expected, it will still provide a useful way to control the situation. Glover et al. (1997) performed an experiment investigating the effect of inferior decision-aid on users' performance in tax calculation domain representing highly structured environment. However, the decision-aid was designed to mainly focus on the "typical" cases. That is, the decision-aid was not robust to consider a variety of cases that can exist in the real world. Their results showed that participants' performance was significantly better with the decision-aid than those without the decision-aid, but only for those typical cases. For non-typical cases, participants without the decision-aid significantly outperformed those with the decision-aid. From the result, it appears that if the decision-aid is not robust to handle many different cases, it is of no use in supporting human operators' judgment and decision-making activities.

Next is the issue of transparency. Transparency refers to the degree to the extent which the inner workings or logic of the automated systems are known to human operators to assist their understanding about the system. Somewhat related to transparency discussed, understandability refers to the operator's understanding of the automation for the expectations that make the operator's trust and use the system appropriately. The operator's intervention will be better and more timely when one's trust is well calibrated to the actual trustworthiness of the system. In designing a machine to aid a human operator, understandability is affected by the degree of transparency of the system which the operator can "see" the underlying mechanisms through the interface. Hanes and Gebhard (1966) conducted a series of studies specifically directing to the question of man–machine interaction. The task setting was a highly realistic simulation of a Navy Combat Information Center in which their subjects were experienced naval commanders. They found that officers were much more willing to accept the computer recommendation if they agreed with the logic used and/or chose the logic themselves, if they understood the computer program or if the recommendations were unambiguously displayed.

Another way to increase the level of operators understanding is to develop an appropriate display interface design. The interface design presumably matches the operators' concept of automated systems, so called mental model. Many studies (e.g., Vicente and Rasmussen, 1992) have claimed that an interface embedded with relationships among the components can enhance operators' understanding, which consequently contribute to increase in performance. On the other hands, Grounds and Wiley (2001) found no significant effect of different display types (configural display, bar graph, and alphanumeric) on operators' decision time to agree or disagree with the recommendations. In the case of a decision-aid which produces the estimate of the environment, there must be a suitable presentation of the uncertainty regarding the system's inferences. This method of presentation becomes the window of opportunity for the operators to reduce the additional uncertainty created by having the automated decision-aid.

Final characteristic is usefulness or utility. The usefulness of data or machines means responding in a useful way to create something of value for operators, eventually developing into trust. Conducting an experiment with two operators as a team within the tactical decision-making environment, Morrison et al. (1998) found that the operators tended to rate those parts of the decision-making support system modules that supported quick decision-making, and thus were more useful, higher than other modules. However, it should be noted that these characteristics are difficult to be hard-coded into a program to manipulate the degree of system's trustworthiness due to their subjective nature.

Sheridan's (1988) list seems comprehensive in that it includes factors beyond those of Muir (1994), or of Rempel et al. (1985), that may have effect on operators' trust indirectly. For example, consider understandability and transparency. These factors may not the ones that operators directly access when they evaluate the level of trustworthiness of an automated system. However, as operators understand the automated system better, they can grasp knowledge about the pros and cons of the system and when to rely on the system. Therefore, it seems that Sheridan's list of trust attributes provide comprehensive understanding of operators' trust and systematic characteristics of trust that may cause operator's trust to develop.

Based on the research framework to study human trust in human–machine interaction setting, several studies performed experiments in a continuous processing domain. Among the prior research, two studies, Muir and Moray (1996), and Lee and Moray (1992), made attempts to identify the role of operator's trust in automated systems. Further, these two studies showed interesting results on the relationship between the operator's trust in automated systems and their reliance on the system measured by the use of the system. For example, Muir and Moray (1996) performed a process control simulation experiment and the results supported this perspective, showing a positive regression coefficient between operator's trust and the use of automation. They also concluded that as the level of operator's trust suffers, their use of automated system decreases. Lee and Moray (1992) conducted very similar experiments and showed that the operators tended to use less of the automated system as the level of trust decreases.

A behavioral counterpart of operators trust in an automated system is an index of reliance, use of automation. This is based on the assumption that the more operators trust, the more they use. Riley (1996) performed an experiment using a simple computer game which participants had to control various simple multiple tasks part of which they could relinquish the control to automated system. Automation reliability (90% and 50%) was controlled by setting the probability that the automation would perform a task correctly. The automation faults were distributed over time and the experimental conditions, which were different in the level of workload and in the level of uncertainty. The level of uncertainty in a task was controlled by introducing abnormalities so that the automated system could not perform the task appropriately. The level of workload was also controlled by making another separate task more complicated. The results showed that participants demonstrated a bias toward manual control, and that both college student and pilots did not delay turning on automation after a failure, and continued to rely on failed automation. Pilots showed greater tendency toward the use of automation under the lower risk. In a subsequent experiment focusing on the dynamics of trust, Riley controlled the level of information that participants were given about the automation prior to performing the experiment. These conditions differed in whether participants were provided with information regarding automation reliability, state uncertainty, or both to reveal the contributions of each element to their automation use decisions. He found that both state uncertainty and trust affect automation use decisions, but only early in the participants' experience with automation.

Although those research projects concentrated on human trust in automation based on the general understanding of supervisory control tasks, the importance of the trust concept seems applicable and has been attempted to define the role of trust in other domains, such as computer supported cooperative work (Jones and Marsh, 1997; Christianson and Harbison, 1997), decision-making in management (Lerch and Prietula, 1989), medical diagnosis expert system (Moffa and Stokes, 1996), computer security problems (Beth et al., 1994), trust in e-commerce (Gefen, 2000), industrial inspection systems (Jiang et al., 2004), in-car navigation systems (Ma and Kaber, 2007).

2.2.1. *Research and models of trust in ADA*

Specific empirical research on operators' trust in automated decision aids is somewhat more limited than general research on trust in automated systems. Several studies used different types of decision aids that might have some relevance with the issues in this study. Glover et al. (1997) conducted an experiment investigating the relationship between the decision aids and the humans. Among the results, they showed that the users' performance was significantly better with the decision-aid than without the decision-aid. This was only true for those typical cases. Reverse was the truth for those cases that were not typical. The result means that participants were able to perform the task better without the decision-aid for the cases for which the decision-aid was not designed and was not expected to encounter. Also, it indicates that participants relied on the decision-aid inappropriately for those cases that the decision-aid could not perform. In conclusion, decision aids that are inferior for non-typical cases can cause operators to inappropriately rely on the automation, which consequently results in the low level of performance.

For instance, one study by Dzindolet et al. (1999) investigated different aspects of automated decision aids on operators' identification performance and revealed that participants decided to ignore the decisions by the decision-aid when no feedback was provided, and when they were told that the decision-aid was actually human rather than an automated aid, regardless of the quality of the decision-aid (better than or equal to participants performance). However, the percentage of participants who decided to ignore the support from the decision-aid when they were told that the decision-aid was actually an automated decision-aid was also not promising, regardless of the presence of the feedback or the quality of the decisions. The average percentages of participants who decided to ignore the decision-aid were 83% and 81% for the feedback and no feedback condition, respectively.

Lerch and Prietula (1989) conducted an experiment consisted of traditional financial management decision problems to investigate effect of inferior quality advice on people's judgments. As might be expected, there was a loss of trust in the agent resulting from faults in the advice, and the recovery of trust was slower than that in performance. They called this "inertia", similar to what others have called a "hysterisis loop". They also found that it was more difficult for humans to recover trust after once it was lowered because of the failures than to build trust initially. In Lerch and Prietula (1989), the level of performance measured by the level of confidence in the decisions the participants made deteriorated after the wrong advice, and never returned to the level of performance where it was before the wrong advice, even at the end of trials.

Finally, Bisantz and Seong (2001) defined different types of failures that could occur in information acquisition and information analysis types of automation, acknowledging that systems could fail due to different intents. For instance, system components or automation may fail due to unintentional hardware or software difficulties. In some environments, failures may be introduced intentionally, through acts of sabotage. They conducted an experiment investigating effects of different types of failure on operators' decision-making performance and trust in two informational sources. One informational source, information window, provided sensed information by various sensors while the other, decision-aid window, provided the environmental estimate of an object's identity. Participants were told that failures due to hardware or software failures or sabotage are possible in addition to the third condition where the probability of decision-aid fault was not mentioned. Note that operators could select whether to consult with the automated decision-aid. Results showed significant impact of fault condition on information window use. Participants in the sabotage condition seemed least likely to select the information window on a track-by-track basis, and participants in the control condition the most likely, while across the three conditions there was a trend toward decreased information window use. They concluded that there appeared to be a trend for participants in the sabotage condition to make less use of the information window than participants in the other two conditions. Also, participants showed a trend across conditions to make less use of the information window and more use of the decision-aid window, indicating that participants attributed the failure to the information window rather than the decision-aid window.

Finally, characteristics of decision aids themselves can impact their utility. Although the importance of automated decision aids in assisting human operators' judgment and decision-making tasks has been emphasized over decades, there has been relatively research directly investigating how decision-aid quality impacts human performance. Will (1991) investigated false dependence on technology, performing an experiment with reservoir petroleum engineers in which misleading suggestions were provided about the appropriate methodology to address a given problem situation. Falsified reasoning explanations in an expert information system were provided to investigate their impact on the decision confidence and performance depending on the level of expertise. Will found no significant difference in decision confidence, and concluded that experts, in general, did not identify the errors in the expert system.

Investigation of trust in automated system has suggested that trust plays an important role in influencing operators' strategies toward the use of automation (Lee and Moray, 1994; Muir and Moray, 1996; Riley, 1996). For instance, trust is related to concepts of automation use and disuse (Parasuraman and Riley, 1997). Misuse refers to inappropriate reliance on automation, because the operator fails to understand the capabilities of the automated system. Disuse implies the non-use of automation when in fact its use is appropriate. In both cases, decisions to rely on automation are linked to operators' understanding of the capabilities of the automated system. A related concept is one of calibration operators' assessments and decisions regarding automated system are considered to be well calibrated when they appropriately use, or not, aspects of automation based on the automation's capabilities.

2.3. Lens Model based feedback: Cognitive feedback

Based on these findings, an important question with respect to the performance of automated decision aids is the degree to which operator's can evaluate the performance of the aid, in terms of its ability to provide appropriate and useful advice. Cohen et al. (1998) emphasize the need for situation specific training to enhance calibration, so that operators' understand the capabilities of automated systems in context. The means of information presentation may also impact operator's ability to appropriately use decision aids. For instance, Grounds and Wiley (2001) conducted an experiment to investigate the effect of different display interface design on operators' trust in the decision-aid. They found no significant effect of different display types (configural display, bar graph, and alphanumeric) on operators' decision time to agree or disagree with the recommendations. On the other hand, several other researchers have found effects of information presentation (Crocoll and Coury, 1990; Lee et al., 1999; Yeh and Wickens, 2001). For example, Yeh and Wickens (2001) performed an experiment focusing on the effect of providing intelligent cuing information to guide attention to targets, and reliability of the automated attention-directing device. They found that the presentation format of cuing significantly impacted the detection accuracy of low-salience objects. However, presentation did not help participants in detecting accuracy of expected, high-salience targets.

A related question is the type of information which may be useful in helping operators' diagnose the capabilities of a decision-aid. Outcome feedback, feedback about whether advice given by the system is correct or not, is one form of such information that could be used in operator training, or testing a decision-aid. However, research on human judgments has indicated that outcome feedback is less effective than other forms of feedback information in improving their own decisions (Balzer et al., 1992). Research on cognitive feedback (see Balzer et al., 1989 for an extensive review) has suggested that more detailed feedback about aspects of judgments, and characteristics of the judgment situation, may lead to improved judgment performance. For instance, information about the judgment environment, such as the manner in which available information (judgment cues) and the situation to be judged are related (i.e., variable weights in regression equations linking available cues to the true state of the situation, as well as the regression coefficient associated with the equation), termed *task information*; or the performance of the judges themselves (i.e., the manner in which judges are combining cues, and making consistent judgments), termed *cognitive information*, may lead to better judgment performance. In this research, the concept of cognitive feedback was adapted to apply to the evaluation of decision aids, as shown in Fig. 2. In this case, cognitive feedback referred to measurements of how the decision-aid—rather than a human judge—produced the estimates. Thus, in this case, "cognitive information" referred to feedback regarding how the decision-aid combined information about the environment to produce its estimates, while task information referred to the way in which that same information was related to the situation. Functional validity information, shown in Fig. 2, refers to the overall performance of the aid, or the degree to which its estimates are correct.

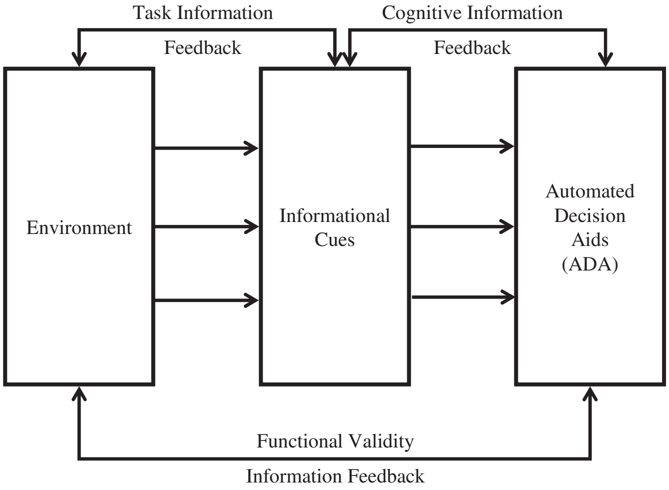


Fig. 2. Three types of cognitive feedback according to Balzer et al. (1992).

Linking to the Sheridan's list of characteristics of automated systems, it seems plausible that cognitive feedback can be a useful source of information to provide deeper understanding of human on how the systems produce estimates for the environment states. Specifically, the cognitive information feedback showing the relationship between the information or cues in the middle and the decision aid's estimates provides parameters such as the weighting scheme of the decision-aid and consistency on generating the estimates as a means to be more transparent or understandable.

2.4. Research hypotheses

Based on previous research on characteristics of automated decision aids and their impact on performance and aspects of trust, the following research hypotheses are addressed in this research:

1. Judgments, when made with inputs from automated decision aids, will depend on the quality of information provided by the aid.
2. Additionally, participant's subjective expressions of trust in automated decision aids will vary (i.e., be calibrated) based on characteristics of the aid.
3. Meta-information, or feedback about the performance of a decision-aid, will be useful in helping participants calibrate their understanding of the aid's performance, and thus impact their judgments performance. Specifically, participants with aid performing poorer will make judgments which are less consistent with the aid, and

thus perform better when provided with meta-information about the aid.

4. Participant's trust in an aid will also be better calibrated when provided with meta-information about aid performance.

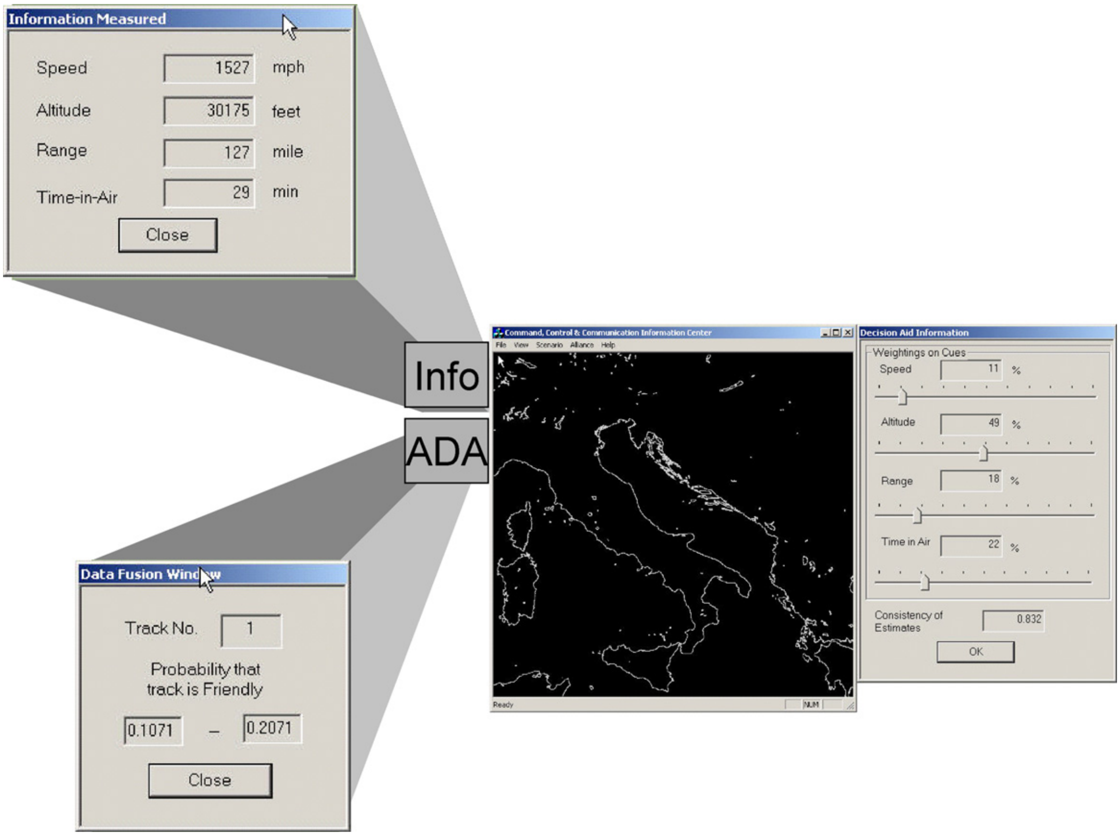
The first two hypotheses are generated to confirm the previous research on human judgment performance and trust in conjunction with automated systems in a different setting, while the rest of the hypotheses are generated within the notion of supporting judgment performance and calibrating their trust in conjunction with automated decision aids. The last two hypotheses are established to test the utility of the meta-information to provide a source for better understanding of ADA which consequently lead to better trust calibration.

3. Method

3.1. Experimental task

An aircraft identification task was chosen in this study. Participants were asked to identify aircraft moving on a simulated radar screen as either hostile or friendly. Participants were given the actual values for the selected aircraft of four parameters: speed, altitude, range, and time in air before making a judgment. Parameter values were presented in a fixed order in a window, called the "Information Measured" window, shown in Fig. 3.

Participants were provided with sets of probabilities before the onset of the experiment. These probabilities indicated the probability of the selected object being friendly given the specific level of each cue, thus indicating that cue values were probabilistically related to an aircraft's identity as hostile or friendly. For example, if the speed of an unknown object was specified between 500 and 1200, it indicated that the probability that the object was friendly was 0.89, while if the speed was between 1100 and 1800, the probability it was friendly was 0.18. Note that the ranges of values overlapped, thus increasing the uncertainty associated with the judgment. Scenarios were constructed so that the identified in the study were given parameters which corresponded to the probabilities given in the training materials. Also, the probability of each aircraft being friendly given the parameters was calculated based on the entire population of aircrafts used in the study. In some conditions, participants were also provided with estimates from an automated decision-aid. These were shown in a separate window, called the Data Fusion Window, and indicated the degree of probability that the selected aircraft was a friendly. Finally, in some conditions, participants were provided with meta-information regarding the performance of the decision-aid. Further details on these manipulations are provided below.


Information Measured

Speed 1527 mph

Altitude 30175 feet

Range 127 mile

Time-in-Air 29 min

Data Fusion Window

Track No. 1

Probability that track is Friendly 0.1071 – 0.2071

Command, Control & Communication Information Center

Ready

Decision Aid Information

Weightings on Cues

Speed 11 %

Altitude 49 %

Range 18 %

Time in Air 22 %

Consistency of Estimates 0.832

Fig. 3. Display screens used in the experiment including the four windows.

3.2. Participants

Fifty-six participants were recruited from the student population at the State University of New York at Buffalo. Participants were paid $6.50 for their participation. An analysis of demographic data indicated that participants had similar ages, time in college, and statistics background across conditions of the aircraft identification task. There were 41 male and 15 female participants. Participants' average age was 24.61 years. There was no significant difference in age between the experimental conditions of the aircraft identification task ($F_{6,49} = 1.489$, $p = 0.202$). Participants' average number of years in college was 4.98. Again, there was no difference in years in college between the seven experimental conditions ($F_{6,49} = 0.435$, $p = 0.852$). The average number of undergraduate or graduate statistics courses taken by participants was 1.11. Again, there was no significant difference in courses between the seven experimental conditions ($F_{6,49} = 0.738$, $p = 0.621$).

3.3. Simulation and display

Participants performed this task using a computer simulation that was written using Visual C++. The simulation was implemented on a Pentium II computer running Windows NT, and was displayed on a 17-in. computer monitor. The display consisted of a radar display, a window for the informational cues, the decision aid's estimate window, and an understandability window showing information about the decision aid's performance (see Fig. 3). The radar window displayed a map on which unknown objects appeared at random time intervals. Participants could click on objects to obtain information or identify them. Objects were represented using military style symbology. Initially, objects appeared as unknown; after identification, they appeared as hostile or friendly. Additionally, if the operator's judgment was incorrect, a line over the normal symbol appeared.

The information window (placed on the upper left side in the screen) contained four cues: speed, altitude, range, and time in air. This window was provided in all experimental conditions. Information in the window changed as participants selected different objects on the radar display. The ADA window (on the bottom left) provided two numbers indicating the lower and the upper limits of a probabilistic estimate that selected object was friendly. These numbers were generated by applying the Bayesian reasoning to the actual probabilities of the cues. This window was provided to participants (in some conditions) on the second day of the experiment. The quality of information provided in this window also varied, depending on the experimental condition. Finally, in some conditions, participants were provided with an understandability window, which provided meta-information about the performance of the decision-aid. The type of information provided is related to cognitive feedback information (Balzer et al., 1992), as described above. In particular, the window shows cue relative weights (utilization validities) corresponding to the judgments being provided by the aid, and a measurement of the degree to which the aid is making judgments consistent with a linear model. The cue utilization validities of the automated decision-aid, shown in the upper part of the window, represent the weighting scheme employed by the decision-aid in generating an estimate. A sliding bar indicated the relative weight that was applied on that specific cue. Information regarding the reliability of the aid was provided through a consistency measurement, which represented the relationship between the actual environmental estimates of automated decision-aid and the statistically predicted estimates of the environmental estimates of automated decision-aid.

Additionally, this field of information in the understandability window (consistency of estimates) provides an indication of the validity of the aid's estimates, because participants could compare the decision aid's weighting scheme with the probability structure provided in the instructions. The information in this window was first shown after participants made 10 identifications. Updates on these parameters were made after every five identifications.

3.4. Independent variables

Three independent variables were controlled to investigate their effects on human operators' judgment performance: the validity of estimates provided by the decision-aid, the reliability of estimates provided by the decision-aid, and the presence of meta-information regarding the decision aid's performance. Meta-information, related to concepts of cognitive feedback, was described in the preceding section. Validity was defined here as the extent to which the aids' estimates appropriately and suggested that the aircraft was hostile or friendly. An estimate was valid if the confidence interval provided by the aid spanned over the actual probability that the aircraft was friendly. An estimate was invalid when the confidence interval spanned over the "opposite" probability. For instance, if the actual probability of the aircraft being hostile equaled 0.7, an invalid estimate spanned 0.3. This transformation was applied to every estimate in the low validity conditions. Validity was operationally measured as the correlation between the midpoint of the interval provided by the aid, and the actual probability that the aircraft was friendly.

Additionally, reliability was defined here as the extent to which the decision-aid provided consistent estimates of the friendliness of aircraft having similar but identical within the predefined categories information parameters (e.g., altitude, speed, etc.). Note that each type of information is categorical. Therefore, even though the actual information may be different, the probability remains identical as long as the information is categorized in the same category, which consequently produces the same estimate. Reliability was manipulated by completely changing the estimates to the opposite end, i.e., from 0.8 to 0.2, for the high reliability condition (HR). On the other hand, for the low reliability condition (LR), only a part of each set of the same estimates was switched to the opposite end. Note that this partial manipulation of the estimates caused the level of validity for the low reliability condition higher than that for the high reliability condition. Reliability was operationally measured as the correlation between the decision aid's estimates given the set of information.

Each variable had two levels: high and low, producing a total of 8 ($2\times2\times2$) possible conditions. However, the combination of high validity and low reliability was excluded because of its infeasibility—such a condition would require that the estimates should accurately indicate the state of the unknown aircraft (have high validity), but be inconsistent (have low reliability). Thus, there were three combinations of validity and reliability, as shown in Table 1. Table 1 also gives the operational parameters used to quantify the levels of validity and reliability of the aid, as defined in this experiment. All three combinations of validity and reliability levels were tested both with and without the meta-information, for a total of six decision-aid conditions. Throughout this text, conditions are specified using "H" for the high, "L" for the low level, "V" for validity, "R" for reliability, and "+U" for the provided understandability information.

It is important to note that while conceptually, the levels of validity and reliability were crossed to form the three possible levels, the levels of validity differed across the two low validity conditions. In the low validity/high reliability (LVHR) condition, estimates were consistently poor—and thus validity, as measured in Table 1, was negative (indicating that it tended to give an estimate that was consistently opposite of the true value). In the low validity/low reliability (LVLR) condition, estimates were inconsistent—not all of the estimates were consistently poor—and thus (somewhat counter-intuitively) had a higher measured level of validity.

3.5. Procedure

Participants performed three scenarios per day for 2 consecutive days. Each scenario had 50 unknown aircrafts to be identified. On the first day, participants were provided with the written task instructions which were read aloud, and were randomly assigned to one of the seven condition including the control condition. There were eight participants per condition. Then, a brief training exercise session designed to familiarize participants with the identification task was performed with the aid of the experimenter. On the first day, all participants performed the identification task without being assisted by the automated decision-aid. On the second day, participants performed the same three scenarios. However, depending on the experimental condition, additional window(s) were provided. One group performed the task in the same format of the first day (control condition). The remaining six groups performed the task using one of the six decision-aid conditions: high validity/high reliability (HVHR); LVHR; LVLR, and with or without the meta-information as shown in the understandability window (+U or −U). A 12-item questionnaire was provided to participants in the six decision-aid conditions, at the end of each of the three sessions, to rate their trust in the decision-aid (Jian et al., 2000). Questionnaire items are shown in Fig. 4.

Table 1
Combinations of validity and reliability present in the decision aid

System characteristic	Validity	High		Low	
	Reliability	High	Low	High	Low
Correlation (ENV–ADA)		0.757		−0.757	0.381
Cognitive control ($R_{s(ADA)}$)		0.736		0.736	0.409
Consistency ($R_{c(ADA)}$)		0.914		0.914	0.409

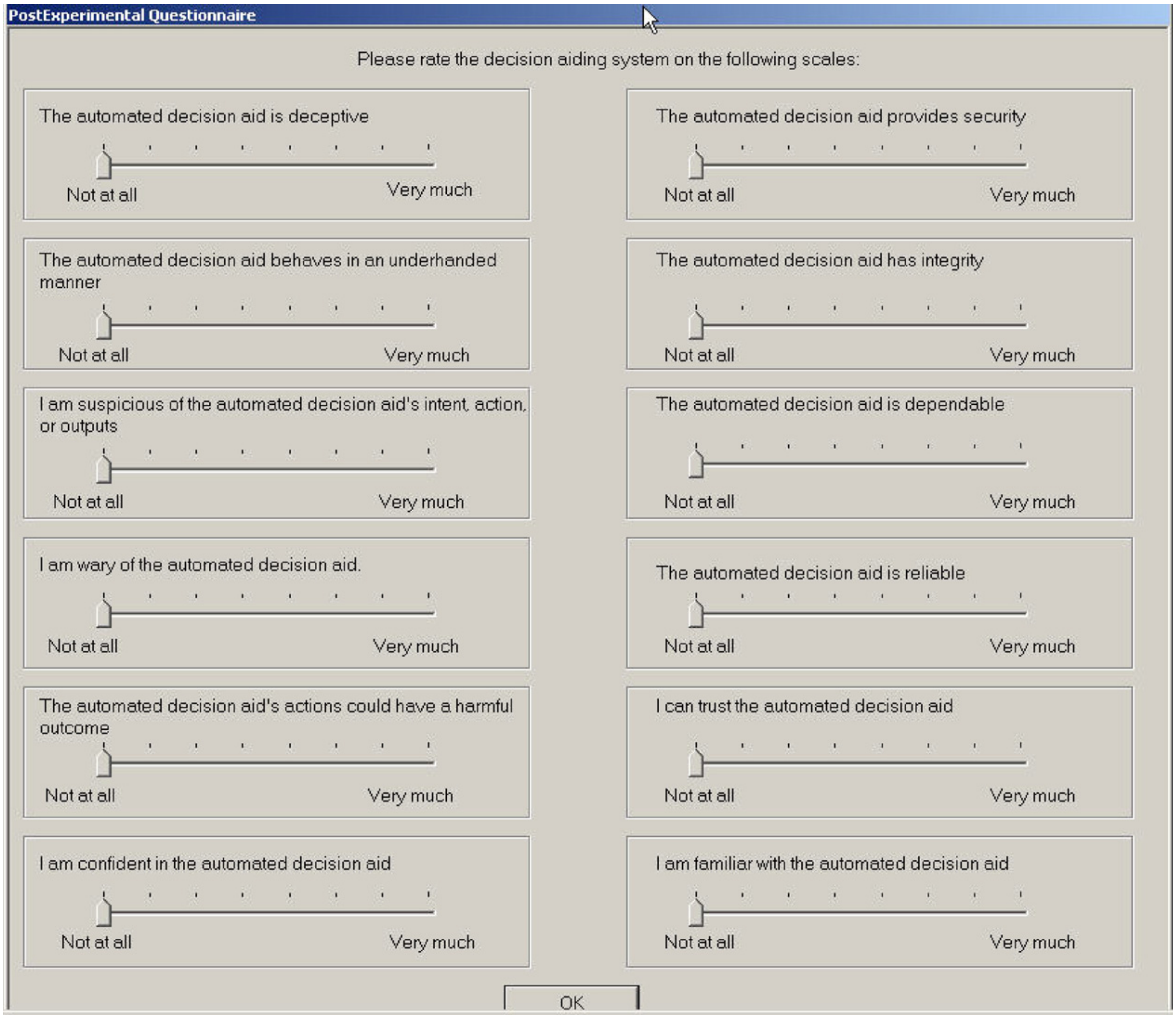
4. Results

4.1. Identification performance

The first dependent measure examined was accuracy. Figs. 5 and 6 depict the number of correct identifications on Day 1 and 2, respectively. Overall, participants performed the identification task well. On Day 1, participants were able to identify the unknown contacts successfully, ranging from 79% to 83%, depending on the conditions. These differences among the conditions were not significant, indicating that the experimental groups were homogeneous ($F_{6,49} = 1.569$, $p = 0.176$) in performing the identification task when provided with only the measured information. Note that the environmental predictability measured by the correlation between the environmental states to be judged and the informational cues that the judgments are based was 0.745, 0.64, and 0.66 for scenario 1, 2, and 3, respectively. The environmental predictability is an index to indicate the degree of linear predictability of the environment based on the given information.

Also, scenario was not significant ($F_{2,98} = 9.363$, $p = 0.176$), indicating that the different levels of the environmental predictability, or experience, did not influence the participants' accuracy. Based on these two results, we conducted an analysis of variance (ANOVA) on the Day 1 identification performance, collapsing across the scenario variable. This analysis confirmed that there was no significant difference among the experimental groups ($F_{6,49} = 1.845$, p=0.110), and provides further evidence that participants across the experimental groups were homogeneous in performing the task.

In contrast, participants' identification performance levels were distinctly different in Day 2, as shown in Fig. 6. An ANOVA on the correct identification score on the Day 2 showed that there was a significant difference among the experimental groups ($F_{6,49} = 8.892$, $p < 0.001$).



PostExperimental Questionnaire

Please rate the decision aiding system on the following scales:

The automated decision aid is deceptive
Not at all — Very much

The automated decision aid behaves in an underhanded manner
Not at all — Very much

I am suspicious of the automated decision aid's intent, action, or outputs
Not at all — Very much

I am wary of the automated decision aid.
Not at all — Very much

The automated decision aid's actions could have a harmful outcome
Not at all — Very much

I am confident in the automated decision aid
Not at all — Very much

The automated decision aid provides security
Not at all — Very much

The automated decision aid has integrity
Not at all — Very much

The automated decision aid is dependable
Not at all — Very much

The automated decision aid is reliable
Not at all — Very much

I can trust the automated decision aid
Not at all — Very much

I am familiar with the automated decision aid
Not at all — Very much

OK

Fig. 4. Twelve item trust scale, adapted from Jian et al. (2000).

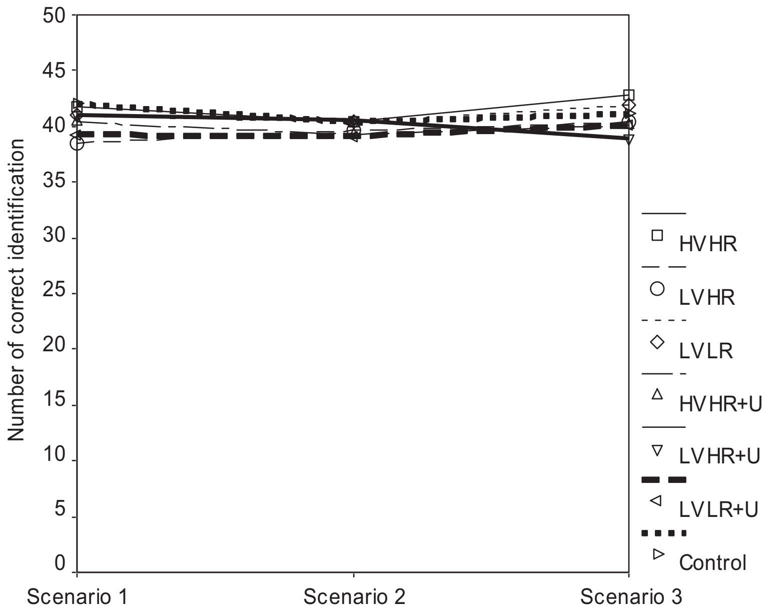


Fig. 5. Identification performance on Day 1.

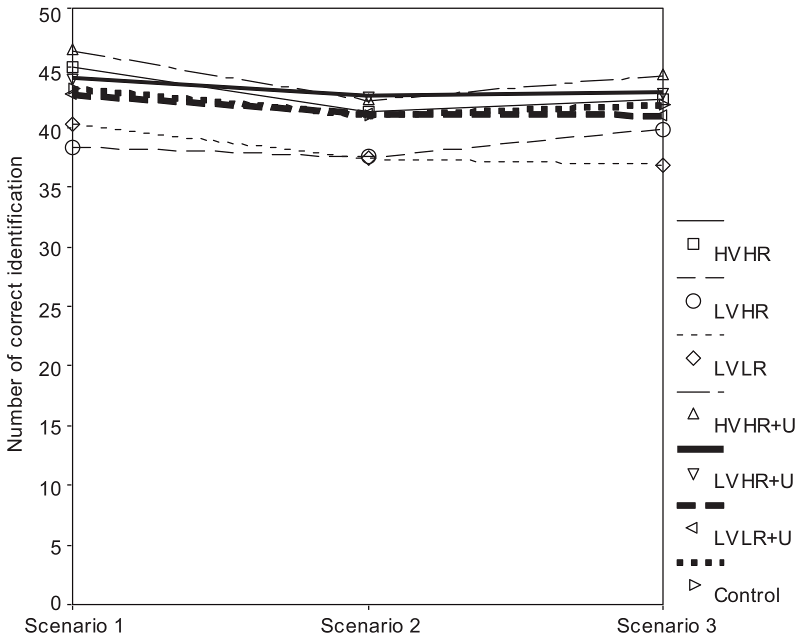


Fig. 6. Identification performance on Day 2.

There was no significant scenario by group interaction ($F_{12,98} = 1.276$, $p = 0.245$). Thus, it seems that participants were differentially influenced by combination of the different levels of quality of the decision-aid and the additional understandability information provided. Close examination of Fig. 6 showed that two groups with the "poor" decision-aid were segregated from the rest of the groups, suggesting that quality of the decision-aid producing the environmental estimates had a large negative impact on participants' judgment, but also that providing additional information of understandability neutralized the large negative influence of the "poor" automated decision-aid. A post hoc analysis on the differences among the experimental groups was conducted, and revealed that participants with a higher quality decision-aid (HVHR) performed the task better than those in either condition with a lower quality decision-aid (LVHR or LVLR), and also that performance with the low quality aid was worse than the control (unaided) condition. This indicates that the low quality decision-aid had a negative impact on participants' identification performance (see Table 2 for mean differences between the groups and the levels of significance). Additionally, there were significant differences between the conditions with the automated decision-aid and those with additional understandability window. As shown in Fig. 6, there were significant differences between the LVHR and the LVHR + U conditions, and between the LVLR and the LVLR + U conditions. Therefore, providing additional information about the behavior of the automated decision-aid, namely the cognitive information feedback, resulted in increasing the identification performance. There was no significant difference between the HVHR and the control condition, or the HVHR + U condition. This indicates that when the decision-aid performed the estimation well, providing additional understandability information did not contribute to increase participants' judgment performance.

Finally, there was no significant difference among the three groups with the additional understandability information, indicating that participants with the "poor" decision-aid along with the additional understandability information were able to utilize the understandability information to evaluate the decision-aid and perform the identification task comparable to those with the "good" decision-aid together with the understandability information.

Additionally, an ANOVA on the number of correct identifications per day, treating the day as a within-subject variable, showed that there was a significant difference in identification performance between Day 1 and 2 ($F_{1,49} = 11.217$, $p = 0.002$), a significant difference between the experimental groups ($F_{6,49} = 4.715$, $p = 0.001$), and a significant Day × Group interaction effect ($F_{6,49} = 7.471$, $p < 0.001$), again indicating that type of decision-aid, and the presence of the understandability information, significantly affected participants' judgment performance. These results are demonstrated in Fig. 7, which shows the change in the number of correct identifications, from Day 1 to Day 2, across the six types of decision aids, and the control condition. Between the groups without the understandability window, post hoc analysis revealed that there is a significant difference between HVHR and LVLR (mean difference = 4.25, $p = 0.026$). Further, there is a significant difference between LVHR and HVHR (mean difference = -5.33, $p = 0.002$), and between LVHR and LVHR + U (mean difference = -4.00, $p = 0.043$). Also, significant differences were found between LVLR and the three groups with the understandability window (mean differences

Table 2
Post hoc analysis on the identification performance on Day 2

PERF	HVHR	LVHR	LVLR	HVHR + U	LVHR + U	LVLR + U	Control
HVHR	*	4.375 (0.005)	4.792 (0.001)				
LVHR		*	−5.750 (0.000)	−4.667 (0.002)	−3.458 (0.046)		
LVLR			*	−6.167 (0.000)	−5.083 (0.001)	−3.458 (0.046)	−3.875 (0.017)
HVHR + U				*			
LVHR + U					*		
LVLR + U						*	
Control							*

Mean differences between the conditions, along with significance levels, are reported. Only significant differences are shown.

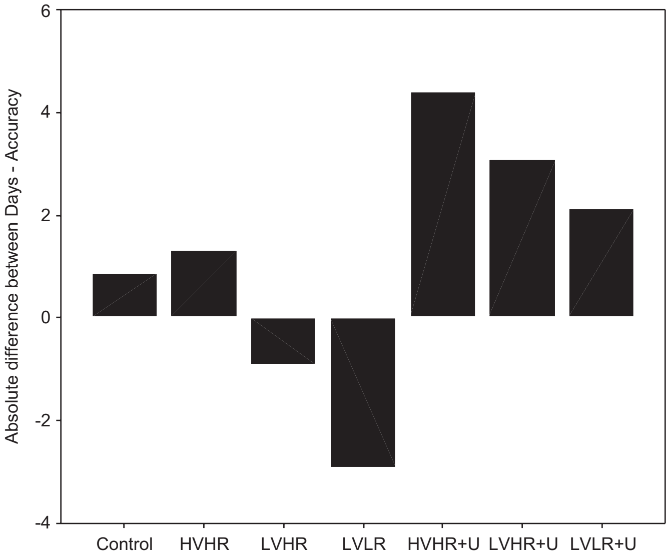


Fig. 7. Change in identification performance across the 2 days.

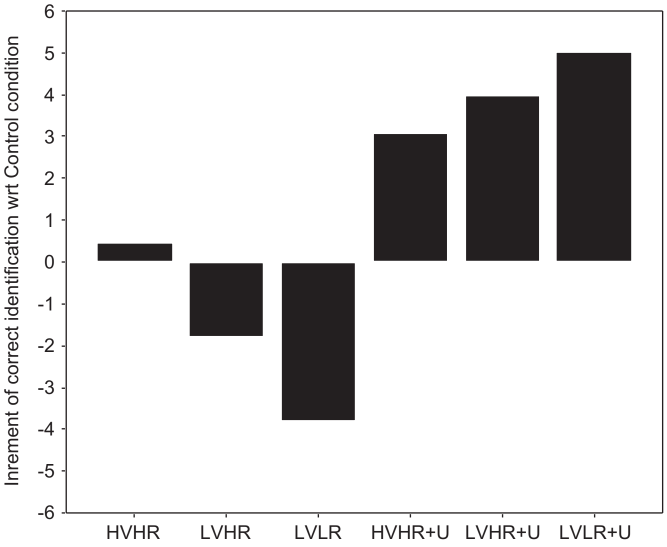


Fig. 8. Degree of benefit of having the automated decision-aid or the understandability information along with the decision-aid.

were −7.33 ($p<0.001$), −6.00 ($p<0.001$), and −5.04 ($p = 0.004$) for HVHR + U, LVHR + U, and LVLR + U, respectively). Note that no significant differences were found between the three groups with the understandability window.

Also of interest was the impact of the decision-aid independent of the learning that occurred from Day 1 to Day 2. The increase of performance in the control group is due to learning. Therefore, taking the improvement of performance in the control group as a reference, the benefit of having the decision-aid or the understandability information along with the decision-aid was calculated by subtracting the performance increment of the control group from those groups with the decision-aid only. Likewise, to investigate the incremental value of the understandability window, the differences between corresponding groups with and without the window were computed. These results are demonstrated in Fig. 8. Results showed that there was a significant difference between the groups ($F_{6,49} = 4.417$, $p<0.001$). A post hoc analysis was performed to identify significant differences among the experimental groups. Results showed that participants with the "good" decision-aid (HVHR) improved more than those with the decision-aid producing poor estimates inconsistently (LVLR) within the three groups only with the decision-aid. Significant differences were also detected within the corresponding pair of the decision-aid and the understandability window (i.e., LVHR vs. LVHR + U). In both cases of the "poor" automated decision-aid (LVHR or LVLR), the understandability window led to greater incremental performance than the decision-aid alone. This indicates that participants were able to utilize the information provided to increase their understanding of the automated decision aid's behavior, and then used that information to make correct judgments. As seen in Fig. 8, the influence of the decision-aid on participants' performance becomes greater with either "poor" decision-aid. However, the benefit of having additional understandability information becomes greater when the decision-aid was both less reliable and less valid. For example, the benefit of the additional information became greater in the LVLR condition than in the LVHR condition due to the poor level of performance in the LVLR condition.

4.2. Correspondence between judgments and decision aid's estimates

Finally, to evaluate the correspondence between participant's judgments and those of the decision-aid, a correlation was performed between the unaided decisions made on Day 1, and the aided decisions (of the same unknown aircraft) on Day 2. These correlation coefficients from participants in different experimental conditions were compared using an ANOVA on transformed parameters values. Cooksey (1996) describes an appropriate transformation for normalizing the parameters (which are correlations) as Fisher's r to z transformation, where r corresponds to the parameters, and z the transformed parameters:

$$Z_{\mathrm{r}} = \frac{1}{2}\log_{\mathrm{e}}\left[\frac{1+r}{1-r}\right].$$

After the analyses of variance, reverse transformation was applied to calculate the meaningful correlations for interpretation:

$$r = \frac{\mathrm{e}^{2Z_{\mathrm{r}}}-1}{\mathrm{e}^{2Z_{\mathrm{r}}}+1}.$$

Results showed that there was a significant difference between the groups ($F_{5,42} = 463.19$, $p<0.001$). As seen in Fig. 9, the two groups with the "good" decision-aid showed a higher correlation (top two lines in the figure). There was no apparent effect of the understandability information when the decision-aid performed the task well. On the other hand, the groups with the "poor" decision-aid had a lower correlation in conditions where they were provided with the understandability window, indicating that the meta-information about the aids performance allowed participants to better calibrate their use of decision-aid. Results of the post hoc analysis revealed that every pair except for the HVHR and HVHR + U pair of groups showed significant difference. Specifically, the mean difference between LVHR and LVHR + U was 0.2085 ($p<0.001$), while that between LVLR and LVLR + U was 0.142 ($p = 0.004$).

4.3. Subjective assessments of trust

Finally, participants' responses to items on the trust questionnaire were analyzed. Responses were collected via computer, on a seven-point scale, and scaled from 0 to 70. Average responses to the positively framed questions (items 6–11), the negatively framed questions (1–5) and the single question on familiarity were analyzed separately. Scenario was treated as a repeated measure. ANOVA results showed significant differences in responses to positively framed questions ($F_{2,42} = 13.37$, $p<0.001$), negatively framed questions ($F_{5,42} = 7.122$, $p<0.001$), and familiarity ($F_{5,42} = 10.810$, $p<0.001$) based on condition, as shown in Fig. 10. Post hoc (least significant difference) tests showed significant differences to positively framed questions between conditions with the good decision-aid and conditions with the poorer decision-aid. Also, there was a significant difference in positive ratings of trust between the HVHR and HVHR + U conditions. Responses to negatively framed questions were significantly lower for the two good decision aids than the four conditions with poor decision aids. For familiarity, post hoc tests showed significant differences among conditions with and without the meta-information, for the HVHR and LVHR, conditions.

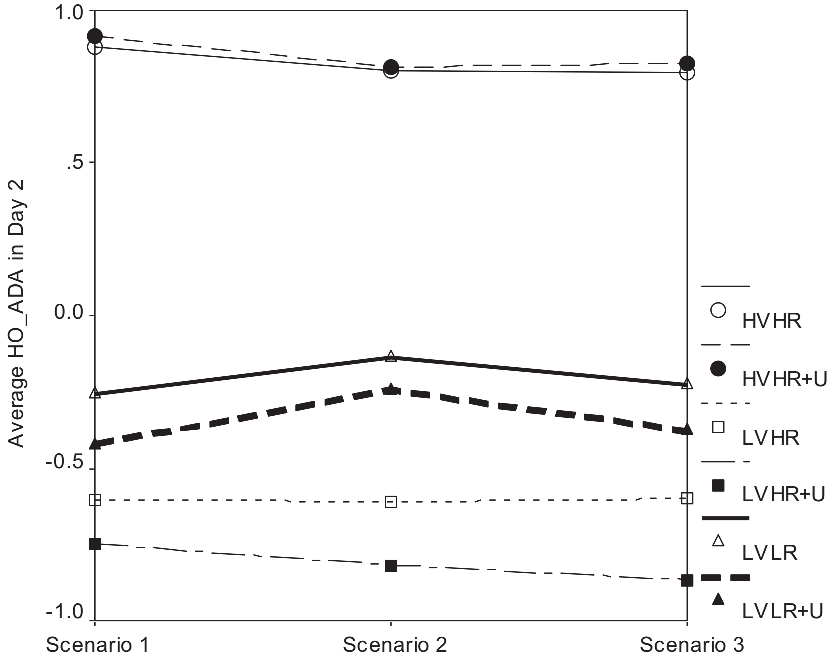


Fig. 9. Average correlation between participants' judgments and the decision aid's estimates.

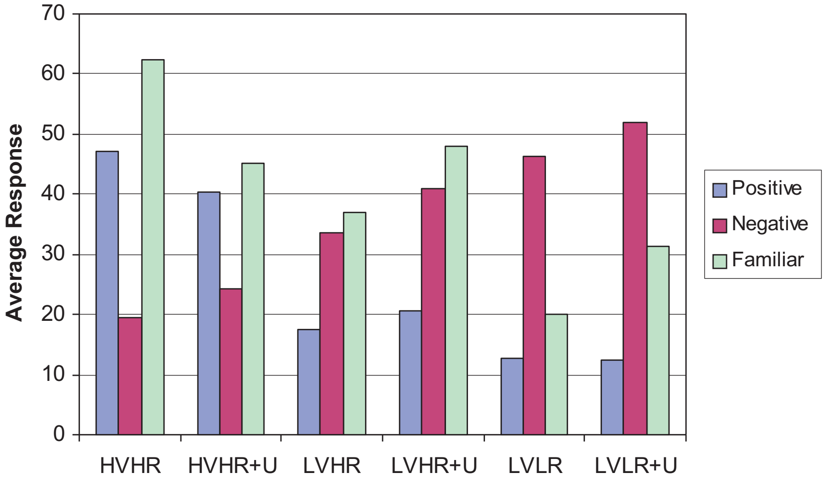


Fig. 10. Subjective ratings of trust, measured by positively framed questions, negatively framed questions, and a question regarding familiarity.

5. Discussion and conclusions

The goal of this research was to increase our understanding of how people make judgments in complex environments, particularly supported with automated decision aids. There are two research problems that arise from this domain of interest. One is to focus on how people make judgments with automated decision aids providing processed information or estimates about the environmental states. Previous research on human judgment and decision-making has focused on how people make judgments given the equivalent level of information in terms of the number of stages involved to process information. However, ADAs process the raw data to transform into a more comprehensible form, the environmental estimates, for human judges. Therefore, a human judge's judgment policy may be different, considering the outputs from the automated decision aids more heavily than other informational cues. Thus, it is necessary to understand for us to the effect of automated decision aids on human judges' judgment policy and performance.

The other focus stems from the first point above. An empirical measure for the degree of effect of automated decision aids on human judges' performance is their reliance on the systems. Many factors can be influential in determining the extent to which human judges rely on automated decision aids, including risk involved, and time pressure. Among them, human judges' trust in automated systems is one potential causal factor: I rely on it because I trust it. Therefore, another research problem is to investigate how people's trust in automated systems change over time and over the quality of automated systems. Previous research indicated that operators' performance using automated systems to control the entire system are impacted by the reliability of control systems, which affect operators' trust in such systems.

Another important aspect that must be considered is that human operators' understanding of the automation has been hypothesized as a necessary condition for the development of trust in automated systems. These understandings include understanding of automated systems' capability, their own ability to control or make judgments, and how the automated systems perform the required tasks. However, research from this perspective has depended on "all-or-nothing" automated systems which human operators need to engage in "automatic mode" to determine whether the automated systems perform the task correctly. This can be a critical factor in a domain where risks involved with human operators' actions or judgments are high or irreversible, so that only well tested and trusted automated systems will be implemented to support human operators' judgments.

Previous research has made specific claims about characteristics of automated systems which may impact trust (Sheridan, 1988). The research presented here is unique in that in manipulated, three of these factors: validity, reliability, and understandability, and measured, the impact of these changes on trust as well as decision performance. While other work has investigated the impact of failures and errors in automated control systems (e.g., Muir and Moray, 1996; Lee and Moray 1994), this research is unique in that it applied similar concepts to decision-aiding technology. Additionally, researchers have suggested a link between calibration of trust and appropriate use of automated systems: this study addresses the relationship between judgment performance, trust, and similarity of judgments with that of the aid. However, little attempt has been made to study the calibration of human operators' trust in automated decision aids.

Findings from this research are significant in three areas. First, the study demonstrated the impact of decision-aid characteristics, including meta-information, on judgment performance. Second, the study showed that decision-aid characteristics also impacted assessments of trust in the aid. Finally, the research extended the application of cognitive feedback to automated judgments.

5.1. Impact of decision-aid characteristics on judgment performance and reliance on the aid

Important research questions addressed by this work concerned the extent to which characteristics of an automated decision-aid impact judgment performance in concert with the aid. Three specific characteristics of decision aids were tested: the validity of the estimates, the reliability of the estimates, and the presences of explanatory or meta-information about the aids' performance. As indicated by measures of judgment correspondence, when the aid was performing well, participants' judgments were similar to the aid. Judgments were less similar for poor performing aids, indicating some calibration of the participants' reliance on the aid. However, judgment performance (as measured by number of correct judgments) still suffered when the aid had poor validity or poor validity and reliability. Judgment correspondence decreased, and judgment performance increased, for the poorer performing aids, when meta-information about the aids' performance was provided. Thus, results indicated that while decision performance may prove sensitive to the reliability and validity of the aid, the negative impact of this sensitivity can be mitigated through the provision of diagnostic information about the aid itself.

5.2. Impact of decision-aid characteristics on trust

Results in the previous section indicated significant differences in ratings of trust aspects, across different decision-aid conditions, thus supporting the second hypothesis that trust would vary based on characteristics of the decision-aid. Several patterns shown in Fig. 9 are of interest, and demonstrate the degree to which participants were able to calibrate their ratings of trust with characteristics of the system. First, notice that positive ratings of trust were generally higher for the better decision aids and lower for the negative decision aids. As expected, ratings of negative aspects of trust were higher for the poorer performing aids, than the better aids. Interestingly, conditions with the meta-information showed a non-significant tendency toward higher negative trust ratings, than the parallel conditions without the meta-information, perhaps because participants were more aware of the shortcomings of the aid. This provides some indication of support for our fifth hypothesis, that assessments of trust would be better calibrated with the presence of meta-information. Finally, ratings of familiarity were higher when participants were provided with the meta-information, for the two poorer performing aids, indicating that the meta-information was increasing participant's understanding of the device. Examination of results for the good aid is interesting. Negative ratings of trust in the good aid increased with meta-information, and familiarity decreased. Because the meta-information showed that the aid was not performing perfectly, trust in even the good aid suffered, indicating that participants were able to calibrate their judgments of trust even about the quality of a generally good decision-aid.

A related, important aspect of this research was the consideration of links between judgment performance, assessments of trust, and reliance on the aid. Positive assessments of trust were lower, while negative assessments were higher for poorer performing aids. If this result reflected a direct correspondence between trust levels and reliance on the aid (that is, if trust were the sole or primary mediating factor influencing use of the aid), then one would expect little or no performance differences across experimental conditions. When the aid was poor, and participants lacked trust in it, they should have ignored the aid, and performed similarly to the first day (without the aid). This, however, was not the case. Participants' performance only recovered from the poor decision-aid when meta-information was present. Subjective assessments of positive trust factors were not influence by this change, while negative assessments showed a non-significant trend toward calibration. Thus, while assessments of trust appeared sensitive to changes in decision-aid characteristics regarding aid quality, with less of an impact of meta-information, judgment performance was primarily impacted by the presence of meta-information, and correspondence between judgments and the estimates produced by the aid showed some impact of both aid quality and meta-information.

5.3. Application of cognitive feedback to decision-aids and future considerations

A final important contribution of this research is the application of cognitive feedback as an indication of decision-aid, rather than human, performance. Extensive past research on cognitive feedback has focused on the role such information can play in judgments made by individuals. Here, the "cognitive information" component of cognitive feedback was used to provide human decision-makers with information regarding the functioning of a decision-aid. Specifically, multiple regression computations were performed relating the estimates produced by the decision-aid, with the values of the informational cues (here; speed, altitude, range, and time-in-air) available about the aircraft. Participants were provided with the resultant cue weights from the regression, as well as the regression coefficient. Thus, participants received information about the manner in which the decision-aid was combining environmental information, as well as how reliably (according to a linear model) that it was doing so. Past researchers have suggested that operators' ability to understand how automation is working may impact their trust in the automation (Sheridan, 1988). Poorly calibrated reliance on automation (misuse and disuse) stem from misunderstandings regarding how well an automated aid might perform in a particular circumstance (Parasuraman and Riley, 1997). As demonstrated here, aspects of cognitive feedback may prove useful in the provision of such information to human operators, for a particular form of automated system: automated judgment and decision aids.

However, this study lacks of several important aspects that need to be addressed to fully provide a good understanding of the process of trust calibration. First, the study did not examine the effect of different types of interface display on human trust in such systems. As briefly discussed previously, there have been some studies showing changes of human interaction behavior with automated systems (i.e., Vicente and Rasmussen, 1992; Grounds and Wiley, 2001). Studying the effect of interface display can be tied with the issue of transparency which was one of the key dimensions addressed in this study. It is feasible that a configural display of critical information can be better understood compared to an alphanumeric based display, which eventually may lead to better calibration of human trust.

Additionally, the issue of usefulness or utility is not examined in this study. Again as briefly discussed, it is somewhat difficult to implement this characteristic to a hard-coded software to manipulate human trust in automated systems due to its subjective nature. Also, it may be inherited from human's previous experience with similar systems, in which the role of tacit knowledge needs to be investigated further.

Acknowledgments

The authors would like to acknowledge the National Science Foundation under Grant no. IIS9984079 to the second author for support of this work. Also, the authors would like to thank Colin Drury and James Llinas for their contributions to this research, and John Lee for thorough review.

References

Balzer, W.K., Doherty, M.E., O'Connor Jr., R.O., 1989. Effects of cognitive feedback on performance. Psychological Bulletin 106, 410–433.

Balzer, W.K., Sulsky, L.M., Hammer, L.B., Sumner, K.E., 1992. Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? Organizational Behavior and Human Decision Processes 53, 35–54.

Barber, B., 1983. The Logic and Limits of Trust. Rutgers University Press, New Brunswik, NJ.

Beth, T., Vorcherding, M., Klein, B., 1994. Valuation of trust in open network. In: Gollman, D. (Ed.), Computer Security-ESORICS 94, Third European Symposium on Research in Computer Security. Springer, Brighton, UK, pp. 3–18.

Bisantz, A.M., Seong, Y., 2001. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. International Journal of Industrial Ergonomics 28, 85–97.

Christianson, B., Harbison, W.S., 1997. Why isn't trust transitive? In: Lomas, M. (Ed.), Security Protocols Proceedings of the International Workshop. Springer, Cambridge, UK.

Cohen, M.S., Parasuraman, R., Freeman, J.T., 1998. Trust in decision aids: a model and its training implications. Technical Report, Cognitive Technologies, Inc., Arlington, VA.

Cooksey, R.W., 1996. Judgment Analysis: Theory, Methods, and Applications. Academic Press, New York.

Crocoll, W.M., Coury, B.G., 1990. Status or recommendation: selecting the type of information for decision aiding. In: Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting, vol. 2, pp. 1524–1528.

Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A., 1999. Automation reliance on a combat identification system. In: Paper presented at the Human Factors and Ergonomics Society 43rd Annual Meeting.

Endsley, M.R., Kaber, D.B., 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42 (3), 462–492.

Gefen, D., 2000. E-commerce: the role of familiarity and trust. Omega 28, 725–737.

Glover, S.M., Prawitt, D.F., Spilker, B.C., 1997. The influence of decision aids on user behavior: implications for knowledge acquisition and inappropriate reliance. Organizational Behavior and Human Decision Processes 72 (2), 232–255.

Grounds, C.B., Wiley, D., 2001. Alternative methods of mitigating automation distrust impacts. In: Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting, Santa Monica, CA, pp. 1728–1732.

Hanes, R.M., Gebhard, J.W., 1966. Information requirements for the control of combat forces. In: Albanes, J.S. (Ed.), Reports of the Committee on Vision, 1947–90. National Academy Press, Washington, DC.

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.

Jiang, X., Khasawneh, M., Master, R., Bowling, S.R., Gramopadhye, A.K., Melloy, B.J., Grimes, L., 2004. Measurement of human trust in a hybrid inspection system based on signal detection theory measures. International Journal of Industrial Ergonomics 34 (5), 407–419.

Jones, S., Marsh, S., 1997. Human–computer–human interaction: trust in CSCW. Ergonomics International 85, 172–174.

Lee, J.D., Moray, N., 1992. Trust, control strategies and allocations of functions in human–machine systems. Ergonomics 35 (10), 1243–1270.

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

Lee, J.D., Gore, B.F., Campbell, J.L., 1999. Display alternatives for in-vehicle warning and sign information: message style, location, and modality. Transportation Human Factors Journal 1 (4), 347–377.

Lerch, F.J., Prietula, M.J., 1989. How do we trust machine advise? In: Slavendy, G., Smith, M.J. (Eds.), Designing and Using Human–Computer Interface and Knowledge Based Systems. Elsevier Science Publishers, Amsterdam, pp. 410–419.

Ma, R., Kaber, D.B., 2007. Effects of in-vehicle navigation assistance and performance on driver trust and vehicle control. International Journal of Industrial Ergonomics 37, 665–673.

Moffa, A.J., Stokes, A.F., 1996. Trust in a medical system: can we generalize between domains? In: Mouloua, M., Koonce, J.M. (Eds.), Human–Automation Interaction: Research and Practice, Cocoa Beach, FL, pp. 218–224.

Morrison, J.G., Kelly, R.T., Moore, R.A., Hutchins, S.G., 1998. Implications for decision-making research for decision support and displays. In: Cannon-Bowers, J.A., Salas, E. (Eds.), Making Decisions Under Stress: Implications for Individual and Team Training. American Psychological Association, Washington, DC, pp. 375–406.

Muir, B.M., 1994. Trust in automation. Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37 (11), 1905–1922.

Muir, B.M., Moray, N., 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics 39 (3), 429–460.

Parasuraman, R., Riley, V., 1997. Human and automation: use, misuse, disuse, abuse. Human Factors 39 (2), 230–253.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics 30(3), 286–297.

Rasmussen, J., Pejtersen, A.M., Goodstein, L.P., 1994. Cognitive Systems Engineering. Wiley, New York.

Rempel, J.K., Holmes, J.G., Zanna, M.P., 1985. Trust in close relationship. Journal of Personality and Social Psychology 49 (1), 95–112.

Riley, V., 1996. Operator reliance on automation: theory and data. In: Parasuraman, R., Mouloua, M. (Eds.), Automation and Human Performance: Theory and Application. Lawrence Erlbaum, Mahwah, NJ, pp. 22–27.

Sheridan, T.B., 1988. Trustworthiness of command and control systems. In: Paper presented at the IFAC Man–Machine Systems.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Vicente, K.J., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics SMC-22, 589–606.

Waltz, E., Llinas, J., 1990. Multisensor Data Fusion. Artech House, Norwood, MA.

Wiener, E.L., Curry, R.E., 1980. Flight-deck automation: promises and problems. Ergonomics 23, 995–1011.

Will, R.P., 1991. True and false dependence on technology: evaluation with an expert system. Computers in Human Behavior 7, 171–183.

Yeh, M., Wickens, C.D., 2001. Display signaling in augmented reality: effects of cue reliability and image realism on attention allocation and trust calibration. Human Factors 43, 355–365.