A SITUATION SPECIFIC MODEL OF TRUST IN DECISION AIDS[1]

Marvin S. Cohen
Cognitive Technologies, Inc.

ABSTRACT

Trust in decision aids can be modeled as a n *argument* regarding the expected performance of the aid. Such a model helps identify the grounds, temporal scope, granularity, risk, and other parameters of trust judgments that underlie interaction decisions at different phases of decision aid use.

Keywords: trust, decision aids, automation, rationality, situation awareness, event trees

SITUATION SPECIFIC TRUST

There is renewed interest in the development of computerized decision aids as key components of the "digitized" battlefield. Decision aids vary in the level of automation they offer — data display and fusion systems, expert systems that recommend options, associate systems that take action unless overridden, and autonomous systems that choose and act without informing the user. However, automation of decision making has not advanced as far or as rapidly as might have been expected. Explanations for this failure tend to center around the issue of trust. According to some advocates, lack of acceptance is caused by irrational technophobia, evidenced by a *lack* of appropriate trust. On the other hand, there are experienced officers who reject decision aids because of the danger of *overtrust*: the tendency to rely on an aid in place of one's own judgment. Significantly, these camps share the same underlying concept of trust, as an enduring attitude toward an aid, akin to love, hate, or faith, rather than a context-sensitive attitude, such as agreement or disagreement.

We will describe an alternative, more differentiated conception of trust. It includes enduring attitudes as a special case, but emphasizes instead variable judgment about the specific conditions under which an aid will and will not perform well. According to the Situation Specific Trust (SST) model, the problem of decision aid acceptance is neither undertrust nor overtrust as such, but *inappropriate* trust: a failure to understand or properly evaluate the conditions affecting good and bad aid performance. To the extent that decision aid acceptance has foundered on the issue of trust, training deserves some of the responsibility. Training focuses on inputting required information, changing modes, and reading outputs. Such training inadvertently reinforces the misconception that trust must be a invariant stance, to accept or reject an aid as a whole. There has been little effort to teach skills for evaluating an aid's performance in real time, and training strategies for interacting with the aid based on that evaluation.

In research sponsored by the Aviation Applied Technology Directorate, ATCOM, Fort Eustis, VA, we developed a systematic and general framework for training users of decision aids. We applied and tested the framework by developing a training strategy for a specific decision aiding environment, the Rotorcraft Pilot's Associate. More details may be found in Cohen, Thompson, Masalonis, Bresnick, Parasuraman, Hess, & Schwartz (2000).

[1] This research was supported by Contract No. DAAH10-98-C-0022 with the Aviation Applied Technology Directorate, U.S. Army Aviation and Missile Command, Fort Eustis, VA. Thanks to Keith Arthur, our technical monitor, and to others who participated in various aspects of the research: Bryan Thompson, Raja Parasuraman, Daniel Serfaty, Steve Hess, Terry Bresnick, and Jeffrey Schwartz. Thanks also to Anthony Masalonis for his contribution to understanding and testing (as well as naming) the Situation Specific Trust model.

BACKGROUND RESEARCH

Lee and Moray (1992, 1994) and Muir and Moray (1996) showed that subjective assessments of trust (1) discriminated appropriately among different levels of system performance and (2) were correlated with operator's use of automation. Muir (1987, 1994) introduced a theoretical framework for trust in machines, borrowing and integrating concepts from social research on trust among humans. The SST framework to be described here was motivated, in part, by two basic observations on this work, which is reviewed in Cohen, et al. (2000): (1) The current models describe different categories of trust, but do not specify fundamental principles that account for the various categories or ensure that they are exhaustive. (2) The categories conflate increasing knowledge of the system with increasing trust, as if that were the only legitimate outcome of experience, and do not allow for decreasing trust or more differentiated trust. Muir's model (1994) postulates that the evolution of trust culminates in blind "faith." In addition, and perhaps connected to this, we noted a fairly pervasive view, explicit or implicit, that actual trust judgments are deeply flawed. For example, Lee and Moray (1992) and Muir and Moray (1996) argued that trust does not change appropriately as more information is acquired; Lee and Moray (1994) argue that decisions about the use of automation lag even further behind changes in trust. Muir (1987) claims that humans are biased toward distrust. Lee and Moray (1992) and Muir and Moray (1996) concluded that trust is not based solely on system performance but is reduced by irrelevant factors like the occurrence of a fault or performance variability even when overall performance is unaffected.

Let us stipulate that trust is not perfectly rational (however we interpret that concept). Nevertheless, taxonomic models of trust do not provide any direct tools for assessing rationality, and do not uphold the specific claims that have been made. On the contrary, the empirical findings are consistent with the following, quite different conclusions: (1) Operators draw on a variety of cues regarding *future* system performance, and do not focus exclusively on *past* performance. These include indicators of system quality such as system variability and the very fact that any flaw has occurred. Operators also use their own compensating actions as "cues" regarding improved system performance. (2) Operators can learn to discriminate finely among levels of on-going system performance (as cues of future performance) by means of graded trust judgments, and they do so relatively quickly, e.g., within a one-hour session. (Intra-session changes have sometimes been overlooked.) Operators maintain these graded discriminations over the course of their experience with the systems, i.e., they do not simplify by reducing all trust assessments to 0 or 1. (3) The degree of trust is affected by the costs of errors. Trust is lower, the more serious the consequences of mistakes. It follows that trust cannot be regarded simply as a prediction of future performance; outcomes are weighted by their desirability. Often, but not always, overtrust has worse consequences than undertrust. The prior probabilities of automation success and failure may also have a legitimate influence on trust. (4) The relationship between trust judgments and the magnitude of system errors may be both non-linear and relative to a baseline expectation. If so, surprising events, whether positive or negative utility, will have the most impact. That is, additional evidence of poor system performance will have diminishing effect the more distrust there already is; and novel system *successes* will have diminishing effect the more trust there is. Such non-linearity may facilitate the role of trust in regulating operator interventions to improve system performance. (5) Interaction decisions are influenced by the difference between trust and self-confidence (as argued by Lee and Moray, 1992). But they are also influenced by costs in time and effort. Costs are likely to be incurred by (a) exploratory behaviors, (b) decisions to shift between automated and manual operation, (c) decisions to monitor the performance of automation, and (d) decisions to verify system recommendations. (6) Faith can be interpreted as reliance on assumptions. Assumptions are inevitable in any prediction of complex system behavior, such as trust, but they vary in importance. Operators may be more likely, on the average, to rely significantly on assumptions early in system use, when a system is applied in novel conditions, and when the costs of gathering information about the system are high. Dependence on assumptions may be associated with relatively low-cost exploratory behavior to find strategies for automation use that avoid such dependence.

TRUST AS AN ARGUMENT

The Situation Specific Trust model has a qualitative, informal core, and quantitative, prescriptive extensions. We will focus here on the qualitative core theory, which describes a mental model that organizes the relationships among components of trust. This structure is based on a schema for arguments in real-world domains (Toulmin, 1958). It identifies the current observations that provide reasons for or against trust in a particular system, the degree of trust based on those reasons, the long-term beliefs underlying the reasons, the sources of information supporting those beliefs, and the potential risks or uninvestigated assumptions associated with the conclusion. Parameters associated with these trust components help characterize a space that differentiates different categories of trust in a systematic way. SST allows us to model the evolution of trust in time and the interactive decisions based on trust. Figure 1 shows the elements of the model: (1) *Warrant:* A Warrant is a general belief that certain conditions are associated with a certain quality of aid performance. The conditions that the user has learned to associate with aid performance may be features or combinations of features of the system, situation, task, or aid conclusions. (2) *Grounds*: To influence an argument for or against trusting an aid, via a Warrant, a feature or combination of features must be observed on the particular occasion for which trust is being assessed. Grounds are the accumulated observations, on a specific occasion, of conditions that influence a user's judgment of the reliability of an aid. Grounds may include any relevant cues regarding future performance for which there is a Warrant, such as the reputation of the aid designer, the past performance of the aid, the fact that the user is able to override it, and specific interaction strategies adopted by the user. (3). *Qualified Claim:* The output of the trust model is a Qualified Claim. This Qualified Claim represents the degree of trust in the system under the conditions specified in the Grounds, according to the Warrant. Qualifications may be precise (e.g., 80%) or, as is more often the case, vague (e.g., very reliable). Degree of trust in a system is the expected desirability of system actions over a given period of time in the future, conditional on the Grounds. (4) *Backing:* Backing is the type of knowledge source or learning method by which a Warrant was acquired. Users can learn what factors influence trust by actual experience of the aid under variations of the relevant conditions, by inference from knowledge of the design of the aid, by testimony of other users, by generalizing from other aids or from other kinds of automation, by making explicit or implicit assumptions (e.g., best case or worst case), or even by assuming that the aid is like a human. (5) *Rebuttals:* Rebuttals are conditions that potentially invalidate the inference from Grounds to Claim, i.e., they are possible exceptions to the rule expressed by the Warrant. A rebuttal is the possibility that implicit or explicit assumptions are false, for example, one's past experience with the aid may not be representative of present conditions, or the aid design may be flawed.

SST allows us to chart how trust varies from one user to another, from one decision aid to another, across phases of decision aid use, and with experience. SST tracks such changes with four parameters.

Granularity. Granularity represents users' ability to discriminate among situations that are associated with differences in aid performance. For example, one might ask how well a target identification aid will hold up as the number of targets is increased, how well an expert system performs in knowledge-based (or novel) tasks as opposed to rule-based (or routine) tasks, or how well a planning aid performs on problems in which there are tradeoffs among goals. To count toward Granularity, a feature must be present in both the Grounds and the Warrant of a trust claim. The granularity of the Warrant reflects the user's long-term knowledge and understanding; it is equivalent to the number of system and situational features that the user has learned or believes to affect trust, and which are pertinent in the present time and place, whether observed by the user or not. Granularity of the grounds reflects the portion of that knowledge that is in fact being utilized in the operator's current situation awareness, i.e., the features that are actually observed and discriminated by the user. Granularity is a matter of effective pattern recognition, and associated

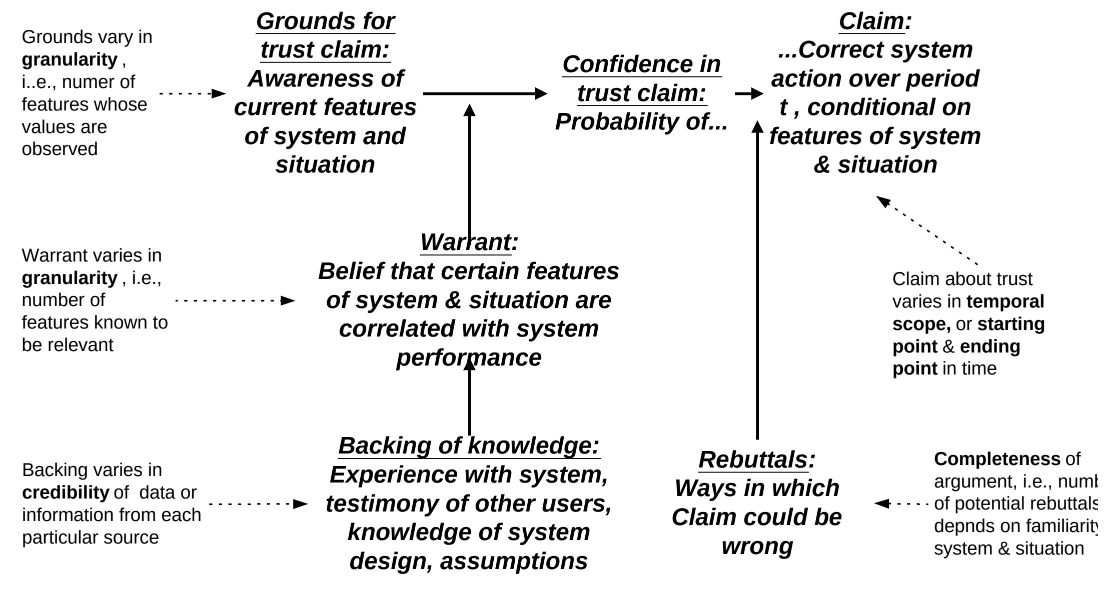
Grounds vary in **granularity** , i..e., numer of features whose values are observed
Grounds for trust claim: Awareness of current features of system and situation
Confidence in trust claim: Probability of...
Claim: ...Correct system action over period t , conditional on features of system & situation
Warrant varies in **granularity** , i.e., number of features known to be relevant
Warrant: Belief that certain features of system & situation are correlated with system performance
Claim about trust varies in **temporal scope,** or **starting point** & **ending point** in time
Backing varies in **credibility** of data or information from each particular source
Backing of knowledge: Experience with system, testimony of other users, knowledge of system design, assumptions
Rebuttals: Ways in which Claim could be wrong
Completeness of argument, i.e., numl of potential rebuttals depnds on familiarity system & situation

Figure 1. Situation Specific Trust model components and parameters.

predictions of success or failure.

Credibility. The credibility of the Backing is influenced by the quantity, representativeness, and consistency of the information that underlies the Warrant. The *quantity* of information in the Backing reflects the amount of experience a user has with the aid in the specific situation relevant to the trust claim, or the accuracy of the design knowledge the user has that pertains to that situation. The *representativeness* of information in the Backing is higher the more varied the situations that the user has experienced, acquired design knowledge for, or made assumptions about. The *consistency* of the information in the Backing is higher to the extent that the different sources of information agree with one another. Conflict among different Backings is a symptom of wrong assumptions, and if identified, may actually promote learning. For example, a conflict between assumptions about design and actually experienced performance may lead a user to understand the design better, and realize that the aid's performance under one set of circumstances (for which it was not optimized) differs from its performance in other environments. Training that includes alertness to such discrepancies may improve the appropriateness of trust and the effectiveness of decision aid use. Although credibility is important, in real-world contexts, the major obstacle to appropriate trust is learning to discriminate relevant features (i.e., granularity), not fine-tuning the precise importance attached to each. In quantitative terms, this means that *resolution* of situations in which trust varies is more important than *calibration* of estimates of degree of trust.

Incompleteness. The incompleteness of an argument is the number of important potential rebuttals to the trust argument, i.e., assumptions which have not been confirmed. This number cannot be known directly, but it varies with the degree to which the decision aid is difficult to operate or understand, unfamiliar to a user, or is being operated under novel conditions. The less familiar the system or the more novel the situation, the more likely the trust argument is flawed and its conclusion incorrect.

Temporal scope. Temporal scope is the span of time that a trust claim covers. Trust might reflect a prediction of the performance of the system across its operational lifetime, over the next hour, or simply with respect to its current conclusion. We can distinguish at least four prototypical phases in the use of a decision aid, corresponding to decreasing temporal scope: (0) trust in people, different types of systems, automation, or decision aids generally, prior to gaining any specific information about the particular system in question; (1) trust in a particular system over all its potential uses, before a specific mission has been assigned or a specific task has been undertaken; (2) trust in the system's capability for a specific mission or task in a specific context; (3) trust in the current recommendation by the aid. Each phase comprises a set of cycles that are embedded in the next higher phase.

THE DYNAMICS OF TRUST

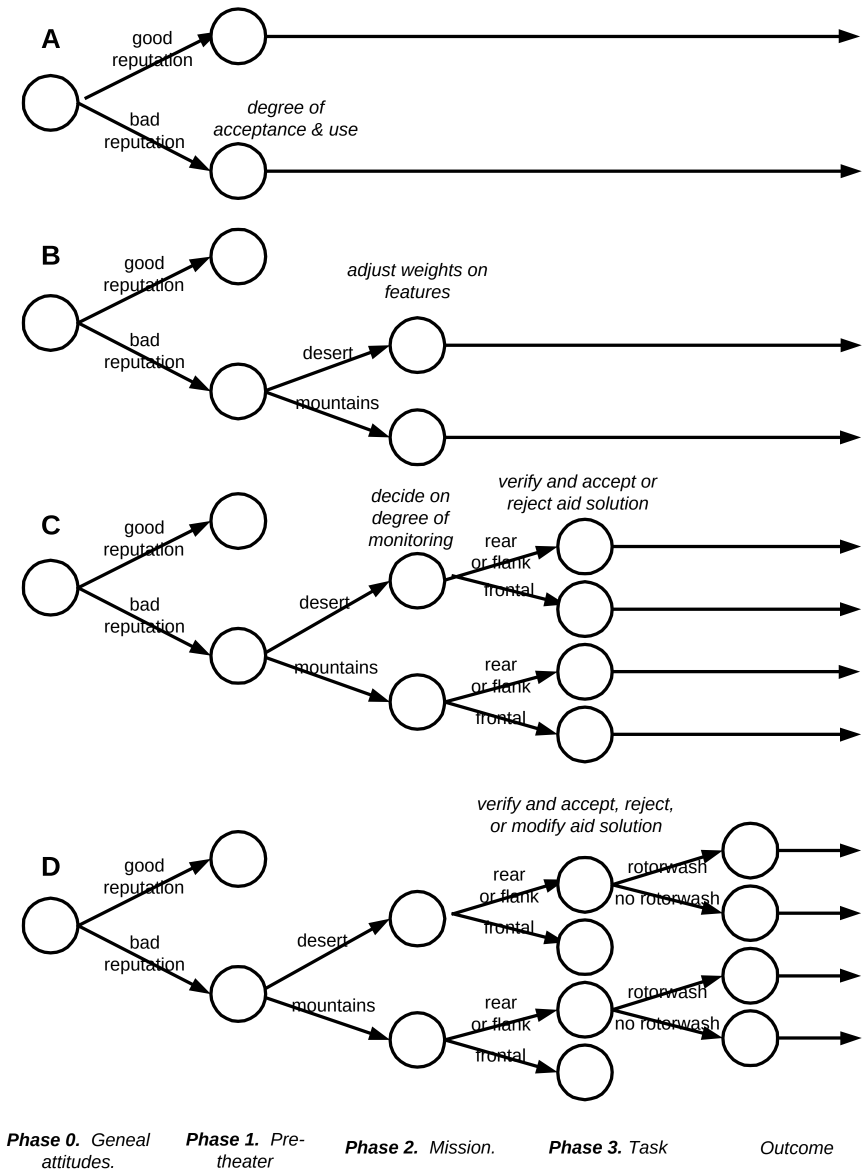

Figure 2. Illustrative event trees varying in granularity.

A key feature of trust is that it evolves as the user gains experience with an aid and moves through the various phases. This process of acquiring information over time can be visualized as a progression along the branching possibilities of an *event tree* (Shafer, 1996), in which different types and quantities of information become available at different phases. Each node in the event tree stands in effect for a *question* that is warranted by the user's beliefs about the factors that influence aid performance and that can be "asked" at that particular time. Each branch emerging from a node represents one of the possible *answers* to the question, and reflects (a) the user's ability to *recognize relevant patterns* and *associate desirable or undesirable outcomes* with them, or (b) the user's ability to *make* relevant facts true by an interaction decision. Each answer becomes part of the Grounds for subsequent trust judgments and interaction decisions. The richer the event tree, i.e., the greater its granularity, the more actively the user can adapt the aid to the situation and fully exploit its value. The tree is a map of the discriminations the user can act on and the interactive strategies available at each phase.

The event trees shown in Figure 2 represent Army aviators using a fictional version of a Combat Battle Position Recommendation (CBPR) aid, as part of the Rotorcraft Pilot's Associate. They increase in granularity from aviator A, who has the least knowledge, to D, who has the most. In Phase 1, before the users are assigned to a theater of operations, awareness of the aid's reputation influences their tendency to accept and use it. Since A has no further knowledge about the performance of the aid, A's degree of trust is fixed after Phase 1 and supports no strategies other than use or non-use. B, C, and D, however, discriminate levels of expected performance based on terrain; after they learn whether a mission is in the desert or mountains, they will adjust weights on features used by the aid to evaluate potential combat battle positions. After Phase 2, B acquires no further information about the reliability of the aid and makes no further interactive decisions. Aviators C and D discriminate whether the recommended battle position provides a frontal versus a rear or flanking angle of attack and know that the chances of an undesirable frontal angle are higher in mountainous than in desert terrain. As a result, C and D will use information about terrain in Phase 2 to decide how frequently to monitor aid conclusions, to verify angle of attack against likely enemy avenues of approach. After Phase 2, Aviator C acquires no further information. D, however, knows that rotorwash (dust, leaves, and snow blown up by helicopter blades) is another feature of combat battle positions that can affect success but is not incorporated into aid algorithms. There is often no way to know whether a position has rotorwash until it is actually inspected. As a result, D may modify the aid's recommendations in Phase 3 to pick regions where there are multiple battle position options. D will then be able to shift battle positions based on close inspection of the terrain.

CONCLUSION

The Situation Specific Trust model helps define the content that training of decision aid users should attempt to convey. It provides an account of the "mental models" required by effective decision aid users at each phase of decision aid use, the monitoring and situation awareness skills required to use those mental models effectively, and the interaction decisions that flow from that situation awareness. SST also provides insight into critical thinking skills necessary to identify and handle novel situations, where assumptions fail. In addition, SST provides tools for generating scenarios in which users can acquire the relevant mental models, and diagnostic measures to assess their progress.

REFERENCES

Cohen, M. S., Thompson, B. B., Masalonis, A., Bresnick, T., Parasuraman, R. Hess, S. & Schwartz, J. (2000). *Trust in decision aids: What is it, how to train it.* Arlington, VA: Cognitive Technologies, Inc.

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

Lee, J D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics, 35*(10):1243-1270.

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., & Moray, N. (1987). Operator's trust in relation to system faults. *IEEE International Conference on Systems, Man, and Cybernetics*, 258-263. Alexandria, VA.

Muir, B., & 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.

Shafer, G. (1996). *The art of causal conjecture*. Cambridge, MA: The MIT Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.