Int. J. Man–Machine Studies (1987) **27**, 527–539

Trust between humans and machines, and the design of decision aids

BONNIE M. MUIR

Department of Psychology, University of Toronto, Canada

A problem in the design of decision aids is how to design them so that decision makers will trust them and therefore use them appropriately. This problem is approached in this paper by taking models of trust between humans as a starting point, and extending these to the human–machine relationship. A definition and model of human–machine trust are proposed, and the dynamics of trust between humans and machines are examined. Based upon this analysis, recommendations are made for calibrating users' trust in decision aids.

1. Introduction

The concept of trust is a critical one in the design of decision support systems. A decision aid, no matter how sophisticated or "intelligent" it may be, may be rejected by a decision maker who does not trust it, and so its potential benefits to system performance will be lost. If forced to use an aid which he does not trust, a decision maker may use any means available, even very demanding and time consuming ones, to direct the output of the aid toward his own decision. On the other hand, a user may trust a decision aid more than is warranted, allowing it to perform functions which would be performed better by the human or which may lead to system failure. Woods, Roth & Bennett (1987) have recently observed all of these reactions from technicians who were using a new expert system designed to support them in trouble-shooting an electro-mechanical device. Thus, the problem we face is how to design decision aids which decision makers will trust enough to use, but will use discriminatingly and effectively. The user's trust must be calibrated to the decision aid, so that he neither consistently underestimates nor overestimates its capabilities.

How can we design decision aids which decision makers will appropriately trust and use? To answer this question, we must examine the concept of trust. We must understand the nature of the trust between humans and machines, and the factors which foster trust and undermine it. And we need to understand how trust changes with experience on a system—how it grows, how it is diminished or destroyed by system failures, and how it may recover. Although the importance of the concept of trust between humans and machines has been mentioned in a number of articles concerning process control (Sheridan, Fischhoff, Posner & Pew, 1983; Sheridan & Hennessy, 1984), it has not been systematically studied yet in any man–machine domain. However, some research has been done by psychologists and sociologists on modelling trust between humans. This research is used as a starting point in this paper to model the trust that may exist between humans and machines in general, and between humans and decision aids in particular.

The remainder of the paper is divided into five sections. In Section 2 a definition and model of people's trust in machines are presented. The third section is a

0020-7373/87/050527 + 13$03.00/0

discussion of the dynamics of trust. This is followed in the fourth section by a discussion of trust, distrust, and mistrust, and the consequences of these in the human–machine relationship. Finally, in the fifth section, recommendations are made for setting users' trust in decision aids to an appropriate level. This latter process will be called "calibrating" trust to a machine. Concluding remarks are presented in Section 6.

2. A definition and model of trust between humans and machines

Several definitions of trust between humans have been proposed in the psychological literature, including:

- "the confidence that one will find what is desired from another, rather than what is feared" (Deutsch, 1973);
- "[an] Actor's willingness to arrange and repose his or her activities on [an] Other because of confidence that [the] Other will provide expected gratifications" (Scanzoni, 1979);
- "a generalized expectancy held by an individual that the word, promise, oral or written statement of another individual or group can be relied on" (Rotter, 1980);
- "a generalized expectation related to the subjective probability an individual assigns to the occurrence of some set of future events" (Rempel, Holmes & Zanna, 1985);
- "the degree of confidence you feel when you think about a relationship" (Rempel & Holmes, 1986).

Each of these definitions seems to capture a different aspect of our everyday usage of "trust"; this suggests that trust is a multidimensional construct. The definitions do, however, have several things in common. First, trust is described as an expectation of, or confidence in, another. Thus, trust is oriented toward the future—future gratifications, behaviours or events. Second, trust always has a specific referent; we trust in someone or something and our trust is particular to that referent. Third, trust may relate to many properties of the referent, including their reliability, honesty, and motivations. The problem with these definitions is that they are too general, and they fail to explicitly acknowledge the multidimensional nature of trust which is suggested by their variety.

The limitations of the foregoing definitions are overcome in a definition of trust proposed by Barber (1983), a sociologist. Barber explicitly recognizes the multidimensional character of trust, defining trust in terms of a taxonomy of three specific expectations:

(1) our very general expectation of the *persistence* of the natural (physical and biological) and the moral social orders (i.e. we expect natural physical laws to be constant, human life to survive, and mankind (and computers) to be good and decent, respectively);

(2) our expectation of *technically competent role performance* from those involved with us in social relationships and systems;

(3) our expectation that partners in an interaction will carry out their *fiduciary obligations and responsibilities*, that is, their duties in certain situations to place others' interests before their own.

TABLE 1

A two-dimensional framework for studying trust in human–machine relationships, produced by completely crossing Barber's (1983) *taxonomy of the component expectations of trust* (*represented by the rows*) *and Rempel, Holmes and Zanna's* (1985) *taxonomy of the dynamics of trust* (*columns*). *This integrated framework is more complete than either taxonomy alone. Statements in the cells exemplify the nature of a person's expectations of a referent* (*j*) *at different levels of experience in a relationship*

	Basis of expectation at different levels of experience		
Expectation	Predictability (of acts)	Dependability (of dispositions)	Faith (in motives)
Persistence			
Natural physical	Events conform to natural laws	Nature is lawful	Natural laws are constant
Natural biological	Human life has survived	Human survival is lawful	Human life will survive
Moral social	Humans and computers act "decently".	Humans and computers are "good" and "decent" by nature	Humans and computers will continue to be "good" and "decent" in the future
Technical competence	*j*'s behaviour is predictable	*j* has a dependable nature	*j* will continue to be dependable in the future
Fiduciary responsibility	*j*'s behaviour is consistently responsible	*j* has a responsible nature	*j* will continue to be responsible in the future

All three of Barber's (1983) meanings of trust seem applicable to the human–machine relationship. Accordingly, the meaning of each of the three expectations in the context of the human–machine relationship is developed in more detail below. They are also presented, along with examples, in the rows of Table 1 (crossed with another dimension which is discussed in the next section of this paper).

Our expectation of the persistence of natural physical laws allows us to understand and create mental models of physical processes, and to use such mental models to predict future events. It is also our expectation of persistence that allows us to construct rule bases for decision support systems.

The more specific expectation of technically competent role performance is at the heart of the trust between humans and machines. Barber (1983) has identified three types of technical competence that one human may expect from another: expert knowledge, technical facility, and everyday routine performance. These correspond closely to Rasmussen's (1983) taxonomy of behaviour into knowledge-, rule-, and skill-based behaviour. Both humans and machines may possess only a subset of these competencies in a particular domain. For example, a physician may expect a patient to be competent to take medication (skill-based behaviour), but not expect him or her to be competent to choose a medication (rule-based) or to interpret any

reaction to it (knowledge-based). Similarly, a human user may expect a decision aid to perform competently the routine task of gathering data, but not expect it to have the technical competence to respond appropriately to it, nor the knowledge required to interpret it.

The third expectation, of fiduciary responsibility, is invoked as a basis for trust when the trustor's own technical competence is exceeded by the referent's, or is not known to him. Unable to evaluate the referent on the basis of competence, he is forced to rely on the referent's moral obligation not to abuse the power that he wields. The trust that we have in politicians or in the domain experts we consult (e.g. physicians, lawyers, teachers) is an example of trust based primarily on fiduciary responsibility. (Trust in domain experts is based to some extent upon competence since we know they have met at least the minimal competency requirements dictated by the licensing and regulatory bodies of their respective professions. But we do not have even these minimal guarantees for politicians—or expert systems.) The issue of responsibility is an important one in designing decision support systems. It is particularly problematic with support systems which are designed as "prostheses" to replace the human in some way or to remedy some human deficiency [see Woods, Roth & Bennett (1987) for a comparison of decision support paradigms]. Presumably, a person consults with a prosthetic machine expert because the machine possesses greater expertise than the human in the domain of interest. Consequently, the person is unable to evaluate the machine expert in terms of its competence, and must rely on his or her assessment of the machine's responsibility (operationalized here as its design-based intentions or purposes). But this too may be problematic for the user. The machine's intentions in asking for certain information or providing certain solutions may be ill-understood by the user either because (1) an explanation capability is not included in the system, or (2) the explanation given is opaque [e.g., if it is expressed in terms of the machine's cognitive system (rule structure) rather than in the terms which the human uses to think about the problem]. As a result, the user will have difficulty in assessing the machine expert's responsibility, and therefore, in calibrating his trust to the machine. The issue of machine responsibility will become more important in human–machine relationships to the extent that we choose to delegate autonomy and authority to "intelligent", but prosthetic, machines†. The more power they are given, the greater will be the need for them to effectively communicate the intent of their actions, so that the people who use them can have an appropriate expectation of their responsibility and interact with them effectively. The alternative is to design systems which give both authority and responsibility to the human user, an approach which is discussed in more detail in Section 5 of this paper.

Because of its relative specificity and completeness, Barber's (1983) definition provides the best definition of trust in human relationships. Accordingly, Barber's taxonomy is adopted in this paper as a basis for the development of a definition and

† Unfortunately, many expert systems of this kind are being built just because they *can* be, with too little consideration given to whether they *should* be [see, for example, Hopple (1986), on methidolatry and the law of the hammer]. The design of these expert systems is driven by technology, rather than by system objectives (Rouse, 1986) or design principles (Woods, Roth & Bennett, 1987). As a result, the emphasis in these expert systems is placed on the machine's performance, with explanation capability and even deception (e.g. Mulsant & Servan-Schreiber, 1983) tacked on more or less to appease the user.

model of trust in human–machine relationships. The following definition of trust in human–machine relationships is proposed:

> Trust (T) is the expectation (E), held by a member (i) of a system, of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but not necessarily isomorphic with, objective measures of these qualities.

Or alternatively,

$$T_i = [E_{i(\mathrm{P_n}+\mathrm{P_m})}] + [E_{i\mathrm{TCP}_j}] + [E_{i\mathrm{FR}_j}]$$

Several aspects of this definition should be noted. T and E are subscripted by the individual holding the expectation. This is to explicitly recognize that trust is based on the *perceived* qualities of another and is therefore subject to all the vagaries of individual interpretation. Thus, the perceived properties, which support an expectation may be quite independent of a referent's actual properties, which are referred to as the referent's trustworthiness. P, TCP and FR are also subscripted to identify the particular referent of the expectations. T is a composite expectation, comprised of the three expectations on the right: P is the fundamental expectation of persistence; TCP includes skill-, rule-, and knowledge-based behaviour, and the reliability and validity of a referent; FR includes the concepts of intention, power and authority.

This definition suggests that a simple, additive model of trust may suffice. However, it seems intuitively reasonable that the three component expectations will be of different relative importance in different circumstances and so should be weighted accordingly. In addition, the three component expectations may interact, with one expectation modifying another. Assuming that these three expectations are exhaustive, and that the model is linear in its parameters and in its independent variables, a model of a human's trust in a machine will take the form:

$$T_i = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3 + B_6X_2X_3 + B_7X_1X_2X_3,$$

where $B_{0\text{–}7}$ are parameters, X_1=P (persistence), X_2=TCP (technically competent performance), X_3=FR (fiduciary responsibility).

3. The dynamics of trust between humans and machines

Rempel, Holmes and Zanna (1985) have suggested that trust between humans is a dynamic expectation that undergoes predictable changes as a result of experience in a relationship. In this section, Rempel *et al.*'s model is extended to develop hypotheses about how a human's trust in a machine changes as a result of experience on a system. The dynamic dimension of trust is represented by the columns of Table 1; the cells contain statements which characterize the different types of trust as they grow.

Early in a relationship, a person bases his trust upon the predictability of another person's behaviours (Rempel *et al.*, 1985). Similarly, early in his experience, a person will judge the predictability of a machine by assessing the consistency of its

recurrent behaviours. The growth of trust will depend on the human's ability to estimate the predictability of the machine's behaviours. Unless its behaviour is completely deterministic, these behaviours will be distributed with a variance about some mean. The human's ability to estimate these properties will depend on his own limitations as a decision maker, and on certain properties of the machine and its environment. For example, we know that humans as decision makers are characterized by certain biases (Kahneman & Tversky, 1973; Sage, 1981). In particular, their tendency to overestimate the representativeness of small samples may cause them to base their judgment of predictability upon too small a sample, a tendency which will usually result in an underestimation of variance, that is, an overestimation of predictability. And as trust develops, probably on the basis of too small a sample (if Kahneman and Tversky are correct), trust itself will tend to reduce the amount of sampling. Hence, knowledge about a machine's behaviour will be inversely related to trust. Also, the more constrained a machine's behaviours are, the greater will be its predictability, and so trust will be inversely related to the degrees of freedom of the machine. The more stable the environment in which the machine operates, the more predictable it will appear to be, so trust will be directly related to the stability of the environment, and inversely related to the amount of disturbance imposed on the machine by the environment. Finally, since the human's trust is based upon observations of machine behaviour, behaviours must be observable for trust to grow.

Later in a relationship, trust in another person is based upon the attribution of a dependable disposition (Rempel *et al.*, 1985). The attribution of dependability may be thought of as a summary statistic of an accumulation of behavioural evidence, which expresses the extent to which another person may be relied upon. The attribution of dependability is based upon perceived predictability, but with a heavy weighting placed on events that involve risk. To decide that another person is dependable, the opportunity must exist for the person to be undependable; they must pass a fair test. In the context of a human–machine relationship, experience with a system generates the behavioural evidence the human requires to make a generalization about a machine's dependability. Some factors which affect the attribution process, such as the tendency (discussed earlier) to gather too little evidence, may lead to overestimations of dependability. However, people's tendency to make the fundamental attribution error—overestimating the role of dispositional factors and underestimating the role of environmental factors in attributing the causes of others' behaviours—may work in the opposite direction: if a person attributes a perceived unpredictability of a machine to the machine's properties rather than to environmental instability, he will tend to underestimate the machine's predictability and dependability. Finally, Rempel *et al.*'s analysis predicts that experience with system events involving risk is a prerequisite for the attribution of dependability to a machine.

The final stage in the growth of trust between humans is the development of faith (Rempel *et al.*, 1985). We cannot know that another person will continue to be dependable in the future. Our attribution of dependability is based on only a limited sample of behaviours, and so it may not be a good indicator of a person's disposition and future behaviours. And the future is filled with uncertainty—circumstances

change and people change. Therefore, a leap of faith beyond the available evidence, a closure against doubt, is required to believe that another person will continue to be dependable in the future. In interpersonal relationships, a referent's history of predictability and dependability contribute to the development of faith, but heavy weighting is given to events which indicate a referent's intrinsic motivation for remaining in a relationship. This stage of trust also has a place in a model of human–machine trust, although the concept of motivation probably is not applicable, at least with today's machines. For example, in complex, highly automated, process control, many processes defy complete understanding and are explicable only in fuzzy terms (Sheridan & Hennessy, 1984). That operators *do* operate these systems in the face of the uncertainty of the future, even though they (1) appreciate the vast number of possible but unforseen interactions that can occur in these systems, (2) realize that their own knowledge of the system is incomplete and perhaps sometimes even incorrect, and (3) recognize the "brittleness" (e.g. Brown, Moran, & Williams, 1982) of procedures in such systems, implies that they must have made a leap of faith. Similarly, the workings of decision support systems in these complex environments (and even in much more narrowly bounded domains) may be so complex that they are never completely understood by their users. Therefore, a leap of faith is also required for users to believe that these decision aids will continue to be dependable in the face of future uncertainty. Since there is no direct analogue of motivation in today's machines, faith in a machine may be based only upon predictability and dependability, and may perhaps depend on extended experience.†

So far this discussion has considered only the factors that promote the growth of trust. We also need to know the factors which undermine trust, and the nature of the recovery process, but much less is known about these, even in human relationships. It is thought that trust between people is "notoriously easy to break down . . . [and] doubly difficult to reestablish" (Rempel *et al.*, 1985, p. 111), but there is no empirical evidence yet to support this intuition. Rempel *et al.* have also made the interesting prediction that trust may give way in the same order in which it was established, that is, expectations of predictability will be lowered before the more resistant attribution of dependability and faith are affected. These predictions need to be tested in the context of human–machine relationships. We need to trace the growth, destruction, and recovery of a human's trust in a machine in response to machine behaviours. In particular, it would be useful to know how different kinds of machine failures affect a human's trust, and whether these effects are specific to affected subsystems or whether they generalize to other subsystems. With respect to decision support systems, the question becomes: When a decision maker discovers a "bug" in one function of a decision support system (e.g., in its deductive inference capability), how is his trust in, and therefore his use of, other functions (e.g. its ability to represent and manipulate possible worlds) affected?

† Usually a lack of understanding of a machine will limit the extent of a human's faith in it. However, an exception may occur, and a lack of understanding may actually *enhance* a human's faith in the machine if he becomes "mystified" (Sheridan, 1980; Sheridan, Vamos & Aida, 1983) by it. This kind of faith can be described as "blind faith".

4. Trust, distrust, and mistrust between humans and machines

A person must decide whether to trust or distrust a machine with which he interacts. Each individual will have a criterion of competence† beyond which a hypothesis of trust will be adopted, and below which a hypothesis of distrust will be adopted. A symbiosis will be achieved when the human trusts and exploits the capabilities of a competent machine (or, in a multifunctional system, its competent functions). Similarly, overall system performance will be improved, and disastrous consequences may be avoided, if the human distrusts and rejects the use of an incompetent machine (or incompetent functions). The human commits an error of *mistrust* if he distrusts a competent machine (or function), or trusts an incompetent one.

In supervisory control environments, operators, especially novices, may be biased toward distrust (Sheridan & Hennessy, 1984). This is at odds with Luhman's (1980) claim that people tend to approach new social relationships with an attitude of trust because it requires less mental effort to trust than to distrust. The basis for the discrepancy may lie in the differential risk associated with the two situations. Automated control systems and decision support systems are usually designed for complex tasks involving some element of risk. Thus, the extra effort required for operators to distrust their machines, including their decision aids, is justified by the potential for disaster in many man–machine working environments, and by the differential consequences of the two errors of mistrust. If an operator distrusts a competent machine, he may perform its function himself, perhaps satisfactorily enough that immediate system performance will be maintained, although other neglected functions might have undefined consequences at some undefined point in the future. In contrast, if an operator trusts an incompetent machine, the consequences may be immediate and catastrophic. (Interestingly, the opposite seems to be true with automatic protection devices, where it is a more devastating error for an operator to override a correct automatic shutdown than it is to allow an incorrect one. The operator is left in the conflicting position of having to trust the automatics the most just when the risk is the highest.) Another reason humans may be biased toward distrust of semi-autonomous machines, including prosthetic decision support systems, is simply because of the fact that system designers saw fit to have a human in the system at all: the human's mere presence implies that the machine may be incompetent or irresponsible, and a distrusting human is required in the system to monitor the machine's output.

A person's trust or distrust of a machine will affect his allocation of functions in the system, and this may, in turn, affect the stability of his trust or distrust. If a machine or function is distrusted, the user will, if possible, do the task(s) himself. This leaves little or no opportunity for him to reevaluate his distrust because (1) the machine (or function) has been removed from the system and is not producing the behavioural evidence needed to support a reevaluation, and (2) the human is left with little or no time for the reevaluation process because he is busy performing the

† For the sake of simplicity, only the dimension of competence is mentioned in the discussion which follows. But it is important to remember that trust is a composite expectation; whether an individual uses a criterion of competence, responsibility, persistence, or some mixture of these to determine his trust in a machine will depend on the specific situation.

task himself. In contrast, if a person trusts a machine (or function), he will allow it to perform its tasks, leaving available both the evidence and the time needed for him to reevaluate its trustworthiness as necessary. Thus, distrust will be relatively more resistant to change because the allocation of function it demands severely restricts the opportunity for the user to gather further, possibly disconfirming, evidence. An implication of this is that a human's trust in a machine, once betrayed, may be difficult to recover.

5. Calibrating trust between humans and machines

Different machines are not equally competent, nor are different functions of a single machine necessarily equally competent. Therefore, it is inappropriate for a user to trust or distrust them all equally. Rather, the user must learn to calibrate his trust, that is, to set his trust to a level corresponding to a machine's or function's trustworthiness, and then use the machine accordingly. In human–machine systems in general, a well-calibrated operator is one who gets the most out of a system; his appropriate trust in competent subsystems allows him to devote his time and effort to compensating for less competent subsystems which he appropriately distrusts. A poorly-calibrated operator is one who overrides competent subsystems and/or fails to override incompetent ones. With respect to decision aids specifically, the well-calibrated user of an instrumental (*vs.* a prosthetic) decision aid maximizes overall system performance in decision making and problem solving tasks by readily accepting the output of competent machine functions and relying on his own competence or that of other resources for functions which the machine handles incompetently. The poorly-calibrated user is one who degrades system performance because he fails to capitalize on, or rejects outright, the output of competent machine functions, and/or is "led down the garden path" to accept incompetent output. The difference between this and the calibration of trust to a prosthetic decision aid lies primarily in the type of output (i.e. behaviour) which the machine produces and which the human user therefore must evaluate. The instrumental decision aid produces information to be used during the human's decision making process, whereas the prosthetic decision aid produces recommended decisions or solutions. This difference has an important implication for the ability of the user to assess the competence of a decision aid; this is discussed in more detail in the third recommendation below.

The question posed at the beginning of this paper was how we should design decision aids so that decision makers will use them with well-calibrated trust. The foregoing discussion of trust allows us to make several tentative recommendations about the calibration of trust to machines in general, and to decision support systems in particular. Calibration could be improved by:

(1) improving the user's ability to perceive a decision aid's trustworthiness,
(2) modifying the user's criterion of trustworthiness,
(3) enhancing the user's ability to allocate functions in a system,
(4) identifying and selectively recalibrating the user on the dimension(s) of trust which is (are) poorly calibrated.

Means for accomplishing these goals are suggested below.

5.1. IMPROVING THE PERCEPTION OF TRUSTWORTHINESS

This strategy would involve, first of all, training the user to understand (to the extent possible in the application) how the decision aid works. It would also involve providing the user with explicit data about the predictability of a decision aid's output in terms of its competence and responsibility. Data on predictability are used because predictability is the foundation upon which trust is built; facilitating the higher-level attribution of dependability would involve presenting summary information about an aid's predictability over time and in different circumstances. Information about competence should also include information about system constraints and environmental stability. Increasing the observability of machine behaviours and the transparency of functions will provide the user with the evidence necessary to support the summary information he is given, and should also help him to improve his own ability to summarize predictability data. The user's ability to perceive the aid's responsibility could be enhanced by improving the decision aid's ability to communicate its intentions. If the aid's intentions are expressed in a fashion that differs from the user's (and the machine's expressions cannot or will not be changed), then training the user to interpret the machine's expressions of intention should help him to perceive more accurately the machine's responsibility. The decision maker needs ample hands-on interaction with a decision aid to develop calibrated expectations of it, so plenty of time should be allowed for this. If this experience can be on a simulator (or in simulated circumstances) which allows the user to experience risky events, all the better, since these are necessary for the growth of trust.

5.2. MODIFYING THE CRITERION OF TRUSTWORTHINESS

This process would involve making explicit the decision aid's domains of expected competence, its actual history of competence and responsibility, and a criterion level of acceptable performance. Providing an explicit comparison of system performance when the user is using and not using the decision aid may also help. Under some circumstances, and especially with prosthetic aids, users may be motivated to set their criterion of machine competence or responsibility unrealistically high or low. Users may set their criterion of competence too high (so that the aid can never reach it) if they do not want to use the aid (perhaps because they fear they will be replaced by the machine); they may set it too low if they feel incompetent to do the task themselves (perhaps because they are novices or have become deskilled due to using the automation) or if they do not want to do a particular task (perhaps because it is too tedious). Users may set their criterion of responsibility too high if they feel their own responsibility and authority are being usurped by the aid, and set it too low if they wish to abrogate responsibility. These motivated "miscalibrations" do not depend on the properties of specific machines per se, and so they will be resistant to change via training or machine "fixes". Rather, they are symptoms of the human's difficulty in dealing with a situation that requires him to relinquish control and authority to a machine, but retain responsibility. Alleviation of these symptoms requires a different approach to decision aiding, which is discussed next.

5.3. ENHANCING THE USER'S ABILITY TO ALLOCATE FUNCTIONS

This analysis of man–machine trust supports the move away from designing stand-alone, prosthetic expert systems and toward the design of decision aids as

instruments (Woods, 1986). In the former case, the machine is in control, and the human is its servant (although, ironically, the human is also expected to act as the machine's master, monitoring it and overruling it when indicated). Because this kind of machine's competence is presumed to exceed that of the human whose job it is to use (and supervise) it, the human decision maker is really in no position to evaluate such a machine's competence, and will have difficulty calibrating his trust to it, and using it effectively. Understandably, he may come to distrust both the competence of the designers (and management) who relegated him to such a paradoxical role, as well as their responsibility if he feels the aid is meant to replace him as a decision maker. This distrust may, in turn, feed back to reduce his expectation (of persistence) that machines and humans are basically good and decent. Many of these effects are described by Sheridan (1980; Sheridan, Vamos & Aida, 1983) in his analysis of automation and "alienation". In contrast, the human is in control of a decision support system which is designed as an instrument (Woods, Roth & Bennett, 1987), and he has the dignified task of dynamically allocating functions to it. The human uses an instrumental decision aid as one of many resources he may call upon to provide information which will help him to make decisions and solve problems (cf. the decision prosthesis which offers solutions). In this approach to decision aid design, the human is explicitly recognized as having the authority and the responsibility for decision making. The competence of the human decision maker will always exceed that of the decision aid in this human–machine relationship, and so the human will be better able to judge the aid's competence (i.e. the competence of its capability to provide information), calibrate his trust to it, and use it effectively. Since the human holds the critical position of deploying power, and is in no danger of having his decision-making task usurped by the machine, many sources of alienation cease to be an issue.

5.4. IDENTIFYING AND SELECTIVELY MANIPULATING THE SOURCE(S) OF POOR CALIBRATION

This analysis has shown that poor calibration may arise from an inaccurate expectation of persistence, competence, and/or responsibility. Thus, calibration training should begin by specifying the expectation (or expectations) that is the basis of poor calibration, and then proceed by selectively improving on that dimension. The model of human–machine trust presented in this paper facilitates this process by defining the relevant dimensions of trust. By analyzing some of the factors which determine the extent of trust in each dimension, it suggests ways to calibrate trust in a human–machine relationship.

The recommendations offered here must be regarded as tentative until the model of trust from which they are derived is tested. It is encouraging to note, however, that this perspective on the human–machine relationship and several other quite different perspectives (e.g. Rouse & Morris, 1986; Sheridan, 1980; Sheridan, Vamos & Aida, 1983) point to many of the same problems in and converge on many of the same recommendations for designing automation to suit people.

6. Conclusion

This paper has taken models of trust between humans and has developed the hypotheses that result if these are extended to the human–machine relationship. We

do not know yet how well models of trust between humans will generalize to the human–machine relationship. But clearly the importance of trust as a moderator between the properties of a machine and a human's use of that machine suggests that a formal model of trust between humans and machines would be of great value to the designers of human-machine systems in general, and decision aids in particular.

This paper has concentrated on the trust between human users and their machines. It has only touched upon the fact that this relationship may be affected by the user's trust in other referents (e.g. fellow workers, designers, management, and society in general) and by their trust in him. An advantage of the model of trust presented here is that it is broad enough to apply to all of these relationships, but at the same time is specific enough to predict problems in the calibration of trust in these relationships and suggest ways to solve these. Another advantage of the model is that it suggests testable hypotheses about the nature and role of trust in human–machine systems. A series of experiments is about to begin in our laboratory to test some of these hypotheses.

Acknowledgement

I would like to thank Neville Moray, Thomas Sheridan, Kim Vicente, and David Woods for their helpful comments on an earlier version of this paper and a related working paper. This work was funded by the Natural Sciences and Engineering Research Council of Canada, partly by a postgraduate scholarship to the author, and partly through Grant A1794 to Neville Moray.

References

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

BROWN, J. S., MORAN, T. P., & WILLIAMS, M. D. (1982). The semantics of procedures, Technical Report. Palo Alto: Xerox Palo Alto Research Center.

DEUTSCH, M. (1973). *The Resolution of Conflict: Constructive and Destructive Processes.* New Haven, CN: Yale University Press.

HOPPLE, G. W. (1986). Decision aiding dangers: The law of the hammer and other maxims. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-16,** 948–964.

KAHNEMAN, D. & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review,* **80,** 237–251.

LUHMAN, N. (1980). *Trust and Power.* New York: Wiley.

MULSANT, B. & SERVAN-SCHREIBER, D. (1983). Knowledge engineering: A daily activity on a hospital ward, Technical Report STAN-CS-82-998. Stanford: Stanford University.

RASMUSSEN, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-13,** 257–266.

REMPEL, J. K. & HOLMES, J. G. (1986). How do I trust thee? *Psychology Today,* February, 28–34.

REMPEL, J. K., HOLMES, J. G. & ZANNA, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology,* **49,** 95–112.

ROTTER, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist,* **35,** 1–7.

ROUSE, W. B. (1986). Design and evaluation of computer-based decision support systems. In: ANDRIOLE, S. Ed., *Microcomputer decision support systems: Design, implementation, and evaluation.* Wellesley, MA: QED Information Sciences.

ROUSE, W. B. & MORRIS, N. M. (1986). Understanding and enhancing user acceptance of computer technology. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-16,** 965–973.

SAGE, A. P. (1981). Behavioral and organizational considerations in the design of information systems and processes for planning and decision support. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-11,** 640–678.

SCANZONI, J. (1979). Social exchange and behavioral interdependence. In: BURGESS, R. L. & HUSTON, T. L., Eds. *Social Exchange in Developing Relationships.* New York: Academic Press.

SHERIDAN, T. B. (1980). Computer control and human alienation. *Technology Review,* October, 61–73.

SHERIDAN, T. B., FISCHHOFF, B., POSNER, M. & PEW, R. W. (1983). Supervisory control systems. In: COMMITTEE ON HUMAN FACTORS, *Research Needs for Human Factors.* Washington, DC: National Academy Press. pp. 49–77.

SHERIDAN, T. B. & HENNESSY, R. T. Eds. (1984). *Research and modeling of supervisory control behavior: Report of a workshop.* Washington, DC: National Academy Press.

SHERIDAN, T. B., VAMOS, T. & AIDA, S. (1983). Adapting automation to man, culture and society. *Automatica,* **19**(6), 605–612.

WOODS, D. D. (1986). The design of decision aids in the age of "intelligence." In: *Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics.* pp. 398–401.

WOODS, D. D., ROTH, E. M. & BENNETT, K. (1987). Explorations in joint human–machine cognitive systems. In: ZACHARY, W. & ROBERTSON, S. Eds., *Cognition, Computing and Cooperation.* Norwood, NJ: Ablex.