This article was downloaded by: [Stanford University Libraries]
On: 13 September 2012, At: 04:34
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Ergonomics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/terg20

Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems

BONNIE M. MUIR [a]

[a] Department of Psychology, University of Toronto, Toronto, Ontario, Canada

Version of record first published: 31 May 2007.

To cite this article: BONNIE M. MUIR (1994): Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems, Ergonomics, 37:11, 1905-1922

To link to this article: http://dx.doi.org/10.1080/00140139408964957

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

ERGONOMICS, 1994, VOL. 37, NO. 11, 1905–1922

Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems

BONNIE M. MUIR

Department of Psychology, University of Toronto, Toronto, Ontario, Canada

Keywords: Trust; automation; Allocation of functions; Attitude to computers; Human–computer interaction; Supervisory control.

Today many systems are highly automated. The human operator's role in these systems is to supervise the automation and intervene to take manual control when necessary. The operator's choice of automatic or manual control has important consequences for system performance, and therefore it is important to understand and optimize this decision process. One important determinant of operators' choice of manual or automatic control may be their degree of trust in the automation. However, there have been no experimental tests of this hypothesis until recently, nor is there a model of human trust in machines to form a theoretical foundation for empirical studies. In this paper a model of human trust in machines is developed, taking models of trust between people as a starting point, and extending them to the human–machine relationship. The resulting model defines human trust in machines and specifies how trust changes with experience on a system, providing a framework for experimental research on trust and human intervention in automated systems.

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

1. Introduction

Toffler's (1980) 'third wave' is upon us. Computerization is widespread, and modern computer systems are so sophisticated that they are capable of running a fully automated nuclear power plant or controlling an aircraft's flight from start to finish. However, in most cases, some element of human control is still retained in automated systems. The human operator's role in these systems is to act as a supervisor of the automation, to monitor its performance during normal operations, and intervene to take manual control when necessary, perhaps to trim a variable or to override faulty automation. Thus, the moment-to-moment, on-line allocation of functions in the system is at the discretion of the human supervisor.

The supervisor's choice of manual or automatic control can have important consequences for system performance. Highly automated systems are designed to run in automatic mode most of the time, for maximum safety and productivity over long time horizons. If the supervisor overrides the automation too frequently or is too hesitant to take manual control, system performance will be compromised, with potentially disastrous consequences. Clearly, the supervisor's moment-to-moment allocation of functions is a critical decision-making process, and it behooves us to understand and optimize this process.

Unfortunately, we do not have a good understanding of supervisors' intervention behaviour. The technological imperative has driven designers to automate whenever possible, but our knowledge of how operators interact with complex, powerful, and 'intelligent' automated systems lags far behind (Rouse 1977). Although a variety of models has been proposed to describe different aspects of supervisory control behaviour

0014–0139/94 $10·00 © 1994 Taylor & Francis Ltd.

(e.g., Moray 1987, Sheridan 1986), no single comprehensive theory of supervisory control has emerged which can describe, explain and predict intervention behaviour. Sheridan (1980, Sheridan *et al.* 1983a, Sheridan and Hennessy 1984, Sheridan *et al.* 1983b) has advanced the intriguing hypothesis that supervisors' intervention behaviour is based upon their trust in the automation. Some support for this hypothesis comes from the case studies of Halpin *et al.* (1973) and Zuboff (1988), who found that operators' trust was related to their acceptance of new automation. Recent experimental studies by Muir (1989, and to be reported in Part II of this article), and an extension by Lee and Moray (1992), were also consistent with Sheridan's hypothesis.

To test the hypothesis that trust is an important factor in determining supervisors' intervention behaviour in automated systems, we need a model of human trust in machines, and no such model exists. This paper is an attempt to address this issue by developing a theoretical model of human trust in machines. Although this paper focuses on supervisory control systems, the model of trust developed here can also be applied to other machines, from the ones we use in everyday life, like watches and dishwashers and cars, to so-called 'intelligent' machines such as decision-aiding computer systems (see Muir 1987).

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

2. The concept of trust in supervisory control systems

Sheridan and Hennessy (1984: 17) have observed that 'supervisory control demands that the system be trustworthy'. If we could not build automated systems that worked and could be trusted, we could not build supervisory control systems at all. Thus, the idea that the automation is trustworthy is implicit in supervisory control systems. Highly automated systems are usually large, complex, capital-intensive, and potentially dangerous, and so it is critical that they run safely and effectively. When human supervisors allow automation to control a process, we may infer that they trust that automation, to some extent at least. However, human operators are charged with the task of overriding the automation when necessary, and so they must carefully monitor its performance and learn when it is appropriate to intervene. According to conventional wisdom (e.g., Sheridan and Hennessy 1984), one of the criteria supervisors use in deciding whether to use or override the automation is their degree of trust in the automation: if their trust in an automatic controller drops beyond some point, they will override it, preferring to perform the task manually. On the other hand, if they trust the automation too much, operators may become complacent and fail to override automatic control even when the automation is faulty.

Although this paper focuses on supervisors' trust in automation, it is likely that this trust is only part of a network of trust that pervades complex, automated systems. This network of trust, involving system designers, the system, operators, management and society, is illustrated in figure 1. Designers of automated systems may allocate functions at the design stage to either the human supervisor or to the automation, depending on their trust in each of them as controllers. Similarly, part of the human operator's trust in the automation may be attributable to the operator's trust in the capabilities of the designers of the automation. Many supervisory control systems have multiple operators who may share and/or trade tasks; the trust that operators have in themselves and in one another to perform tasks may affect how tasks are allocated among them. Trust may also be an important part of the relationship between operators and management. Management hires and trains people who can be trusted to control the system, and

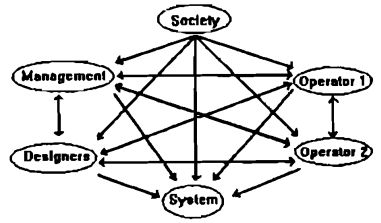

Figure 1. Network of trust in supervisory control systems.

operators are asked to trust and implement management's policy decisions regarding procedures and priorities, for example, safety/productivity tradeoffs. Finally, the plant is allowed to operate because of the trust that society has in the ability of everyone and everything involved to run it safely and effectively.

3. Why we need trust as an intervening variable

Do we need to infer trust as an intervening variable in supervisors' use of automation? Do we need an intervening variable at all? Perhaps operators simply base their allocation behaviour upon the properties of the automation: if it works, they will use it, and if it is faulty, they will not. A number of factors argue against this view.

First of all, it is too simplistic. Figure 2 is a Venn diagram of the complex, hierarchical interactions that can occur in a supervisory control system. The fact that the automation is not fully nested in the human supervisor's area means that there are some properties of the automation which supervisors will never know. How then could supervisors conditionalize their allocation behaviour on the unknown properties of this automation? Incomplete nesting also makes it difficult for supervisors to know when the automation is faulty. If the automation fails in an area outside the supervisor's knowledge base, the supervisor will fail to detect the fault, and fail to override the automation. Supervisors know that they can never have complete knowledge of the properties of the automation; they know that it may fail in unforeseen and potentially disastrous ways, and yet they realize that they still must use it most of the time. The fact that supervisors do use the automation under these circumstances implies that something else, something outside the system, is guiding their allocation behaviour.

Second, this view does not address the issue of individual differences in the use of automation. There are many examples of this in everyday life, for instance, some people use automated banking machines, while others refuse. Clearly, the source of these individual differences does not lie in the properties of the automation, since these properties remain constant across users. The source of this disparity must lie in the individuals themselves, in something they bring to the situation.

People do not simply base their decision to use automation on the properties of the automation. If we are to account for supervisors' use of automatic control in the face of necessarily incomplete knowledge of the system, and individual differences in allocation behaviour, we must appeal to something within the individual, some intervening or organismic variable that mediates between the automation and the supervisor's responses to the automation.

Experts in the field of supervisory control (e.g., Sheridan 1983a, Sheridan and Hennessy 1984) have suggested that trust is the intervening variable that mediates

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

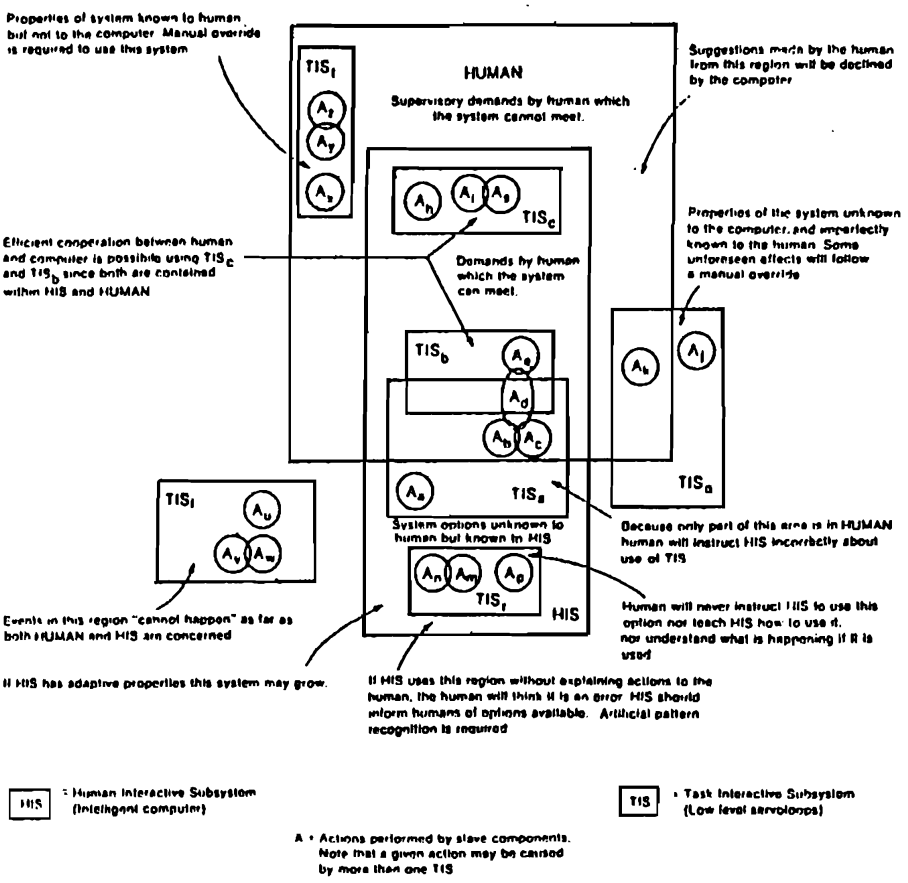

Figure 2. Venn diagram of interactions of components in supervisory control systems (*Source:* Moray 1987. Reprinted by permission of John Wiley & Sons, Inc., Copyright © 1987).

supervisors' intervention behaviour. This view has intuitive appeal because it is consistent with our own personal experiences as supervisors of other people, perhaps as a parent or a teacher or a boss. Imagine, for instance, teaching a young child to perform a task such as hammering a nail. We teach her how to hold the nail, how to use the hammer. At first, we hold our hands over hers to protect her. Then gradually, if we see her skill improving, we withdraw, leaving her to work with only minimal supervision because we trust her to be able to do the job safely and effectively. If we do not trust her to perform the task satisfactorily, we watch her closely and step in to offer help, or, at some point, decide to do the job ourselves. In the workplace, supervisors of other employees simply cannot (nor would they wish to) know every detail of their subordinates' work, yet they are still responsible for their output. To say that supervisors let staff members do their jobs without intervention to the extent that they trust them is consistent with our natural usage and understanding of the word 'trust'. The supervisory control of human staff members bears a close resemblance to the supervisory control of automation (Sheridan and Hennessy 1984), and so perhaps

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

our natural usage and understanding of the concept of trust can be applied to automated subordinates as well.

But is it sensible to speak of trust in a *machine*—an inanimate object? In fact, in everyday life, we often express our attitude toward machines in terms of trust. For example, it sounds sensible to hear a person say 'I don't use banking machines because I don't trust them', and we have a general idea of what they mean.

The hypothetical construct of trust does seem to capture a certain aspect of our attitude toward automation, and so trust may indeed be an important determinant of people's use of automation. If we wish to avoid the consequences of bad allocation decisions, then we need a theory of human trust in machines. This theory should describe the nature of human trust in machines, and how trust changes with experience on a system. It should also describe how supervisors' perceptions of the automation affect their trust in the automation, and the relationship between their trust in the automation and their use of it. Such a theory would allow us to predict supervisors' allocation decisions on the basis of their trust or on the basis of the system variables which affect their trust. It could be used to identify problems in this decision process and suggest solutions, preventing the consequences of bad allocation decisions. No such theory of human trust in machines exists. Rather than start *de novo* to develop such a theory, the approach adopted here was to take models of trust between people as a starting point and extend these to apply to the human/machine relationship.

4. A definition and model of trust in machines

There has been surprisingly little research done on the concept of our trust in other people. Furthermore, the little work that has been done is theoretical, with only a few substantiating empirical studies. It is difficult to do experimental research on trust because it is a hypothetical construct, and cannot be observed or measured in any direct sense physically. It is similar in this way to other important constructs in human factors such as mental workload and mental models. All three are 'intervening', or 'organismic', variables which reside in the human mind and mediate the human's observable responses to environmental stimuli, and many of the same problems are encountered in investigating them (Gentner and Stevens 1983, Johnson-Laird 1983). For example, it is not possible to prove the existence of an intervening variable; its existence in the operator's mind can only be inferred. It is also difficult to describe the nature of an intervening variable because it cannot be accessed or observed or measured directly. Indirect measures, such as subjective scaling or verbal protocol analysis, may not access the relevant information or may bias the operator's response in some way by virtue of the task demands they impose. Finally, operators may be unable to describe their own knowledge (Broadbent 1977) or personal experience of internal states.

In the psychological literature, trust in another person has been defined as:

- '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 *et al.* 1985);

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

- 'the degree of confidence you feel when you think about a relationship' (Rempel and Holmes 1986).

Each of these definitions seems to capture a different aspect of our everyday usage of 'trust', suggesting that trust is a multidimensional construct. The problem with these definitions is that each is too narrow, some are too vague to be testable, and they fail explicitly to acknowledge the multidimensional nature of trust which is suggested by their variety.

The different definitions do, however, have several things in common, and these commonalities are useful for identifying the necessary attributes of a broader definition of trust. First, trust is described as an expectation of another; our trust is oriented toward the future, toward predicting future gratifications, behaviours, or events. Second, our trust always has a specific referent; we trust in someone or something and our trust is specific to that referent. It is not simply a matter of one person being globally more or less trusting, in an absolute sense, than another, although, of course, this may be true. In addition, each individual discriminates among referents in a relative sense, and can simultaneously trust some referents to varying degrees and distrust others. Third, trust may relate to many different properties of the referent, including such things as their reliability, honesty, and motivations.

Barber (1983) has proposed a definition which explicitly recognizes the multidimensional character of trust, and at the same time includes the necessary attributes identified above. Barber defines trust in terms of a taxonomy of three specific expectations:

- our general expectation of the *persistence* of the natural physical order, the natural biological order, and the moral social order;
- our specific expectation of *technically competent role performance* from those involved with us in social relationships and systems;
- our specific expectation that partners in an interaction will carry out their *fiduciary obligations and responsibilities*, that is, their duty in certain situations to place others' interests before their own.

These expectations are discussed in more detail below, and are extended to the human/machine relationship.

4.1. *Persistence*

An expectation of persistence is an expectation of constancy, one so fundamental as to be overlooked in other definitions of trust, but according to Barber, it is the foundation for trust. Our expectation that nature works in a lawful, predictable way reduces the complexity in our world by limiting possible outcomes. Our expectation of constancy in natural physical laws is expressed in statements like 'the heavens will not fall' (Barber 1983: 9). It is the constancy of physical laws that allows us to understand and create mental models of physical processes, and to use these models to control a physical process and predict future system states.

4.2. *Technical competence*

The more specific expectation of technical competence is central to the meaning of trust in automation. According to Barber, there are three types of technical competence we may expect from another person: expert knowledge, technical facility, or everyday

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

routine performance. It is interesting to note that these correspond closely to Rasmussen's (1983) taxonomy of behaviour into knowledge-, rule-, and skill-based behaviour. Humans (and machines) may possess only a subset of these competencies in any given knowledge domain, and different subsets in different domains. For example, physicians may expect their patients to be competent to take medication (routine performance), but not to choose a medication (technical facility), or to interpret any response to it (expert knowledge). Similarly, supervisors may expect a human-interactive control computer to be competent to perform the routine task of communicating directives to lower level task-interactive computers, or even to be competent to alter certain system variables in response to specific 'if...then' rules, but not expect the computer to be competent to interpret or respond to unusual system states. The expectation of technical competence is probably closest to our intuitive understanding of what it means to trust a machine; machines are designed to perform a task, and we expect them to work properly—to be competent.

4.3. *Fiduciary responsibility*

The 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 when the competence of another is completely unknown. For example, when we consult domain experts such as physicians or lawyers, we are unable to evaluate them on the basis of competence, and we are forced to rely on their moral obligation not to abuse the power that they wield.

The dimension of responsibility may not seem to be applicable to machines, since there is no analogue of morality in even the most sophisticated human-interactive computers. In today's machines, the closest analogue of responsibility may be the operator's expectation that the automation will meet its design-based criteria for performance in domains where the machine has superior knowledge, autonomy, authority, and power, or when the competence of the automation is unknown.

Barber's definition of trust is detailed enough to apply to specific circumstances, and yet comprehensive enough to apply to the variety of interactions that may occur in a complex supervisory control environment. Therefore, Barber's taxonomy is adopted here as a basis for the following definition of trust in machines (and people) in automated systems:

> Trust (T) is the expectation (E), held by a member of a system (i), 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 is not necessarily isomorphic with, objective measures of these properties.

This definition can be summarized using the equation below:

$$T_i = [E_{i(Pn+Pm)}] + [E_{iTCPj}] + [E_{iFRj}]$$

In this definition, 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; FR includes notions of intention, power and authority. Expectations T and E are subscripted by the individual (i) holding the expectation to recognize explicitly that trust and its component expectations are subjective variables; the *perceived* properties which support an expectation may be quite different from a

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

referent's true properties, and so a clear distinction is made between the subjective trustworthiness and the objective trustworthiness of a human or machine referent. The expectations of *TCP* and *FR* are subscripted by the referent (j) of the expectations since expectations are specific to referents.

The arithmetic operators in the above equation suggest a simple, additive model of trust, when in fact a multiplicative model or a more complex model may turn out to be a more accurate mathematical representation. For example, the three component expectations need not be equally important; each component expectation may have to be weighted according to its importance in a particular context. For instance, the technical competence of an expert system may be more important than the issue of power and authority under normal operating conditions, but the reverse may very well be true under unforeseen emergency conditions. In addition, it is likely that the three component expectations will interact. For example, it seems intuitively reasonable that an individual's expectation of persistence might colour the expectations of competence and responsibility: one who does not believe that automation is basically good and desirable may discount any evidence of competence and/or responsibility in it. Assuming that these three expectations are exhaustive, and that the model is linear in its parameters and in its independent variables, a regression model of a supervisor's trust in a human or machine referent 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_{0-7} are parameters, $X_1 = P$ (persistence), $X_2 = TCP$ (technically competent performance), and $X_3 = FR$ (fiduciary responsibility).

This complex model is hypothetical at this point. It may turn out that a considerably less complex model will give a good fit under some sets of circumstances, for example, when one component expectation clearly dominates the others. It would be premature at this point to advocate any particular mathematical model of trust in automation, but the above model is testable and is proposed here as a subject for future research.

5. The dynamics of trust in machines

A complete model of human trust in machines must be able to account for changes in operators' trust as a result of experience on a system. There is presently no model of the development of trust in a machine, but Rempel *et al.* (1985) have proposed one for the development of trust in another person, and so their model is extended here to the human–machine relationship. According to Rempel *et al.* (1985: 111), trust in another person is a dynamic expectation which follows a certain 'developmental sequence' as a relationship progresses. They describe their model as a hierarchical stage model: at each stage, trust is based upon the outcome of earlier stages, and the stages occur in a fixed order, with *predictability* dominating early in a relationship, *dependability* dominating later, and *faith* dominating in a mature interpersonal relationship. These stages and their relevance to human–machine trust are described in more detail below.

5.1. *Predictability*

Early in a relationship, a person's trust in another person is based upon the predictability of the other's recurrent behaviours (Rempel *et al.* 1985). The perception of predictability is enhanced if a person's behaviour is both consistent and desirable, and if the functional reinforcements and restraints on behaviour are known.

According to this model, supervisors will judge the predictability of the automation

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

by evaluating the consistency and desirability of its recurrent behaviours. The assessment of predictability, and therefore the growth of trust, will depend upon three factors:

5.1.1. *The actual predictability of the machine's behaviours:* Some machines or subsystems operate within very small tolerance limits, or with few degrees of freedom, while others have larger limits or more degrees of freedom. In general, the more constrained a machine's actual behaviour is, the more predictable it is, and so trust will be inversely related to the degrees of freedom of the machine. Also, unless a machine is completely deterministic, its behaviour will be distributed with a variance about some mean, and the smaller the variance, the greater its predictability.

5.1.2. *The operator's ability to estimate the predictability of the machine's behaviours:* To estimate predictability, the machine's behaviours must be observable. A transparent system, that is, one that is easily observed and understood, should foster trust because not only are system behaviours easily observed, but also the functional relationships supporting the observable behaviours are accessible and clear. Experience may also enhance the accuracy of operators' perceptions of predictability. Experienced operators performing real life tasks are very good at recognizing the properties, or 'signatures', of systems (Bainbridge *et al.* 1974, Woods 1986), and so may be good judges of machine predictability, whereas novice operators may exhibit more of the limitations and biases exhibited in laboratory studies of judgement under uncertainty (Kahneman, Slovic and Tversky 1982, Sage 1981), such as overestimating the representativeness of small samples of machine behaviour, and anchoring on initial observations (Kahneman and Tversky 1973).

5.1.3. *The stability of the environment in which the system operates:* A machine which is sensitive to its environment may appear highly predictable in a stable environment, but suddenly appear to be unpredictable when the environment becomes unstable. Operators must learn to distinguish the apparent unpredictability which results from an unstable environment, and which should not cause distrust, from inherent unpredictability, which should.

5.2. *Dependability*

As a relationship with another person progresses, the focus changes from an assessment of their specific behaviours to an assessment of their stable disposition, specifically their dependability, or the extent to which they can be relied upon (Rempel *et al.* 1985). The process of attributing dependability is based upon the accumulation of behavioural evidence that supports perceived predictability, but with a heavy weighting placed on events that involve risk and personal vulnerability. To decide that a person is dependable, there must be an opportunity for the person to be undependable.

Extended to the human–machine relationship, we may expect that considerable experience with the automation will be required to support the attribution process, and this process will be facilitated by allowing operators to push the automation beyond its usual limits into uncertain and risky scenarios to see how dependable it is. Once supervisors decide on the dependability of a machine, they no longer need to monitor the predictability of its behavioural acts so closely, and the amount of sampling will

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

be reduced. Hence, this model predicts that knowledge about a machine's specific behaviours will be inversely related to assessments of dependability.

5.3. *Faith*

Trust has been defined as an expectation and as such it has an orientation to the future. Rempel *et al.* (1985) suggest that the final stage in the growth of interpersonal trust is a leap of faith, a closure against doubt in the face of an uncertain future. One must go beyond the available evidence to believe that another person will continue to behave in the future as they have in the past. Another person's 'past predictability and dependability . . . provide an important basis for generalizing to future situations' (p. 98), with heavy weighting given to events which provide information about the person's motives for being in the relationship. The development of faith may also be affected by certain personality variables, such as personal security and self-esteem, which contribute to a person's willingness to take emotional risks (Barry 1970).

The notion of faith also has a place in a model of human–machine trust, although the concept of motivation needs to be modified somewhat, at least with today's machines. Most processes under supervisory control are so complex that they defy complete understanding (Sheridan and Hennessy 1984). Supervisors realize the potential for unforeseen interactions in these systems and the 'brittleness' (Brown *et al.* 1982) of procedures for dealing with them. To control a process under such uncertainty implies that a supervisor must have made a leap of faith beyond the behavioural evidence generated by the automation. In supervisory control involving an intelligent mechanical subordinate, faith in automation may be based upon predictability and dependability and a supervisor's perception of the appropriateness and flexibility of the software which defines the goals and directs the behaviour of the machine subordinate. Certain personality variables may also influence a supervisor's ability to develop faith in automation.

The foregoing discussion describes trust as an ascending function of experience, fostered by a perception of consistent and desirable behaviour, particularly in risky and uncertain circumstances, and by relationship-appropriate goals. But in interpersonal relationships the road to trust may be a rocky one; we may never get past the first stage of trust with some individuals, and mature trust in another person can be betrayed. Little is known about the process by which trust is diminished, or how it may recover, although Rempel *et. al.* (1985: 111) have speculated that it may be 'notoriously easy to break down . . . [and] doubly difficult to re-establish'. They also speculate that trust, when betrayed, may be destroyed in the same order in which it was built up, with the perception of predictability giving way first, before the more resistant dimensions of dependability and faith. These are important issues for the human–machine relationship. We need to understand human–machine trust in normal operating conditions, and also how trust changes as a result of automation failure. Does trust decline gradually, or is it brittle, shattering in the face of failure? Does distrust of one subsystem generalize to reduce trust in other subsystems? If so, which subsystems, and how are they related? Does trust ever recover, and to what level? Does it ever return to its pre-failure level? To optimize supervisory control behaviour, we need to understand fully the growth, destruction, and recovery of supervisors' trust in the automation with which they work.

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

6. An integrated model of trust in machines

Barber's (1983) model of the meanings of trust and Rempel *et al.*'s (1985) model of the dynamics of trust are not inconsistent, and both perspectives have something unique to contribute to a comprehensive model of trust in automation. Barber's model provides the broader context and richness of meaning needed to characterize the myriad interactions in a complex and hierarchical supervisory control environment. Rempel *et al.*'s model provides the dynamic factor needed to predict how trust may change as a result of experience on a system.

Table 1 shows the taxonomy formed by completely crossing the two models. This integrated model is more complete than either model alone, and provides a framework for studying trust in machines in general, and trust in automation in supervisory control in particular. The strength of the integrated model lies in its ability to apply to a wide variety of interactions and circumstances, while still being specific enough to characterize any particular situation. The integrated model appears quite complex, but in reality each situation will involve only a small subset of the total number of cells. For example, the dimension of competence will likely dominate our trust in most machines, and this trust will have progressed to one of the three stages of development, thus, at any particular point in time, it will be described by a single cell in the model.

The crossing of Barber's and Rempel *et al.*'s models implies that they are orthogonal dimensions. Certainly, it would be desirable to have a model of trust in machines that incorporated both dimensions, one that could describe, explain and predict trust at different points in time. An examination of the cells in the model reveals that each appears to be unique, thus justifying the assumption of orthogonality, theoretically at least. There may appear to be some overlap in the concepts of responsibility and faith, since they both involve aspects of morality, including elusive concepts like intent and motive. However, as they are used here, they are distinct concepts. Responsibility is invoked as a basis for trust when an assessment of competence cannot be made, thus it is used by necessity, not by choice. Responsibility is a property of the referent, not the trustor. Responsibility may also be used as a basis for trust in a new encounter, as an assumption, before behavioural evidence has been generated. In contrast, faith is an expectation within the trustor, and is bestowed upon the referent, voluntarily, after extended and varied experience with the referent. There is a theoretical need to differentiate between our expectations of moral behaviour in someone we are meeting for the first time, and in someone with whom we have shared a wealth of experience.

7. Differentiating trust from other concepts

There is a problem with confusing terminology in the study of trust. To facilitate research, there is a need to operationalize the relevant concepts. Figure 3 shows how the proposed model of trust differentiates between trust, confidence, predictability, and accuracy, four concepts which are often used in this area. For example, the figure shows that perceived *predictability* is one of the bases for trust, which, in turn, is the basis for an operator to make a *prediction* about the future behaviour of a referent. The *accuracy* of that prediction may be assessed by comparing it with the actual behavioural outcome. In addition, an individual who makes a prediction may associate a particular level of certainty, or *confidence*, with the prediction. Thus, confidence is a qualifier which is associated with a particular prediction; it is not synonymous with trust.

There is a great deal of evidence to show that the accuracy of people's predictions under uncertainty can, under some circumstances, be systematically biased with respect

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

Table 1 An integrated model of trust in human–machine relationships, created by crossing Barber's (1983) model of the meanings of trust (rows) and Rempel *et al.*'s (1985) model of the dynamics of trust (columns). 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 contorm 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

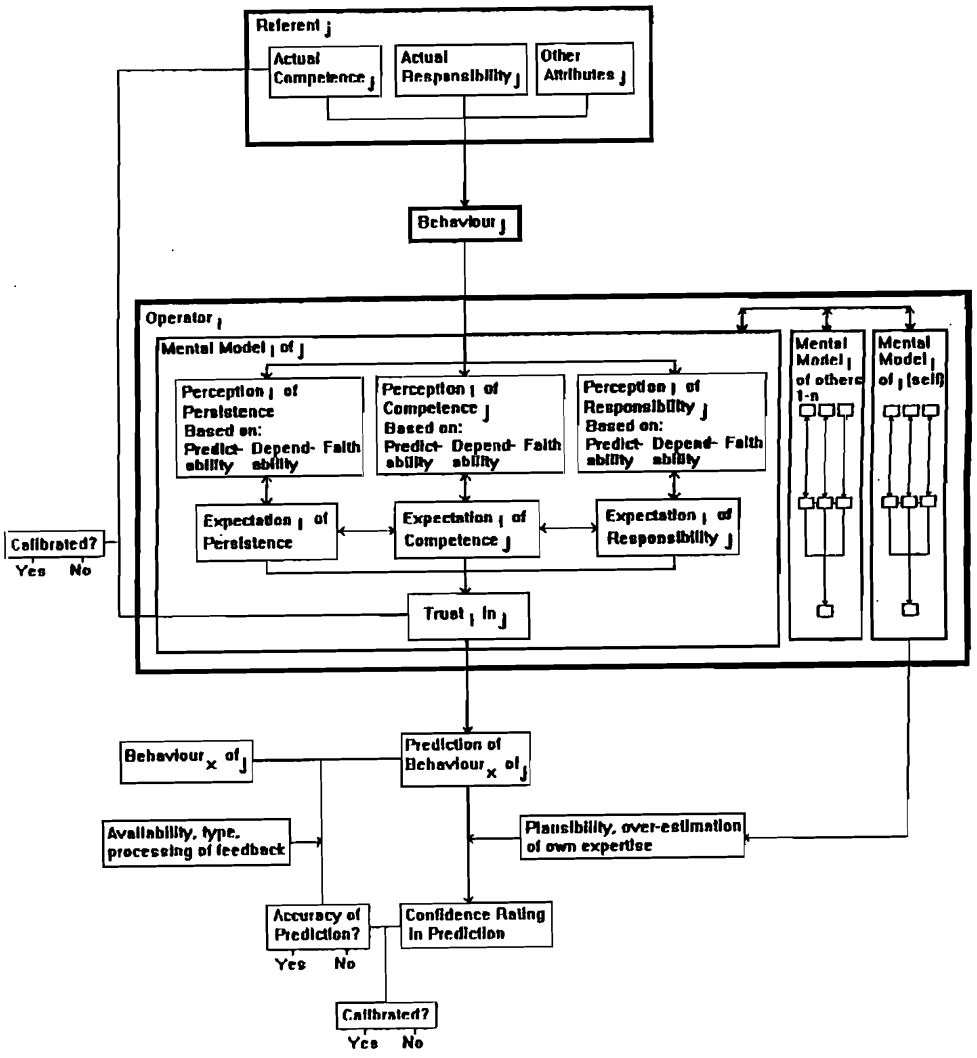

Figure 3. Model of the relationship between the automation, the operator's trust, and predictions about the automation's behaviour. The operator's mental models of others and of him or herself are shown in an abbreviated form, and contain all the cells seen within the mental model of *j*.

to normative models (Kahneman *et al.* 1982), and that their expressed confidence in predictions is not well calibrated with their accuracy (Lichtenstein *et al.*, 1982, Oskamp 1982). Wagenaar and Keren (1986) suggest that this poor calibration results from different underlying mechanisms: accuracy is related to aspects of feedback which affect the quality of the experts' mental models, whereas confidence is related to both the plausibility (rather than the accuracy) of their inferences, and the experts' overestimation of their own expertise. When supervisors make predictions about the automation's behaviour based upon their trust, we need to know whether their trust is affected by feedback on the accuracy of their predictions, or by the plausibility of their predictions, or by their estimate of their own expertise in running the system. Lee and

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

Moray (1992) are currently investigating the latter—whether supervisors' trust in their own expertise affects their intervention behaviour. The model presented in figure 3 provides a theoretical framework for interpreting this line of research, and integrating it with other research in the area of human–machine trust.

8. The calibration of trust in machines: trust, distrust, and mistrust (errors in trust)

In semiautomated systems, supervisors must choose the level of human involvement that optimizes system performance under the circumstances. If indeed supervisors' use or rejection of the automation is influenced by their trust in the automation, then their use will be optimized when their trust is at a level which corresponds to the objective trustworthiness of the automation. The process of adjusting trust to correspond to an objective measure of trustworthiness is referred to here as the calibration of trust.

Automation varies in quality; some designs are better than others at performing a given task, and even within the same machine, some functions may be performed more competently than others. Therefore, it is inappropriate for operators to trust all automation equally, or all functions or components within a single machine equally. The operators' task is to adjust or calibrate their trust to the true properties of each specific referent and then to use it accordingly. Each individual operator will have a subjective criterion of machine performance beyond which an attitude of trust will be adopted, and below which an attitude of distrust will be adopted.

Table 2 shows how a supervisor's trust and consequent choice of automatic or manual control interact with the quality of the automation to affect system performance. It is assumed here that trust and use are tightly coupled. For the sake of illustration, the quality of the automation is described qualitatively as 'good' or 'poor' although in reality quality will vary on a continuum between these two extremes. Well-calibrated operators will optimize system performance by trusting and using good automation and distrusting and rejecting poor automation. Their appropriate trust in competent and responsible automation will allow them to switch their attention to, and manually compensate for, less competent and less responsible automation which they appropriately distrust. On the other hand, poorly-calibrated operators will trust and use poor automation, and this false trust and consequent failure to reject poor automation can lead to an automated disaster. Poorly-calibrated operators may also reject good automation, based on false distrust. In a system optimized for automatic control, not only will the benefits of competent and responsible automatic control be lost, but also operators may become involved in demanding manual control, preventing them from performing their other functions in the system, and increasing the likelihood of introducing human error into the control loop.

In this analysis, errors in trust are called 'mistrust' and are of two kinds: false trust and false distrust. Accident analyses show that supervisors make both types of errors in the *use* of automation, both failing to override faulty automation, and disregarding or overriding perfectly good automation (e.g., Wiener and Curry 1980, Green and Skinner 1987). But the operators' trust in the automation was not specifically examined in these investigations, and so mistrust can only be inferred, post hoc.

One of the functions of trust is the reduction of complexity and uncertainty (Luhman 1980). Through experience, we build up expectations of other people's competence and responsibility, and an expectation of persistence. These expectations serve to limit the universe of possible behaviours of a referent to a small subset of potential behaviours,

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

Table 2. How the operator's trust in and use of the automation interact with the quality of the automation to influence system performance. The cells exemplify appropriate trust, appropriate distrust, and the two errors of mistrust (false trust and false distrust).

Operator's trust and allocation of function	Quality of the automation 'Good'	'Poor'
Trusts and uses the automation	*Appropriate trust*, optimize system performance	*False trust*, risk automated disaster
Distrusts and rejects the automation	*False distrust*, lose benefits of automation, increase operator's workload, risk human error	*Appropriate distrust*, optimize system performance

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

thus reducing uncertainty and complexity. But distrust also reduces uncertainty, and thus it may be viewed, in a psychological sense, as a functional alternative to trust. If we distrust another, our expectations are of technical incompetence or failure to meet fiduciary obligations and responsibilities, and these, too, are a subset of all possible behaviours. Thus, distrust is not the absence of an expectation, but rather it is the presence of an expectation of incompetence and/or irresponsibility.

The functional equivalence of trust and distrust in reducing uncertainty suggests some interesting predictions about supervisors' monitoring behaviour. Specifically, it suggests an inverted U-shaped function, with automation that is highly distrusted or highly trusted being monitored infrequently because uncertainty is low, whereas middling values of trust will be associated with high uncertainty and result in close monitoring of that automation. The prediction that supervisors will infrequently monitor distrusted automation seems, at first, to be counterintuitive. But the reason for this is that, if possible, supervisors will override distrusted automation, and perform the task manually. This effectively eliminates the distrusted automation from the system and thus prevents operators from monitoring its behaviour. If supervisors must continue to use distrusted automation, then they will no doubt continue to monitor its behaviour closely. It is also likely that supervisors will closely monitor all new or unfamiliar automation until they have gathered enough evidence to reduce their uncertainty enough to adopt an expectation of trust or distrust toward it.

Distrust may be more resistant to change than trust because of the different allocation of functions appropriate for each expectation. Distrust induces supervisors to take manual control which, in turn, effectively eliminates the distrusted automation from the system and thus prevents operators from gathering any further, possibly disconfirming, evidence on its behaviour. In contrast, an expectation of trust releases supervisors from lower-level control and affords them the opportunity to continue to observe the behaviour of a trusted subsystem and to update their expectations if indicated.

9. Recalibrating trust in machines

With experience on a system, operators will learn to calibrate their trust to the automation. However, they may be less than perfect in this process, and instances of mistrust will surely occur. Some supervisors may have a systematic bias toward trust or distrust; that is, their expectations may reflect the relative competence of different subsystems, but be too trusting or too distrusting in an absolute sense. Alternatively,

a supervisor may be well calibrated in general, but be prone to being overly trusting or distrusting of a particular subsystem. To the extent that trust and the operator's allocation of functions are related, these biases will result in inappropriate intervention behaviour. To avoid errors of mistrust and inappropriate allocation choices, these biases must be overcome; the supervisor's trust must be recalibrated to correspond more closely with objective measures of the automation's trustworthiness.

How can a supervisor's calibration be improved? Since developing an expectation of trust involves a learning process, it should be possible to modify operators' trust by retraining them. The integrated model developed in this paper can be used to make some suggestions for improving calibration, outlined below. These recommendations must be regarded as tentative, pending empirical support for the proposed model of trust. They are, however, consistent with other researchers' recommendations for improving the fit between people and automation (Rouse and Morris 1986, Sheridan 1980, Sheridan *et al.* 1983b, Zuboff 1988).

9.1. *Identify the referent, and selectively recalibrate the operator on the dimensions of trust which are poorly calibrated*

Identify the specific referent for which the operator's trust is poorly calibrated. Determine whether the source of the problem lies in the expectation of persistence, competence, or responsibility, and selectively retrain on this aspect.

9.2. *Improve the accuracy of the supervisor's perception of the trustworthiness of the automation*

Provide objective data about the behaviour of the automation, including its performance over time, and the system constraints and environmental disturbances which affect it. If necessary, improve the transparency of the automation's behaviour. Strive to achieve a match between the level of abstraction at which the behaviour of the automation is transparent and the level of abstraction of the supervisor's expectation of trust, which increases as trust grows. For example, a novice supervisor's trust will be at the relatively concrete stage of establishing the predictability of components, and for this purpose the supervisor needs very detailed information, at the level of individual behavioural events. Educate the operator about the design-based goals and intentions of the automation. Allow the operator to experience how the automation responds to risky and unfamiliar scenarios, using a simulator if necessary. Provide plenty of time and practice for the operator to explore the system, to expand the operator's knowledge of it, and to provide a basis for the development of accurate expectations.

9.3. *Modify the supervisor's criterion of trustworthiness*

Does the supervisor expect too much or too little from the automation? If so, is this an honest difficulty, or is it motivated by other factors, such as a fear of being replaced by the automation? Educate the operator about the machine's domains of competence, its actual history of competence and responsibility, and make explicit the criterion level of acceptable performance.

9.4. *Enhance the supervisor's ability to allocate dynamically functions in the system*

If supervisors are to accept responsibility for system performance, they must also have autonomy and authority over the automation. Emphasize that the supervisor is in control

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

of the automation, no matter how apparently 'smart' it may be, and design systems accordingly, giving supervisors the dignified task of dynamically allocating functions at their discretion. Recognize that supervisors will intervene, perhaps sometimes inappropriately, and provide support for them in manual control mode.

9.5. *Recognize that distrust may be difficult to overcome*
This may be particularly true if the operator experiences the failure of trusted automation, and can manually override it.

9.6. *Introduce automation carefully*
The careful, informed introduction of automation should prevent errors of mistrust and reduce the need for later recalibration.

10. Conclusion

The integrated model of human trust in automation which is presented in this paper is based upon models of interpersonal trust. The latter seem to transfer quite well to the human–machine relationship, but empirical testing is needed to confirm this. The proposed model provides a theoretical foundation for experimental studies of human trust in automation, including two by the author which are described in Muir (1989) and to be reported in Part II of this article.

Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council of Canada. The author is grateful to her doctoral dissertation supervisor, Neville Moray, and to two anonymous reviewers for their helpful comments on an earlier draft.

References

BAINRIDGE, L., BEISHON, J., HEMMING, J. H. and SPLAINE, M. 1974, A study of real-time human decision making using a plant simulator, in E. Edwards and F. Lees (eds) *The Human Operator in Process Control* (Taylor & Francis, London).

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

BARRY, W. A. 1970, Marriage research and conflict: an integrative review, *Psychological Bulletin*, **73**, 41–54.

BROADBENT, D. E. 1977, Levels, hierarchies, and the locus of control, *Quarterly Journal of Experimental Psychology*, **29**, 181–202.

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

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

GENTNER, D. and STEVENS, A. L. 1983, *Mental Models* (Erlbaum, Hillsdale, NJ).

GREEN, R. and SKINNER, R. 1987, *The Log* (British Airline Pilots' Association monthly journal, October).

HALPIN, S., JOHNSON, E. and THORNBERRY, J. 1973, Cognitive reliability in manned systems, *IEEE Transactions on Reliability*, **R-22**, 165–169.

JOHNSON-LAIRD, P. N. 1983, *Mental Models* (Harvard University Press, Cambridge, Massachusetts).

KAHNEMAN, D., SLOVIC, P. and TVERSKY, A. (eds) 1982, *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).

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

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012

LEE, J. and MORAY, N. 1992, Trust, control strategies, and allocation of function in human–machine systems, *Ergonomics*, **35**, 1243–1270.

LICHTENSTEIN, S., FISCHHOFF, B. F. and PHILLIPS, L. D. 1982, Calibration of probabilities: the state of the art to 1980, in D. Kahneman, P. Slovic and A. Tversky (eds) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).

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

MORAY, N. 1987, Monitoring behaviour and supervisory control, in K. Boff, L. Kaufman and J. P. Thomas (eds) *Handbook of Perception and Human Performance* (Wiley, New York).

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

MUIR, B. M. 1989, Operators' trust in and use of automatic controllers in a supervisory process control task, Doctoral dissertation. University of Toronto, Canada.

OSKAMP, S. 1982, Overconfidence in case-study judgments, in D. Kahneman, P. Slovic and A. Tversky (eds) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).

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. and HOLMES, J. G. 1986, How do I trust thee? *Psychology Today*, February, 28–34.

REMPEL, J. K., HOLMES, J. G. and 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. 1977, Human computer interaction in multitask situations, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-7**, 384–392.

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

SAGE, A. P. 1981, Behavioural 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 behavioural interdependence, in R. L. Burgess and T. L. Huston (eds) *Social Exchange in Developing Relationships* (Academic Press, New York).

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

SHERIDAN, T. B. 1986, Supervisory control, in *Handbook of Human Factors* (Wiley, New York).

SHERIDAN, T. B., FISCHHOFF, B., POSNER, M. and PEW, R. W. 1983a, Supervisory control systems, in *Research Needs for Human Factors* (National Academy Press, Washington, DC).

SHERIDAN, T. B. and HENNESSY, R. T. (eds) 1984, *Research and Modeling of Supervisory Control Behavior* (National Academy Press, Washington, DC).

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

TOFFLER, A. 1980, *The Third Wave* (William Morrow, New York).

WAGENAAR, W. A. and KEREN, G. B. 1986, Does the expert know? The reliability of predictions and confidence ratings of experts, in E. Hollnagel, G. Mancini and D. D. Woods (eds) *Intelligent Decision Support in Process Environments* (Springer-Verlag, Berlin).

WIENER, E. L. and CURRY, R. E. 1980, Flight-deck automation: promises and problems, *Ergonomics*, **23**, 995–1011.

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*, Atlanta, 398–401.

ZUBOFF, S. 1988, *In the Age of the Smart Machine: The Future of Work and Power* (Basic Books, New York).

Downloaded by [Stanford University Libraries] at 04:34 13 September 2012