Laboratory studies of trust between humans and machines in automated systems

by Neville Moray* and T. Inagaki†

In supervisory control, operators are expected to monitor automation and to intervene if there is an opportunity to improve system productivity or if faults develop which cannot be managed by the automation. Central to how humans interact with automation is the degree to which they trust the system to perform well and handle unforeseen events. This paper summarises recent laboratory experiments and theoretical models, both quantitative and qualitative, of the dynamics of trust between humans and machines and discusses the calibration of trust and the problem of allocating responsibility for control between human and machine.

Keywords: Human–machine interaction; trust; automation; function allocation.

1. Introduction

The increasing use of automated systems has implied a change in the role of operators from manual controllers to supervisory controllers (Sheridan, 1976, 1992, 1997). A central problem of supervisory control is deciding when operators should intervene. Automation is optimised on the assumption that operators will not control the plant manually, but even so they are expected to intervene in three situations, (1) when the plant drifts from set points despite the automatic controllers, (2) when an opportunity arises to increase productivity beyond what the optimised algorithms can realise, and (3) in fault management of beyond-design-basis conditions. In all plants the human operator is regarded as the last line of defence against hazards caused by equipment failures.

This mixture of roles carries with it well known problems, since remaining tasks which are not automated become ever more difficult for the operator (Bainbridge, 1983), and central to the decision of when to intervene is the problem of *trust*. If operators trust automation they will let it control the plant. If they distrust it, then they will tend to intervene. The importance of trust was recognised early by Sheridan but only recently has systematic research into the role of trust appeared. This paper will summarise some recent research and modelling. It concentrates on experiments directly relevant to process control where quantitative modelling is most advanced, but we may note that currently there is great interest also in other systems where operators monitor highly automated systems for long periods, especially in aviation.

2. Domains of research on trust

No systematic taxonomy has been developed for research on trust but some distinctions can be made which may be important, since modes of intervention differ considerably. For example, a chemical process which develops a fault can often be shut down quickly and then need relatively little intervention after shutdown, whereas an automated civilian airliner which develops a fault must be brought safely to the ground. On the other hand, no flights last longer than about 15 h, whereas a nuclear power reactor must be monitored for many days after a scram because of the decay heat which is generated from residual radioactive by-products of fission. Take-off lasts a few seconds: the start-up of a process control plant may take several days.

The following kinds of tasks can be distinguished:

1. Continuous closed loop systems with slow dynamics (process control);
2. Continuous closed loop systems with fast dynamics (aircraft and vehicles in general);
3. Continuous closed loop discrete systems (discrete manufacturing);
4. Open loop cognitive decision making aids ('expert systems').

It is also useful to distinguish three aspects of a system to which trust may be given:

1. Displays;
2. Information processing and decision making;
3. Control systems and effectors.

Finally it is useful to distinguish four kinds of investigations:

1. Field studies of real or simulated systems such as aircraft or industrial plant (Sarter and Woods, 1997; Zuboff, 1988).
2. Simulators which provide high fidelity physical and dynamic copies of real systems, such as those used in aviation and nuclear power plants training.
3. 'Microworlds' which resemble real systems and include closed loop coupling of humans and automated control in real time, involving simulated physical causality, but are not in general simulations

*Department of Psychology, University of Surrey, Guildford, Surrey, UK.
†Institute of Information Science, University of Tsukuba, Japan.

of any real system (Lee and Moray, 1992, 1994; Moray *et al*, 1999). Such research is the main topic of the present paper.

4. Arbitrary laboratory tasks which may not have any control element, and whose structure is arbitrary, designed simply to provide an experimental environment for making behavioural measurement. No realistic coupling or causality is involved (Parasuraman *et al*, 1993, 1996; Riley, 1994).

In general the possibility of applying the results of research to real industrial processes decreases from 1 to 3 of this list. For a particularly interesting discussion of the relation of these three types of research environment, see Rasmussen *et al* (1995). It is encouraging that research on trust seems to produce broadly similar results in all three kinds of environments. Where differences are found, they can be related to the environment in a meaningful way.

Finally, we need to define *trust*. We use the term to refer to an attitude of mind towards an agent with whom a human operator is collaborating. The agent may be another human or a machine (automated sensor, automated controller, computer hardware, software programs or software 'agents', etc.). By trust we mean an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of its designers' intentions, be relied on to achieve the design goals. No formalism has been developed for trust, but for discussions of what is implied and qualitative models see particularly Muir (1994) and Cohen *et al* (1998).

Trust is typically measured by subjective ratings. It is important to note that, when properly constructed using sophisticated scaling techniques, subjective scales are reliable and valid, and can be shown to have equal interval or even ratio metrics. See, for example, McDonnell (1969). A typical scale will ask operators questions such as, 'On a scale from 1 ("not at all") to 10 ("completely") how much do you trust the feedstock pump to keep the inventory level constant?'. Reliable and repeatable data can be obtained when appropriate subjective scales are used. The scores thus obtained can be used as dependent or independent variables in investigating how trust affects operator behaviour. Throughout this paper when we refer to ratings of trust we will assume that some such subjective scale has been used.

3. A summary of empirical findings

Zuboff (1988) in her sociological study of the impact of office automation on workers' attitudes and practices includes important field observations about the way in which trust in equipment is affected by the way in which automation is introduced; but major experimental work on trust began with the studies by Moray and his collaborators (Lee, 1991; Lee and Moray, 1992, 1994; Muir, 1989, 1994; Muir and Moray, 1996). This group used a microworld PASTEURISER in which supervisory control was exercised over a feedstock pump, a steam pump, and the power supply to a boiler in order to maximise the output of pasteurised product of a simulated pasteurisation plant. Each of the three subsystems could be run either under automatic or manual control, and faults such as random perturbation of pump rates, leaks, or incorrect displays of sensor values could be introduced either transiently or for long periods. Operators were trained on the task over periods from a few minutes to several hours, and experimental runs varied from a few minutes to about an hour. Throughout these studies productivity was measured as the percentage of the raw feedstock which was successfully pasteurised.

Using PASTEURISER, Muir (1989; Muir and Moray, 1996) found that operators were able to identify which parts of the plant were faulty, and to change their trust in those components while leaving their trust in the reliable components unchanged. Muir examined cases where the feedstock pump could be unreliable due either to a faulty display (the displayed value was not the actual pump speed), or to a faulty control (the actual rate was not that commanded), using both constant errors and variable errors in addition to a fully normal pump. Trust was affected both by faulty displays and faulty controls, and a fault in controls changed the way in which faulty displays affected operators and vice versa. Relatively small faults of varying magnitudes reduced trust more than did a larger constant error, and with experience constant errors could be compensated manually and trust tended to recover. Faults consisted of random variations in the feedstock pump rate, in the display which represented the pump rate, or both, depending on the experiment. Random numbers were added to the actual rates in the range from $\pm 5\%$ to $\pm 30\%$ (small to large faults).

Faults in one aspect of a subsystem therefore can affect trust in other aspects, but Muir found that the effect of unreliability did not spread beyond the subsystem to the rest of the plant. This may be related to the phenomenon of 'cognitive lockup' or 'cognitive tunnel vision' (Moray, 1986). During fault management operators tend to fixate on one part of the plant and do not consider the possibility that the real cause of the problem is elsewhere, looking for confirmation of their diagnosis rather than testing their hypothesis. Although this is usually taken as evidence of poor fault management strategy, Moray (1981) argues that it is a rational strategy, since faults are most likely to be caused by components to which the faulty variables are tightly coupled. Muir's work suggests that trust in a component is restricted to others to which it is tightly coupled.

Using a 100 point scale to measure trust, she found the following relations:

$$\text{trust} = 83.0 - 49.0 \log_{10} (\text{fault magnitude})$$

for constant faults, and

$$\text{trust} = 70.9 - 42.0 \log_{10} (\text{fault magnitude})$$

for variable faults. The constants probably depend on the particular fault magnitudes, whereas the logarithmic relation probably generalises.

An important result was that the frequency with which the displayed pump rate was monitored also varied as a function of trust and error magnitude. The pump rate was monitored every 1.5 iterations of the simulation when a large constant error was present and at the same frequency when a small random error was added to the constant error; but this sampling rate fell to one sample every 9.1 iterations when there was no error. This is particularly interesting in the light of so-called 'automation induced complacency' which we shall consider later. Finally, she found that the proportion of time for which the automated control was used varied as a function of trust:

% time in automatic control
  = 3.0 + 0.80 (subjective trust rating).

Muir had allowed operators to choose the mode of operation only of the PASTEURISER feedstock pump, but Lee and Moray (1992, 1994) allowed their operators to manage all three controllers. They did not find Muir's simple relation between trust and time in automation, but found that they had to take account also of the operators' self-confidence in their ability to control the plant in manual control mode. Lee (1991) fitted a time series model to the data, and found the following relationships for trust in a feedstock pump which showed occasional random disturbances:

$$\begin{aligned} T_n = {} & c_1 T_{n-1} + c_2 \text{Productivity}_n \\ & + c_3 \text{Productivity}_{n-1} + c_4 \text{FaultSize}_n \\ & + c_5 \text{FaultSize}_{n-1} + v \qquad \ldots(1) \end{aligned}$$

$$\begin{aligned} \%AU_n = {} & k_1 AU_{n-1} + k_2 AU_{n-2} + k_3(T - SC)_n \\ & + k_4 \text{FaultSize}_n + k_5 \text{Bias} + v' \qquad \ldots(2) \end{aligned}$$

where T is the subjective rating of trust, SC the subjective rating of self-confidence, AU the percentage of time in automation control, and Bias a factor indicating that some operators prefer one mode to the other, the $c(\,.\,)$ and $k(\,.\,)$ are weighting coefficients, and v and v' are residual error terms.

Although they did not develop an equation for self-confidence as a function of plant state, we assume that a similar equation could be developed. Indeed Moray *et al* (1999) working with a simulated air-conditioning plant called SCARLETT found the following equations:

$$\begin{aligned} T_n = {} & 0.028R + 0.69T_{n-1} - 1.01F_n - 0.33F_{n-1} \\ & - 0.46D_n - 0.32D_{n-1} \qquad \ldots(3) \end{aligned}$$

$$SC_n = 0.007R + 0.88SC_{n-1} - 0.92A_n + 0.72A_{n-1} \ldots(4)$$

where R is the % reliability of the automated fault diagnosis, F the presence or absence of a false diagnosis, D the presence of a disagreement between the operator and the automated diagnosis, and A the occurrence of an accident to the plant as the result of a fault. F, D and A are (0,1) variables.

Figure 1 shows data averaged over a group of controllers for the acquisition of control skill and trust in PASTEURISER. On Trial 16 a transient fault developed, and after Trial 40 the fault returned and remained for the rest of the experiment. Note the lags in the effect of faults on trust, due to the inertia represented by the memory of trials at $n - 1$ and $n - 2$ in the time series model. Because of this performance recovers faster than does trust. The trust recovery curves following faults seem to be identical to the original acquisition curves in Lee's experiment, suggesting that trust recovery is no harder than its initial acquisition.

Figure 2 shows data from a single operator. At Trial 16 faults began to appear in the feedstock pump when it was in manual mode. Self-confidence decreased, trust increased, and the operator switched to automatic control. On Trial 20 faults began to occur only during automatic control. The relation between trust and self-confidence was reversed, and when $T - SC$ became negative, manual control was again used.

All studies that have looked for it have found a relation of trust to system reliability (Hiskes, 1994; Moray *et al*, 1999; Riley, 1994; Tan and Lewandowsky, 1996). The question of whether trust alone determines the time spent in automated control, or whether self-confidence in the ability to use manual control is also important seems to vary as a function of the system, and may be also a function of the amount of experience the operators gain in the two modes of control, manual and automatic. It is important to note that if the time spent in automation is strongly affected by trust and self-confidence, that is, the relation of the latter determines when manual intervention takes place, then it becomes vital that operators have a chance to operate the plant under manual control, since without such experience their self-confidence will not be well calibrated. Some

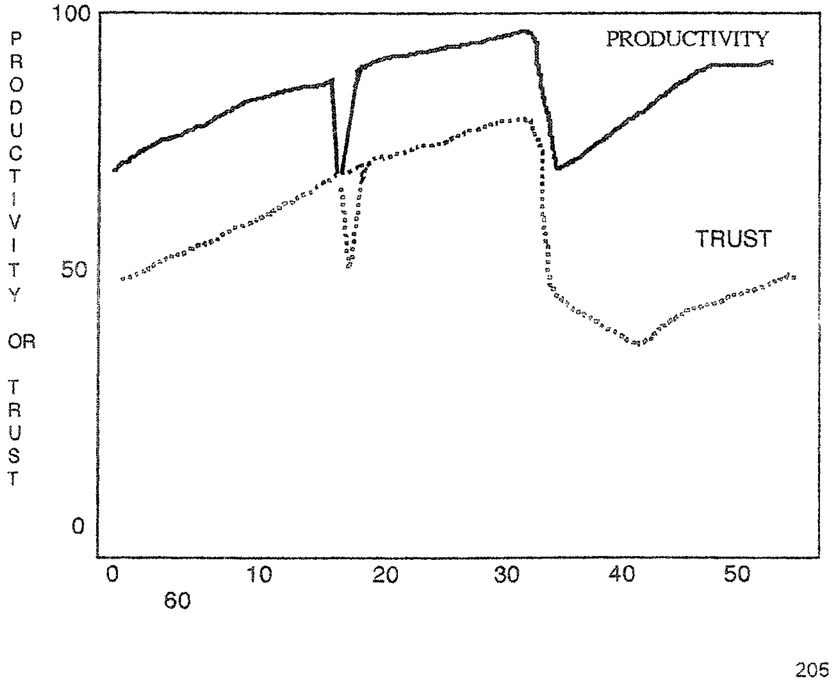

Fig 1 The acquisition of trust and performance in PASTEURISER process control. The curves are based on the average for a group of 16 operators. A transient fault occurred on Trial 16, and a permanent fault on Trial 40. After Lee (1991)

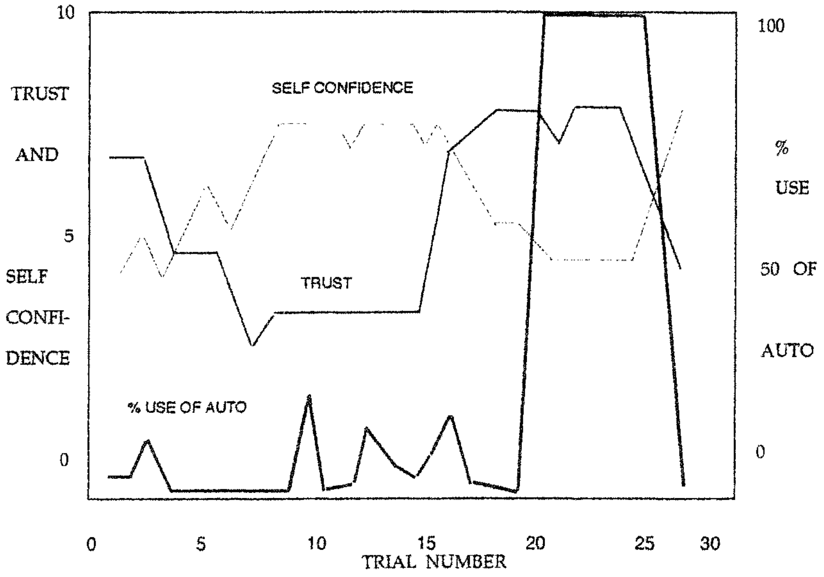

Fig 2 Dynamic function allocation by a single operator in the face of faults under automated or manual control. At Trial 16 faults appeared during MANUAL control, T-SC went positive, and AUTO was selected. At Trial 26 Manual faults ceased and AUTO faults began, and the trend was reversed, leading to a return to MANUAL control. After Lee (1991)

evidence suggests that lack of experience may lead to under-confidence, and hence an unwillingness to intervene when it is necessary. This need to practise manual control, either on the plant or on a high fidelity simulator, is in addition to the need which is already recognised to develop and maintain manual control skills for fault management. Other workers using PASTEURISER have found similar results, although Eidelkind and Papantonopoulos (1997) found much less recovery of trust. However, their task was a pure monitoring task rather than a control task closed either through the operator or the automation, and this may account for the difference. Such methodological differences must be noted when generalising from laboratory studies to real systems.

Most research shows that trust in automation increases during early experience of a system. One exception is some of Muir's data. In addition to asking for ratings of trust, she asked for ratings of 'faith in the ability of the system to perform well', meaning to what extent did operators believe that in future situations as yet unencountered would the automation handle any problems which might occur. Surprisingly, faith was high at the start of the experiment, then fell, and later began to recover. It seems likely that this is due to assumptions which operators bring to new tasks. If asked to operate equipment, they may well assume that it will be good and effective or they would not be asked to use it. Experience then shows that the equipment may be less good than expected in some situations, and so faith falls, later to recover as the operators learn how to deal with the idiosyncrasies of the automation.

Lee's time series model indicates that only the very recent past has an effect on the levels of trust, and hence on the extent to which automation is used, but all the studies show that trust is not at zero on the first trial and it is likely that operators bring to any new task some initial level of trust in machines based on their past general experience. Eidelkind and Papantonopoulos (1997) varied the cost of poor performance and found that this biased initial trust: the more costly a failure of the automation to perform the task, the less it was trusted prior to any experience of its quality.

All studies have also found very large individual differences among operators, both for their ratings and their behaviour. The models described in this paper are models of grouped data, not of individual operators. It is not currently known whether individual differences reflect inherent stable psychological differences in attitudes, or rather depend on different rates of learning and constructing mental models of how the system works. This is an important issue, since if they are stable traits it would mean that the importance of personnel selection is increased and training should be tailored to individuals (Cohen *et al*, 1998; Riley, 1994).

All studies have found that while the level of trust asymptotes to a fixed final value in the absence of plant failures, it is very volatile in the face of faults and perturbations in the system and the experience of even transient automation failures. As Muir states in Muir and Moray (1996), '... consequences need not be *consistently* negative to diminish trust; trust is fragile and any instances of even *temporary* loss of control are sufficient to reduce it ... Operators' trust is apparently conditionalised on the worst observed machine behaviours ... behaviour must be both desirable and consistent to foster trust.'

Several studies have found that an important determinant of trust is disagreement between the automation and the operator. This may occur particularly where there is automated detection and diagnosis of faults. Moray *et al* (1999) investigated automated fault management in a simulated central heating system, and found that when operators disagreed with the diagnosis made by the automation, trust declined, even if the fault was successfully managed by the automation. A similar effect was found by Muir, and by Hiskes (1994). Hiskes used a discrete manufacturing microworld called UIUC-CIM, in which parts were conveyed from one machine to another in an assembly task by an automated guided vehicle (AGV). Operators could intervene to change the schedule governing movements of components from machine to machine. As with PASTEURISER, faults developed after some hours of initial experience had been gained by operators with reliable automation. The AGV occasionally failed to take a part to the machine which it was scheduled for, and instead dropped it off at the first machine which had a vacant buffer.

Hiskes found results similar to those found in the continuous process control of PASTEURISER, and while he could not develop a time series model for technical reasons to do with identifying suitable time indices for system state, he developed a multiple regression model which was rather similar to those developed by workers using PASTEURISER. He added an extra subjective variable which had not been required by Lee, finding that an important variable was the extent to which operators believed that it would be possible to improve productivity by intervening. This varied even though self-confidence in their ability to intervene did not change in response to levels of reliability. A major difference between Hiskes's results and those of Lee was that Lee's operators took over manual control almost as soon as the value of T – SC went negative, whereas Hiskes's operators did not take over until T – SC approached –5.0 (averaged over all operators). It seems likely that the great differences between the nature of discrete manufacturing and continuous process control is the reason, or that manual control of the discrete process was extremely difficult, but this result deserves more research.

These models are strong. Those of Lee accounted for more than 75% of the variance in the use of automation, those of Moray *et al* for around 75% of the variance in trust and self-confidence, and Hiskes's model, when restricted to those operators who actually did intervene manually (since some did not), for nearly 70% of the variance. One is therefore inclined to think that factors which have been found in these studies influencing trust between human and machine and the effects of that trust in determining whether automation will be used, may well play roles in real processes beyond the laboratory.

The final empirical study to be considered is by Tan and Lewandowsky (1996). It is particularly interesting because it suggests that as software 'agents' become more common, machines begin to behave more and more like humans in their relations to human operators, and the 'socialisation' of machines continues, there may be important effects arising from the sociotechnical dynamics of automation. It is common to hear of 'human-centred automation', and for analogies to be made between intelligent software agents and humans. Tan and Lewandowsky (1996) show that there may be subtle effects on human–machine relations from such factors which need further study.

Tan and Lewandowsky repeated the experiment of Lee and Moray (1992, 1994) but compared their operators' response to two forms of co-operative supervisory control. In the first condition they duplicated Lee's experiment and obtained almost identical results. In the second condition they told their operators that they could either allow another human operator in an adjacent room to run the system, or they could take control. That is, they were exercising supervisory control over another human (who was very experienced) rather than over an automated agent. The most important result was that operators were more tolerant of error when they believed they were co-operating with a human. In fact, they were always co-operating with the automation, and the systems were identical in the two conditions: only the 'cover story' had been changed. No human was present in the other room, but their belief that their co-worker was human changed the dynamics of trust, and the amount of time they passed control to the other operator.

4. General models of trust between humans and machines

Five kinds of models have been developed to describe and predict trust between humans and machines.

4.1 Regression models

Most studies use multiple regression to identify the independent variables that affect trust or the percentage of time spent in automatic control (Hiskes, 1994; Lee, 1991; Lee and Moray, 1992, 1994; Muir, 1989; Muir and Moray, 1996). These models do not capture the dynamics of trust and allocation of function.

4.2 Time series models

Lee (1991), Lee and Moray (1994) and Itoh (Moray *et al*, 1999) model the dynamics of trust using time series. While the significant variables differ from task to task, the models are similar, and a striking finding is the agreement among researchers that not more than one or two steps back in the past need to be included in the time series models: that is, while the recent past does affect current trust and current allocation of function, it is only the very recent past which is relevant. There are reasons to think that operators bring to their task long term expectations which heavily bias trust and self-confidence, and that the high frequency dynamics are modulating a low frequency long-term attitude to automation in general and the particular plants with which operators work. The models agree that the current value of trust affects the future value for one or two periods, and that current and recent past measures of automation effectiveness (productivity, magnitude of faults, quality of fault diagnosis, etc.) similarly affect future trust, and hence dynamic allocation of function. There is some indication that disagreements between operator and automation affect trust.

4.3 Qualitative models

Muir (1994) and Riley (1994) have each proposed a general qualitative model for trust in automation. These are useful for summarising what is known about variables which affect trust, and can be used as heuristics for guiding research, but are not sufficiently precise to make quantitative predictions.

Muir (1994) based her model on the literature concerning trust between humans, and applied the same concepts to the growth of trust between humans and machines. Drawing on the work of Remple *et al* (1985) and Barber (1983), she proposed that predictability and dependability are crucial requirements if a system is to be trusted. Predictability is the ability to know what the system is about to do, and dependability is the ability to rely on the system continuing to behave as it has in the past even when the situation changes within design parameters. If these are present, ultimately the operator comes to have faith in the system in the sense that it will be trusted to perform even in situations which have never been encountered before. Muir's data (Muir and Moray, 1996) support these predictions except that faith was higher than expected on first encountering the system, and then fell before recovering in accordance with the model. As stated earlier, this may indicate an assumption by operators that any system they are asked to control is likely to be well designed and effective, and that discovering the reality of a system causes a transient loss of trust until its idiosyncrasies have been mastered.

An important prediction from Muir's model is that in order to develop and calibrate their trust in a system, operators need to experience the system when it is dealing with problems such as faults and perturbations. Trust may not be well-calibrated in the absence of opportunities for the automation to fail.

It is interesting in the light of Muir's model to look at recent work on automated avionics systems in civil aircraft. Sarter and Woods (1997) found that in highly automated cockpits many pilots did not trust the automation precisely because it was unpredictable. Many reported that they had been surprised by how the automation behaved on at least one occasion, and several reported that they did not ever come to understand why it behaved as it did. There is good evidence that several crashes of modern airliners have been due to a failure to understand the behaviour of the automation, and it is clear that a critical problem is that the feedback to the pilot of system state is currently insufficient for state identification. The result is both a loss of trust and inefficient coupling, resulting in competition for control between human and automation. In conversations with pilots and plant operators we have found that they speak of the difficulty of operating if their trust in the automation has suffered due to poor feedback. Muir's model captures well several of the findings in this domain, although it does not lead to detailed design predictions.

Riley's model is a generalised influence diagram. Although developed on the basis of laboratory experiments which did not simulate mechanical systems, it seems to capture many characteristics of what is known about trust and human–automation coupling, and suggests further candidate variables for more elaborate quantitative models, including workload, perceived risk, skill, task complexity, etc.

4.4 APT – Argument-based probabilistic trust

Recently Parasuraman and his group (Cohen *et al*, 1998) have proposed a qualitative model of trust which can in principle be realised as a computational model to predict human–automation interaction. Its domain of application is particularly the use of automated decision aids, or so-called 'expert systems' which can be thought of as automated cognition, rather than direct control of continuous or discrete industrial processes. Their model is based on information value theory, and uses evidence to reduce uncertainty by looking at the backing assumptions and warrants for the evidential claims. The output of the model is a degree of certainty, a probability that a particular course of action will succeed, that a particular diagnosis is correct, and how much one can trust the decision aid's suggestions. The model was developed in the context of military tactical decision making, but has clear applications to fault diagnosis and fault management, particularly in complex high technology systems where the system state is difficult to assess.

4.5 Neural net model

Lewandowsky (personal communication) has recently applied neural net modelling to the development of trust in the PASTEURISER microworld. Neural nets looked at plant state and calibrated themselves for trust by recognising normal and unusual states of the plant. This work is in its early phase, but is of considerable interest.

These models for trust seem to make different assumptions about the degree to which the operator is making conscious decisions. Neural nets do not require conscious deliberation. The time series models of Lee, and the regression models in general, such as that of Hiskes, are models of data, and while predictive, make no assumptions about how the assessment of evidence and estimates of trust and self-confidence are made. Being models of data, the coefficients will change from one data set to another, although it seems likely that the general form of the equations has application across tasks and even across domains. These are not models of psychological process but of the behaviour of human–machine systems. On the other hand, both Muir's and Riley's models, and APT are models of psychological process, and at least in the form in which it has been described (Cohen *et al*, 1998), APT implies a high level of conscious cognitive activity by those assessing the evidence. One might relate the types of trust models to Rasmussen's taxonomy of behaviour (Rasmussen, 1986). The regression, time series and neural net models require no more than skill-based behaviour, APT is a model of knowledge-based behaviour, and the models of Muir and of Riley contain elements of both, and of rule-based behaviour (Fig 3).

These distinctions are important in that if APT requires deliberate and detailed assessment of evidence by operators before they alter their trust, it also implies that although APT is a powerful and rational model for changing trust it will suffer from the well-known problems of knowledge-based behaviour, including the need to take a considerable time to collect and assess evidence, the well-known limits of the rate of processing information, unresolvable alternatives, etc. On the other hand, if the dynamics of trust can, at least in certain situations, become so well practised that they operate at the level of skill-based behaviour, then changes in trust will occur rapidly, but perhaps unconsciously.

Such differences have strong implications for system design. In general, the more clearly the system state can be displayed to operators, and the easier designers can make the problem of state identification, the more rapidly evidence can be assessed, and the more likely it is that skill-based tuning of trust and confidence will be correct, based on pattern recognition rather than on conscious reasoning (Reason, 1990; Zsambok and Klein, 1997). The forms of the models emphasise the deficiencies in the design of some high technology automated systems from the point of view of human–system coupling, as noted by Bainbridge (1983) in her classic article, and as emphasised by Sarter and Woods (1997).

KNOWLEDGE-BASED	MUIR, 1994. QUALITATIVE MODEL RILEY, 1994, QUALITATIVE MODEL
RULE-BASED	COHEN, PARASURAMAN AND FREEMAN, 1998. "IF – THEN" RULES ZSAMBOK AND KLEIN, 1997, RECOGNITION BASED DECISIONS
SKILL-BASED	LEWANDOWSKY, 1998, NEURAL NETS LEE, 1991, REGRESSION AND TIME SERIES LEE AND MORAY, 1994, TIME SERIES MORAY, INAGAKI AND ITOH, 1999, TIME SERIES

Fig 3 The relation of different models of trust to Rasmussen's hierarchy. Different models concentrate on different levels of representation, skill, rule, or knowledge-based behaviour. Some (such as Muir, 1994) contain components at all levels, although the main bias is to the level indicated in the figure

5. Special problems

5.1 Complacency

It would seem very desirable that trust between operators and their automated systems should develop to a high level, so that operators use the automation in the manner for which it was designed and optimised. Recently however several workers have suggested that when using highly reliable automated system operators may become 'complacent' and too trusting. This has been related particularly to the problem of detecting faults in a normally highly reliable system; that is, to monitoring tasks. It is suggested that when a system is trusted, operators cease to monitor it efficiently, and hence fail to notice deviations from set points, warning lights, etc. (Parasuraman *et al*, 1993, 1996).

Muir's findings that the frequency of monitoring is inversely related to trust predict that as a system becomes trusted a supervisor will observe it less and at increasingly long intervals, supporting the position of those who worry about complacency. But in fact there is as yet no empirical evidence that operators are complacent. To claim that trust results in complacent undersampling one must first identify the optimal frequency at which to monitor the trusted system. If a system has never failed, and hence is completely trusted, is it a sign of complacency not to monitor it? The problem with the published research on complacency is that no study has identified the optimal sampling interval. There are a number of bases for defining the sampling interval, from the Nyquist theorem to more psychological models taking into account the rate at which forgetting of last observations occurs (Moray, 1986). One could in principle define a suitable sampling interval, and hence define three classes of monitoring behaviour. If operators monitored a source less often than is indicated by a model of the optimal observer, they would be said to be *complacent*. If operators oversampled, thus wasting time on unnecessary observations which could be used for other aspects of plant management, they would be described as *sceptical*. If they sampled at the optimal rate, they would be described as *eutactic*.[1] A close examination of the work of Parasuraman *et al* (1993) shows that the operators were not undersampling, although it is difficult from their paper to establish an exact optimal rate. Since no research so far has defined the eutactic sampling rate, we do not know whether, with growing trust, the monitoring rate will fall below the eutactic rate and become complacent. Moray *et al* (1999) showed that their time series equation for trust will, in the absence of any disturbances, faults or disagreements between operators and automation, asymptote to a level of trust slightly lower than the objective system reliability, which does not suggest that complacency is a problem in completely trusted highly reliable systems. There is, however, certainly a need for more work, both theoretical and empirical, on monitoring behaviour, and in particular industrial field studies of eye movements are needed.

It is important to note that even if monitoring is eutactic, not all abnormal signals will be detected. The occurrence of abnormal states, and hence abnormal signals on the interface can be assumed in the first instance to be randomly distributed in time, probably with a Poisson distribution. Monitoring requires operators to distribute their attention over many displays, either on the wall of a control room, or on different pages of computerised displays. As a conservative practical approximation we assume that they monitor only one source (page) of information at a time. Whatever sampling rate is chosen within the constraints of known facts about attention, there is a non-zero probability that a fault will appear on a channel which is not at that moment being monitored. Furthermore, if sampling of a channel is based on trust, as Muir reported, and if trust is well calibrated so that the sampling rate is (correctly) low, the probability that a fault will be rapidly detected is given by the convolution of the probability of occurrence and the probability of sampling, which means that the probability of prompt detection must be lower than the probability of occurrence, and hence there are bound to be missed or at least delayed responses to abnormal incidents, even when monitoring is eutactic.

The research issue of interest is whether in fact the decline of monitoring with growing trust asymptotically approaches the eutactic rate for a given system, or whether it does indeed become complacent. The results of Moray *et al* (1999) do not indicate complacency, but whichever is true, one cannot guarantee a prompt response to abnormal incidents on the basis of spontaneous monitoring. Alarms are essential to attract the operators' attention. Quantitative work on complacency is urgently needed, as this is a fundamental problem for supervisory control. There is always a non-zero probability of plant failure, but if no failure has ever occurred, at what interval should the plant be

[1] From the Greek *ευτακτος*, 'suitably or well trained'.

monitored? It may, for example, be possible to classify monitoring regimes into those which require almost continuous monitoring of basic process inputs which are fundamental to the process of safety critical, those which are important for the timing of processes, and those which are mainly important if the operator feels a need to update his or her information about the status of the plant, but which are not safety or productivity critical. But at present research is inadequate, except for some field studies which while interesting have not led to exact modelling (Cellier *et al*, 1996).

5.2 Allocation of authority

The research reported in this paper supports the idea that trust between human and automation plays an important role in efficient supervisory control. Empirically we find that when automation is trusted operators allow it to operate as designed, but as trust declines they are prone to intervene. Operators will not always correctly assess the trustworthiness of automation, or their own competence to take control by intervening, and hence the problem of ultimate authority appears. Which agent, human or automation, should have the final authority and responsibility for decisions in an automated system under supervisory control?

Recent papers on the human factors of automation have argued strongly that the final responsibility must reside with the human operators (Billings, 1997; Woods and Roth, 1988), and the analysis of problems with highly automated aircraft has shown that where responsibility is ambiguous, or poorly indicated in the control station, severe problems arise. Moray (1986) has described how less than perfect understanding of each other's abilities and characteristics could lead to misunderstanding between operator and machine and hence competition for control. It is interesting that one hears much less of this problem in process control, although an analysis of some nuclear power incidents suggests that there may also be problems in highly complex process plants during fault management.

There is not space here to review the problem in detail, but we claim there are some situations where there is absolutely no doubt that the final authority must rest with automation, and humans must accept the automation's decisions. A formal approach to this problem has been developed by Inagaki (1993, 1995), who provides a logical proof that conditions exist for which automation should be allowed to perform autonomous safety-control actions even when the human has given it no explicit directive to do so. One example is where time-critical events require a response so rapid that humans cannot handle the task. Another, as Wei (1997) has pointed out, is where it is expected that humans are particularly error-prone, as in the '20-minute rule' governing intervention following a nuclear power plant scram. Even when trust in automation is high and well calibrated, there will be events which call for intervention, and when the balance between trust and self-confidence swings to favour operator intervention, there are situations where it should not be allowed and automation should have absolute authority.

6. Conclusions

Research into trust between human and machines has only recently begun, but has already shown solid results of practical importance for systems design, for training, and for the design of operating procedures. This research is reaching maturity, and in the next few years should be able to play a role in the design of safer and more productive human–machine systems. While the present paper is based on experimental studies in microworlds, we hope that it will serve to stimulate interest in the important role which 'social relations' between humans and machines play in the efficient and safe use of high technology automation. Methodology is available to go far beyond traditional ergonomics in the study of human–machine relations.

7. References

Bainbridge, L. 1983. 'Ironies of automation', *Automatica*, **19**, 755–779.

Barber, B. 1983. *Logic and the Limits of Trust*, Rutgers University Press, New Brunswick, New Jersey.

Billings, C. E. 1997. *Aviation automation: the search for a human-centered approach*, Lawrence Erlbaum, Mahwah, New Jersey.

Cellier, J.-M., DeKeyser, V., and Valot, C. 1996. *La gestion du temps dans les environments dynamiques*, Presses Universitaires de France, Paris.

Cohen, M. S., Parasuraman, R., and Freeman, J. T. 1998. 'Trust in decision aids: a model and its training implications'. Technical Report, Cognitive Technologies, Inc, Arlington, Virginia.

Eidelkind, M. A., and Papantonopoulos, S. A. 1997. 'Operator trust and task delegation: strategies in semi-autonomous agent systems', In: Mouloua, M. and Koonce, J. M. (eds), *Human–Automation Interaction*, Lawrence Erlbaum, Mahwah, New Jersey.

Hiskes, D. J. 1994. 'Trust, self-confidence, and the allocation of function in discrete manufacturing systems'. Technical Report EPRL-94-02, Engineering Psychology Research Laboratory, University of Illinois at Urbana-Champaign.

Inagaki, T. 1993. 'Situation-adaptive degree of automation for system safety', *Proceedings of 2nd IEEE International Workshop on Robot and Human Communication*, pp. 231–236.

Inagakai, T. 1995. 'Situation-adaptive responsibility allocation for human-centered automation', *Transactions of SICE of Japan*, **31**(3), 292–298.

Lee, J. D. 1991. 'Trust, self-confidence and operators' adaptation to automation', Unpublished Ph.D. thesis, University of Illinois at Urbana-Champaign, Engineering Psychology Research Laboratory Technical Report EPRL-92-01.

Lee, J. D., and Moray, N. 1992. 'Trust, control strategies and allocation of function in human–machine systems', *Ergonomics*, **35**, 1243–1270.

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

McDonnell, J. D. 1969. 'An application of measurement methods to improve the quantitative nature of pilot rating scales', *IEEE Transactions on man–machine systems*, **MMS-10**(3), 81–95.

Moray, N. 1981. 'The role of attention in the detection of errors and the diagnosis of failures in man–machine systems', In Rasmussen, J. and Rouse, W. B. (eds), *Human Detection and Diagnosis of System Failures*, Plenum Press, New York.

Moray, N. 1986. 'Monitoring behavior and supervisory control', In: Boff, K., Kaufmann, L., and Beatty, J. (eds), *Handbook of Perception and Human Performance*, Wiley, New York.

Moray, N., Inagaki, T., and Itoh, M. 1999. 'Adaptive automation, trust, and self-confidence in fault management of time-critical tasks', *Journal of Experimental Psychology: Applied*, in press.

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

Muir, B. M. 1994. 'Trust in automation: part 1 – Theoretical issues in the study of trust and human intervention in automated systems', *Ergonomics*, **37**(11), 1905–1923.

Muir, B. M., and Moray, N. 1996. 'Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation'. *Ergonomics*, **39**(3), 429–461.

Parasuraman, R., Molloy, R., and Singh, I. L. 1993. 'Performance consequences of automation-induced "complacency"', *International Journal of Aviation Psychology*, **3**, 1–23.

Parasuraman, R., Mouloua, M., and Molloy, R. 1996. 'Effects of adaptive task allocation on monitoring of automated systems', *Human Factors*, **38**(4), 665–679.

Rasmussen, J. 1986. *Information Processing and Human–machine Interaction: An Approach to Cognitive Engineering*, North-Holland, Amsterdam.

Rasmussen, J., Peitersen, A.-M., and Goodstein, L. 1995. *Cognitive Engineering: Concepts and Applications*, Wiley, New York.

Reason, J. 1990. *Human Error*, Cambridge University Press, Cambridge.

Remple, J. K., Hoimes, J. G., and Zanna, M. P. 1985. 'Trust in close relationships', *Journal of Personality and Social Psychology*, **49**, 95–112.

Riley, V. 1994. 'A theory of operator reliance on automation'. In: Mouloua, M. and Parasuraman, R. (eds), *Human Performance in Automated Systems: Recent Research and Trends*, Lawrence Erlbaum, Hillsdale, New Jersey, pp. 8–14.

Sarter, N., and Woods, D. D. 1997. 'Team play with a powerful and independent agent: operational experiences and automation surprises on the Airbus A-320', *Human Factors*, **39**(4), 553–569.

Sheridan, T. B. 1976. 'Towards a general model of supervisory control'. In: Sheridan, T. B. and Johannsen, G. (eds), *Monitoring Behavior and Supervisory Control*, Plenum Press, New York, pp. 271–282.

Sheridan, T. B. 1992. *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, Cambridge, MA.

Sheridan, T. B. 1997. 'Supervisory control'. In: Salvendy, G. (ed.), *Handbook of Human Factors* (2nd Edition), Wiley, New York, pp. 1295–1327.

Tan, G., and Lewandowsky, S. 1996. 'A comparison of operator trust in humans versus machines', *INTERNET: CybErg International Electronic Conference. http://www.curtin.edu.au/conference/cyberg/centre/cognitive.cgi?state = 000100#Workload*

Wei, Z.-G. 1997. *Mental Load and Performance at Different Automation Levels*, Delft University of Technology, Delft, Netherlands.

Woods, D. D., and Roth, E. M. 1988. 'Cognitive systems engineering', In: Helander, M. (ed.), *Handbook of Human–Computer Interaction*, North-Holland Elsevier, Amsterdam.

Zsambok, C. E., and Klein, G. 1997. *Naturalistic Decision Making*, Lawrence Erlbaum Associates, Marwah, New Jersey.

Zuboff, S. 1988. *In the Age of the Smart Machine*, Basic Books, New York.