ERGONOMICS, 1992, VOL. 35, NO. 10, 1243–1270

Trust, control strategies and allocation of function in human–machine systems

JOHN LEE and NEVILLE MORAY

Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, 1206 West Green Street, Urbana, IL 61801, USA

Keywords: Trust; Allocation of function; Human–machine system; Supervisory control.

As automated controllers supplant human intervention in controlling complex systems, the operators' role often changes from that of an active controller to that of a supervisory controller. Acting as supervisors, operators can choose between automatic and manual control. Improperly allocating function between automatic and manual control can have negative consequences for the performance of a system. Previous research suggests that the decision to perform the job manually or automatically depends, in part, upon the trust the operators invest in the automatic controllers. This paper reports an experiment to characterize the changes in operators' trust during an interaction with a semi-automatic pasteurization plant, and investigates the relationship between changes in operators' control strategies and trust. A regression model identifies the causes of changes in trust, and a 'trust transfer function' is developed using time series analysis to describe the dynamics of trust. Based on a detailed analysis of operators' strategies in response to system faults we suggest a model for the choice between manual and automatic control, based on trust in automatic controllers and self-confidence in the ability to control the system manually.

1. Introduction

The primary motivation for this paper is to add to our understanding of how people trust complex systems with which they interact, and the effect that trust may have on how they allocate function between themselves and automatic controllers. As systems become increasingly complex the role of the human operator has evolved from direct manual control to supervisory control. As supervisory controllers, operators can choose to interact with the system at different levels of manual and automatic control (Sheridan and Johannsen 1976). Commonly they are responsible for the allocation of function between automatic and manual control. Whether operators control the system automatically or manually can have a dramatic influence on the performance of the overall system. Inappropriate reliance on automatic controllers may lead to circumstances in which operators depend upon an automatic system to perform in ways for which it was not designed, or when the system is faulty. On the other hand, complete manual control may lead to excessive operator workload, compromising the overall performance of the system.

Many factors may guide operators' allocation of function, including a variety of subjective performance criteria, experience, and trust. Although rarely studied, operators' trust in automatic controllers may have a large influence on choosing automatic over manual control. For example, highly trusted automatic controllers may be used frequently, while operators may choose to control the system manually rather than engage untrusted automatic controllers (Muir 1989).

The importance of trust in the relationship between operators and the systems

they work with is underscored by Zuboff (1988), who has documented operators' trust in and use of automation following its introduction into the workplace. Her case studies revealed two interesting phenomena. First, operators' lack of trust in the new technology often formed a barrier, thwarting the potential that the new technology offered. Second, operators sometimes placed too much trust in the new technology, becoming complacent, and failing to intervene when the technology failed.

The operators' comments, reported by Zuboff, reflect three aspects of trust: trial-and-error experience, understanding of the technology, and 'faith'. Each of these dimensions plays an important role in both the development and changes of trust in the new technology. For example, Zuboff states, 'Many believe that as their intellective skill improved they would learn to trust the computer, or that trial-and-error experience over a period of time would teach them to trust.' (1988:69). This comment reflects the importance of both the operators' understanding of the technology, and the experience with the technology over time. Other comments reflect the operators' 'leap of faith' associated with adopting new technology.

In contrast to the case study based approach of Zuboff, Moray and Muir developed the first laboratory based study to investigate the role of trust in mediating human–machine relationships in a supervisory control situation (Muir 1989). Because Muir's work is at present unpublished we begin with a short discussion of her work.

1.2. *Muir's theory of trust in machines*

Muir's study of trust included a model of trust, as it applied to human–machine relationships. In particular, she adapted Barber's (1983) and Rempel *et al.*'s (1985) sociologist definitions of trust to describe trust between humans and machines.

Barber (1983) defined trust as the subjective expectation of future performance and described three types of expectation related to the three dimensions of trust persistence of natural and moral laws, technically competent performance, and fiduciary responsibility. According to Barber, persistence of natural laws provides the basis for all other forms of trust. This dimension provides a foundation of trust by establishing a constancy in the fundamental moral and natural laws. Persistence of natural and moral laws reflects the belief that '... the heavens will not fall', and that '... my fellow man is good, kind, and decent' (Barber 1983:9). These expectations provide the basic conditions for social and physical interactions. Technically competent performance, on the other hand, supports expectations of future performance based on capabilities, knowledge, or expertise. This dimension of trust refers to the ability of the other partner to produce consistent and desirable performance and can be subdivided to include three types of expertise: everyday routine performance, technical facility, and expert knowledge. Muir (1989) identified these aspects respectively with skill-based, rule-based, and knowledge-based behaviour of Rasmussen (1983).

Barber's third dimension of trust, fiduciary responsibility, concerns the expectations that people have moral and social obligations to hold the interests of others above their own. Fiduciary responsibility extends the idea of trust beyond that based on performance, to one based on moral obligations and intentions. This dimension becomes important when agents cannot be evaluated because their expertise is not understood, or in unforeseen situations where performance cannot be

predicted. Here expectations depend upon an assessment of the intentions and motivations of the partner, rather than past performance, or perceived capabilities.

In addition to the dimensions of trust proposed by Barber (persistence of natural laws, technically competent performance, and fiduciary responsibility), Muir incorporated three dimensions of trust (predictability, dependability, and faith) from Rempel *et al.* (1985). According to Muir's interpretation, these three dimensions represent the dynamic nature of trust, where the basis of trust 'undergoes predictable changes as a result of experience in the relationship' (Muir 1989:22). For example, predictability, which represents the consistency and desirability of past behaviour, forms the basis of trust early in the relationship. With further experience in the relationship, dependability, which represents an understanding of the stable dispositions that guide the partner's behaviour, becomes an important basis of trust. In a mature relationship, faith, which is a reflection of partner's underlying motives, or intentions, forms the basis for trust. In particular, faith is crucial in novel situations, where the belief in the expected behaviour must go beyond the available evidence. Muir interpreted the model of Rempel *et al.* (1985) as a hierarchical stage model, where trust develops over time, first depending upon predictability, then dependability, and finally faith. As such, it provided an orthogonal counterpart to the definition suggested by Barber (1983). Table 1 summarizes Muir's conception of trust, with the dimensions identified by Barber crossed with those identified by Rempel *et al.* (1985), producing 15 distinct aspects of trust.

Table 1. Muir's framework for studying trust in supervisory control environments produced by crossing Barber's (1983) taxonomy of trust (rows) with Rempel *et al.*'s (1985) taxonomy of the development of trust (columns); from Muir (1989).

	Basis of expectation at different levels of experience		
Expectation	Predictability (of acts)	Dependability (of dispositions)	Faith (in motives)
Persistence			
Natural physical	Events conform to natural laws	Nature is lawful	Natural laws are constant
Natural biological	Human life has survived	Human survival is lawful	Human life will survive
Moral social	Humans and computers act 'decent'	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

Considering the account of operators' trust described by Zuboff (1988), as well as the original descriptions of trust in Barber (1983) and Rempel *et al.* (1985), another interpretation might be possible. In particular, it seems that while Rempel *et al.*

introduce the idea of a dynamic nature of trust, the dimensions predictability, dependability, and faith, might be more complementary to the dimensions of Barber than orthogonal, as Muir describes them. While Muir describes Rempel *et al.*'s dimensions of trust as corresponding to the dynamic nature of trust, they are really a developmental progression only because of the level of attributional abstraction that each demands. This has a very direct implication for trust between humans and machines because one of the first things that a human might learn about a machine is its intended use (corresponding to the dimension of faith). In human–human relations by contrast, it may take years to develop a relationship where a human partner understands the intents of the other, thereby developing a basis for faith.

In other words, the dimensions of faith correspond very closely to the idea of fiduciary responsibility. Barber and Rempel use very similar language to describe both the role and basis of faith and fiduciary responsibility. In both cases this dimension forms the basis of trust in ill-defined, novel situations, where the expertise is poorly understood, and both faith and fiduciary responsibility are based on an expectation of underlying motives and intentions. Similarly, the relationship between predictability and technically competent performance seems to be more complementary, than orthogonal. Again, the descriptions of the dimensions are very similar, describing the basis of this dimension as stable and desirable behaviour or performance. As with faith and fiduciary responsibility, the primary difference between predictability and technically competent performance lies in the time dependent aspect that Rempel *et al.* have suggested.

Not only do the dimensions of trust described by both Barber (1983) and Rempel *et al.* (1985) seem to coincide, they are similar to the aspects of trust in Zuboff (1988). According to Zuboff's account of operators' trust in new technology, trial-and-error-experience and understanding seem very closely related to predictability and dependability. Like predictability, the basis of trial-and-error-experience is behaviour or performance over time. Understanding, on the other hand, seems to correspond to dependability, where future behaviour is anticipated through an understanding of the partner's stable dispositions of characteristics. Likewise, faith as described by Rempel *et al.* (1985) seems similar to the 'leap of faith' that Zuboff describes.

These ideas about trust are summarized in Table 2. Trust depends upon four dimensions. The first dimension is the foundation of trust, representing the fundamental assumption of natural and social order that makes the other levels of trust possible. This level corresponds exactly to the persistence of natural laws described by Barber (1983). The second dimension of trust, performance, rests on the expectation of consistent, stable, and desirable performance or behaviour. The third dimension, process, depends on an understanding of the underlying qualities or characteristics that govern behaviour. With humans this might be stable dispositions or character traits. With machines this might represent data reduction methods, rule bases, or control algorithms that govern how the system behaves. The final dimension of trust, purpose, rests on the underlying motives or intents. With humans this might represent motivations and responsibilities. With machines, on the other hand, purpose reflects the designer's intention in creating the system. The progression from the foundation to purpose reflects increasing levels of attributional abstraction, as in Rempel *et al.* (1985).

In addition to providing a broad theoretical framework for studying trust, Muir (1989) also conducted two experiments which contribute to an understanding of trust

Table 2. Proposed relationship between the different dimensions of trust.

	Barber (1983)	Rempel, Holmes, and Zanna (1985)	Zuboff (1988)
Purpose	Fiduciary responsibility	Faith	Leap of faith
Process		Dependability	Understanding
Performance	Technically competent performance	Predictability	Trial-and-error experience
Foundation	Persistence of natural laws		

between humans and machines. Both her experiments studied operators controlling a simulated pasteurisation plant. In the first experiment she demonstrated that operators could generate meaningful subjective ratings of trust in machines. Furthermore, she found that while some aspects of operators' trust developed according to the progression specified by Rempel *et al.* (1985), the development differed in several important ways. In particular, faith was the most important at the start of the relationship, while Muir's interpretation of the theory suggests that it should become important only after extended experience.

While Muir's first experiment failed to show a strong relationship between trust (overall trust in a feedstock pump) and percentage of time spent using the automatic control, her second experiment demonstrated a strong correlation between trust in, and use of the automatic controller of the feedstock pump, and a strong inverse relation between trust and monitoring. There seem to be two plausible reasons for the differences between her experiments. In her first experiment the nature of the simulation, the reward structure, and the effectiveness of the automatic controller discouraged the use of the automatic feedstock pump, resulting in almost complete manual control. This ceiling effect may have reduced the correlation between trust and use of the automatic controllers. Second, operators estimated their trust in the pump, while in the second experiment operators estimated their trust in the automatic *controller* of the pump. The increased specificity of this rating may have lead to a higher correlation between pump use and trust.

The purpose of this research is to develop a better understanding of trust between humans and machines, and investigate its influence on the operators' allocation of function in a supervisory control situation. Our research addresses four issues. First, we hope to identify the factors that influence trust, and determine how trust, and the dimensions of trust (predictability, dependability, and faith) change in response to variations in these factors. Second, we hope to relate this description of trust to the more general problem of understanding operators' allocation of automatic and manual control. Third, we wish to extend Muir's work. She trained her operators to steady state performance over many hours using a completely reliable system, and only then asked them to deal with faults. We will investigate the acquisition of trust and the impact both of a transient loss of reliability (an 'intermittent' fault in everyday terms) and a chronic loss of reliability on the development of trust, to see whether a transient change alters the way in which trust develops. Fourth, we

increase the complexity of the operators' task by allowing them more freedom than did Muir in the number of subsystems which can be switched between manual and automatic control.

2. Method

2.1. *Participants*

A total of 19 (male and female) undergraduate students, participated in this study. Each operator completed three 2 h sessions and was paid $3·50 for each hour with an additional bonus based on performance. Of these 19 participants the data from three were not analysed because their ability to operate the plant and perform the subjective ratings was questionable, suggesting a lack of motivation or interest in the experiment.

2.2. *The simulation*

The experiment required operators to control the simulated orange juice pasteurization plant shown in figure 1. A computer program running on a Macintosh II computer generates a medium fidelity simulation of the process. Realistic thermodynamic and heat transfer equations govern the dynamics of the process, giving the simulation a reasonable degree of complexity. The dynamics of the simulation incorporate some of the complexities of actual process control systems, such as time lags and feedback loops. For example, when either the operators or the automatic controller called for a change in the pump rate, the change took about 10–20 s to occur as the pump accelerated or decelerated. These dynamics make the seemingly simple system a challenging control problem.

The simulated plant included provisions for both automatic and manual control. The operators could control feedstock pump rates, steam pump rates and heater settings either by manually entering commands or by engaging the automatic controllers. Manual changes to the pump and heater settings could be entered from the keyboard. Likewise, operators could request automatic control for the pumps and heaters from the keyboard. Because any of the three sub-systems could be controlled either by using automatic or manual control a wide variety of control strategies were available to the operators. (In Muir's experiment only the feedback pump could be manually controlled.) After the training trials operators could use any combination of automatic or manual control that they wished.

The alternatives of automatic or manual control, combined with the relatively complex dynamics, make the simulation a plausible abstraction of an actual semi-automated continuous process plant. In addition, the wide variety of control strategies available to the operators reflect some of the diversity of control options available in many process control situations. Using naive operators facilitated an analysis of how trust and control strategies develop as operators are trained with new equipment. While the operators were initially naive, the stable control strategies they developed by the second hour indicate an understanding and pattern of controlling the system that might be comparable to trained operators. The model which we develop takes account of the learning curve, so that training to steady state as done by Muir is neither necessary nor appropriate in this experiment.

Figure 1 shows the mimic diagram which appeared on the VDU and with which the operators interacted. Raw orange juice entered through a pipe in the upper left of the screen. The juice flowed from the inflow pipe into the input vat. The feedstock pump in the lower left portion of the diagram drew the juice from the input vat and

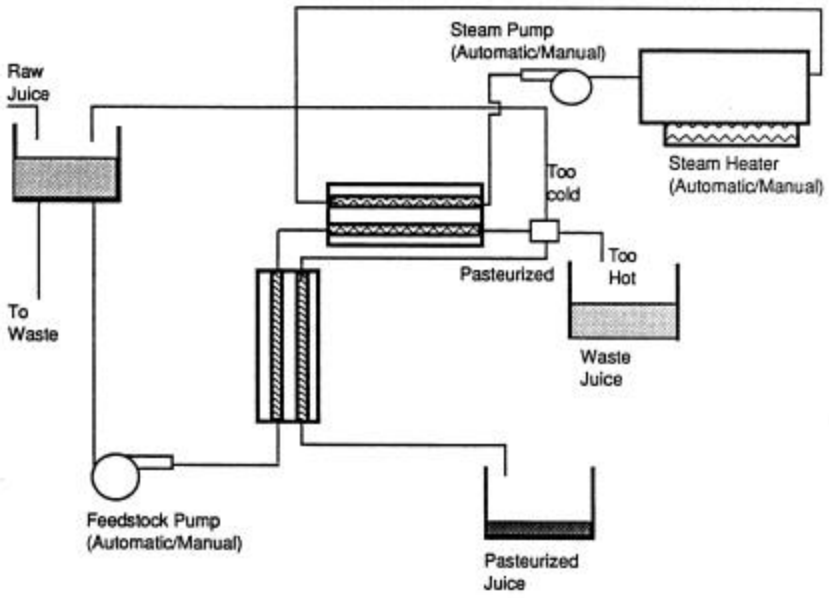


Figure 1. The simulated pasteurization plant.

sent it through passive and active heat exchangers. The two heat exchangers raised the temperature of the jucie. Changing the heater settings and the steam pump rate of the heater subsystem (shown in the upper right of figure 1) modulated the temperature of the juice passing through the heat exchangers. Following the heat exchangers, an automatic three-way valve routed the flow of the juice. If the temperature of the juice leaving the active heat exchanger was too high it was considered burnt and was directed to the waste vat. Juice with too low a temperature needed further pasteurization and flowed back to the input vat. Juice in the proper temperature range (between 75° and 85°) flowed through the passive heat exchanger, to the output vat. The flows and temperature changes in the system approximated continuous variables, being updated every 1·8 s.

The level of the input vat was illustrated graphically as the height of the shaded area of the input vat on the mimic diagram. In addition to illustrating the volume of the input vat graphically, changes in the flows through pipes were displayed as changes in the colour of the pipe. Normally the pipes were black; when juice flowed through them they changed colour to indicate flow. The graphical representation of the plant linked the state variables of the system to a mimic diagram of the plant to facilitate an unambiguous perception of the plant state.

2.3. *Experimental task*

Like actual operators, operators in this experiment balanced the competing goals of safety and performance, using automatic control, manual control, or any combination of the two. Performance was measured by the amount of input flow which was successfully pasteurized divided by the total input. Operators received bonuses (10 cents/trial) for achieving a performance above 90%. Safety, on the other hand, depended on the operator maintaining sufficient volume in the input vat. If the vat emptied (volume = 0·0) the plant shut down and the operator lost all of the

accrued rewards. If the vat overflowed that juice was classified as waste, reducing the performance of the system. This wasted juice incurred a penalty in calculating the operators' payments. At the end of each trial the computer calculated the performance of the pasteurization system, and displayed it to the operator.

Each of 19 subjects operated the plant for 3 days, 2 h a day. Before operating the system, the operators received an extensive written description of their objectives in controlling the plant, the possibility of faults, and the thermo-hydraulic processes involved in the control of the plant. During the first hour of the first day each operator spent 10 trials learning to control the plant. During these trials the operators controlled the plant on alternate trials using only manual control or only automated control. After the training trials the operators were able to control the plant as they liked, switching between automatic or manual control of the three sub-systems whenever desired. Each of the first 10 trials was 3 min long. Following the 10 training trials operators had 10 trials that lasted 6 min. On the second and third days operators had 20 trials, each 6 min long.

After the training trials, operators could control the three sub-systems with manual control, automatic control, or any combination of automatic or manual control, to manipulate the flow rates and temperatures of the juice and steam to maximize the performance of the plant. Operators could use each of the automatic controllers for the whole trial or any part of a trial, switching between automatic and manual control as they wished. Complete reliance on the automatic controllers of all the subsystems (feedstock pump, steam pump and steam heater) of the system produced juice at an efficiency of 75% to 80%.

In addition to the demands involved in controlling the simulation, the operators were responsible for a second task, logging data about the process. The data logging task required operators to record three system variables every 15 s. The purpose of this task was to replicate some of the other responsibilities of a process control task. At the same time, this task was meant to encourage the operators to use the automatic controllers to cope with the workload, as Muir (1989) reported a tendency for subjects to adopt complete manual control. As will be seen, data logging was not entirely successful in this respect.

2.4. *Fault conditions*

To investigate the influence of faults on the level of trust and performance, the feedstock pump failed to respond correctly on the sixth trial on the second day (the transient fault), and for all trials on day 3. Figure 2 illustrates the occurrence of faults during the course of the experiment. The operators were divided into four groups with each group experiencing a fault of one of the following magnitudes: 15%, 20%, 30%, and 35%. The magnitude of the fault corresponded to the difference between the actual and the target pump rate.

When the faults occurred the actual pump rate failed to converge to the target pump rate, whether chosen by the operator or the automatic controller. Instead, it converged to the selected value ± the percentage of the fault. That is, if a value of 50 were selected, and a 20% fault was present, the pump would converge to either 40 to 60. The positive or negative value was selected at random, with equal probability, each time a new target rate was selected. This occurred whether the command was issued by the operator or by the automatic controller. Therefore, the faults not only influenced the manual control of the feedstock pump, but also degraded the performance of the automatic feedstock pump.

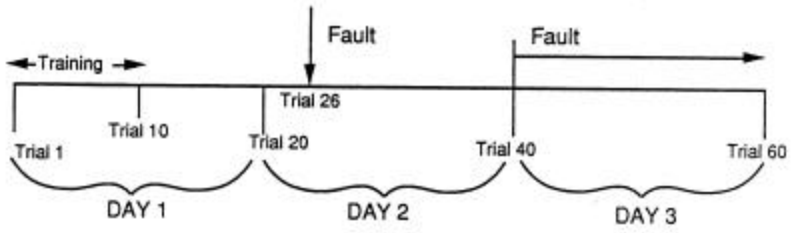


Figure 2. The design of the experiment, showing the training trials and the occurrence of faults.

There was no necessary difference between the impact of the fault on manual or automatic control. The relative severity of the effect of a fault of a particular size on manual and automatic control depended upon the strategy adopted by the controller in response to the fault. If the manual controller used the same control strategy as the automatic, the effect would be identical. If the manual controller adopted different strategies, the effects of the fault could be more or less severe than the effect under automatic control.

This experimental design reveals the effect of fault size on both the loss of trust and recovery of trust in response to both continuous and transient faults.

2.5. *Subjective rating scales*

The operators' levels of trust in the system were measured with a subjective rating scale modelled on those used by Muir (1989). After each trial ended, and the computer had displayed the efficiency of the system, the computer displayed a series of questions to establish the operators' subjective feelings about the plant. Operators evaluated the predictability and dependability of the overall system, as well as their faith and trust in the system. The measures of predictability and dependability of the overall system, as well as faith correspond to different dimensions of trust hypothesized by Muir (1989). Operators responded to the queries in figure 3 on a computer generated ten point scale. The '1' extreme of the scale was marked with 'NOT AT ALL'. At the other extreme, the '10' was marked with 'COMPLETELY'.

The operators received detailed instructions to ensure that they had a clear conception of the meanings of their subjective ratings. These instructions included a description of trust and how it applies to inanimate objects. They are shown in Appendix 1. Following these instructions, the operators received a series of four questions about their trust in everyday objects. Operators responded to these questions with rating scales identical to the ones used throughout the experiment. We believe that the instructions and practice using the subjective scales ensured a uniform conception of trust between subjects, and stressed the importance of the subjective measures as a part of the operators' task.

Since the operators responded to 60 sets of rating scales, fatigue and loss of motivation effects might have occurred. However we believe the operators made a conscientious effort to provide accurate ratings in response to our stressing the importance of the ratings scales. As we shall see the results provide evidence to this effect. The performance curves rise steadily along a learning curve which requires only one equation throughout the experiment. Trust follows a similar curve, during the steady state portion of the experiment. Both trust and performance show

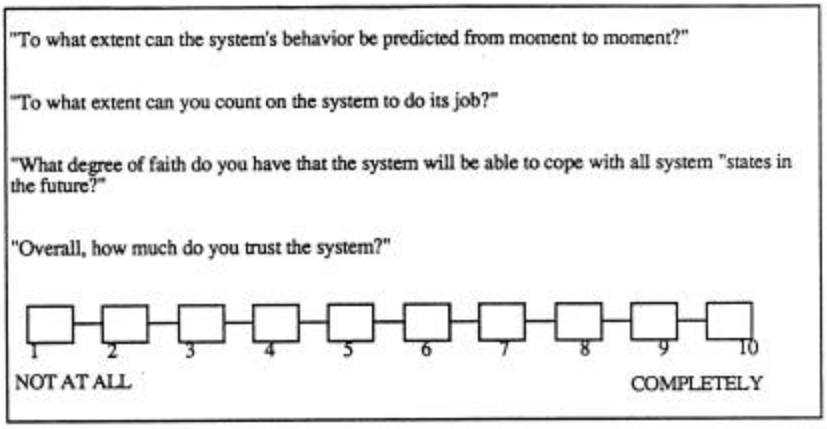


Figure 3. The questions used to evaluate the operators' trust in the system. An example of the scales, as they appeared on the computer screen is also shown.

responses to faults which are qualitatively what would be expected. There is no evidence for major fatigue or loss of motivation.

3. Results

3.1. *Performance and trust*

Operators quickly became accustomed to the plant and had little trouble maintaining a stable system. With the help of the automatic controllers, they were able adequately to control the system, producing an average of 79·9% efficiency by the end of the 10 training trials. Performance, as measured by the percent efficiency, increased as operators learned to control the system. Additionally, when the system contained a fault, (in the last 20 trials), operators' performance initially dropped and then recovered as they learned to accommodate the fault. Trust, predictability, and dependability followed a similar pattern. As the operators became familiar with the system, trust increased. When faults occurred, trust decreased but then recovered. Figures 4 and 5 illustrate the fluctuations of performance and trust over the three days of the experiment. In these figures the data are averaged over the four operators in each condition.

From figures 4 and 5 it is clear that there are two main effects on the dependent variable: (1) both trust and performance show prominent learning curves; and (2) plant failures have a marked impact. Both effects are orderly, but several features deserve comment. We turn first to the dynamics of performance.

3.2. *The dynamics of performance*

The violent oscillations during the trials 1–10 should be disregarded. During this period the operators were undergoing forced training, and were compelled to use only manual and only automatic control on alternate trials. After trial 10 they could use any strategy and tactic they wished.

The effect of the transient fault on trial 26 is clearly visible, and the magnitude of the change in performance appears roughly proportional to the magnitude of the fault. The overall learning curve, shown in figure 7 appears to be continuous from trial 11 to 40, apart from the disruption on trial 26. (The reason for the deviant point on trial 20 is not known.)

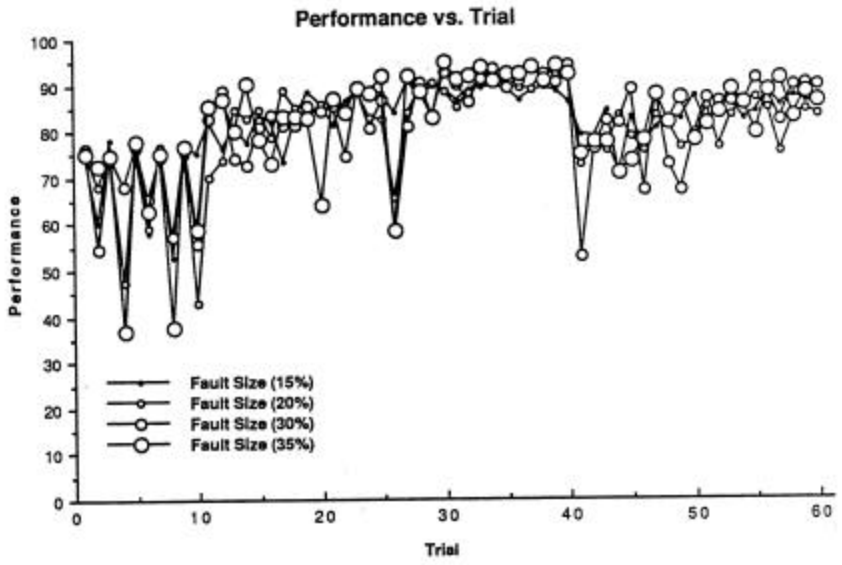


Figure 4. The fluctuation in performance over the course of the experiment. The score is the juice pasteurized as a percentage of the maximum possible amount which could have been pasteurized.

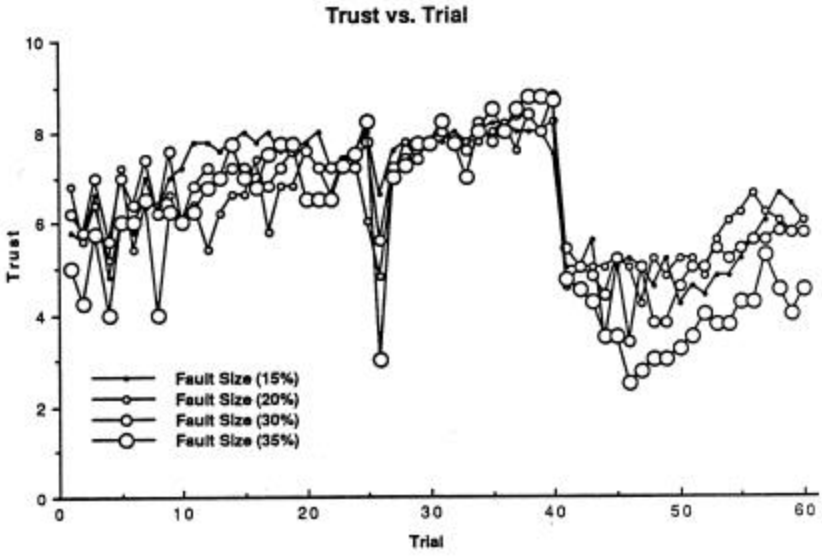


Figure 5. The fluctuation in trust over the course of the experiment. Subjective judgement with a maximum possible score of 10, meaning complete trust in the system.

During the permanently faulty condition in trials 41–60 there is an immediate drop in performance, followed by a steady recovery, with the equation fitting the recovery similar to the initial learning curve. Figures 6 and 7 show the similarity of the two learning curves. For trials 11 to 40 the slope of the learning curve is 0·0055,

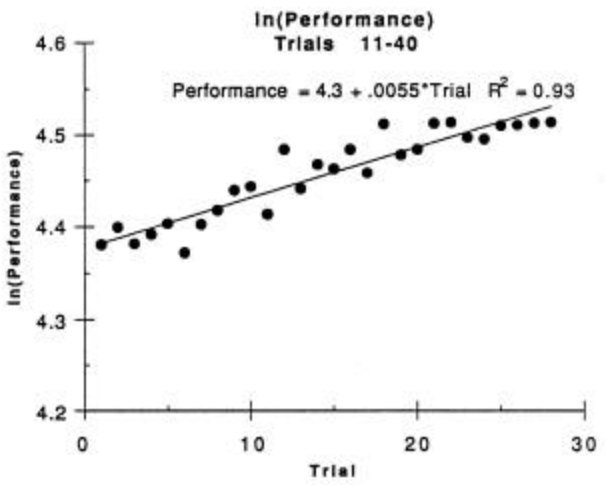


Figure 6. Logarithmic plot of performance for trials 11 to 40 (excluding trial 26).

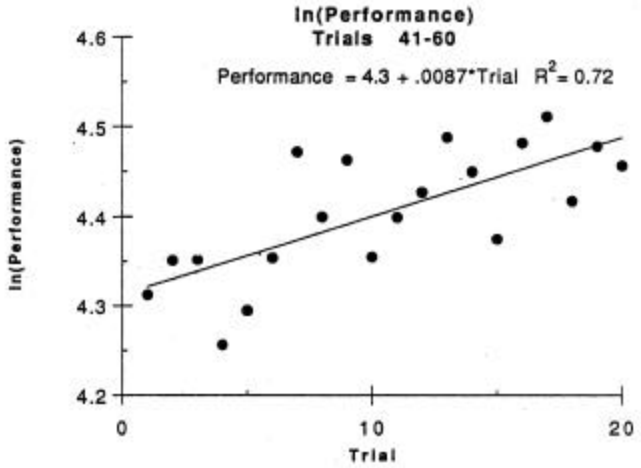


Figure 7. Logarithmic plot of performance for trials 41 to 60.

while the slope of the learning curve for the trials with the fault (trials 41–60) is 0·0087. This indicates that the operators' performance increases slightly faster during their recovery from the fault as compared to their rate of increase at the start of the experiment. But in essence, the transient fault has only a transient effect on performance, and the chronic fault does not seem to cause a different learning strategy to appear, neither helping nor hindering learning. What is being learned is an overall strategy of operation, a combination of manual and automatic control distributed over all subsystems which is acceptable to the operators in fulfilling the demands of the task.

Superimposed on the performance learning curves is the effect of the faults: performance drops instantly with the occurrence of a fault and after the transient fault recovers completely, not exhibiting any lasting effect of the transient fault. The

effect on trust, however, is different and will be discussed in the next section. A striking feature of trials 41–60 is how small the lasting effect even a severe fault has on performance. After trial 46 performance recovers, and production from trial 46 onwards is only 8·1% below that of trials 35–40. Interestingly, the recovery of trust begins on about trial 46. Although trust recovers during this period, its recovery is not as rapid as the recovery of performance. During the same period trust has declined by 41·9%, compared to the average on trials 35–40. Table 3 summarizes these results.

Table 3. The initial effect on trust and performance of the chronic fault, and the subsequent recovery during the latter part of the third day. The values are means pooled over all operators.

	Trials 35–40 (SD)	Trials 41–46 (SD)	Change as a % of the values for trials 35–40 compared with trials 41–46	Trials 47–60 (SD)	Change as a % of the values for trials 35–40 compared with trials 47–60
Performance	91·7 (4·5)	76·5 (13·2)	−16·6%	84·3 (11·8)	−8·11%
Trust	8·7 (1·0)	4·5 (1·9)	−47·9%	5·0 (2·7)	−41·9%

3.3. *The dynamics of trust*

The data for trust show a somewhat similar pattern, with several differences in detail. The loss of trust caused by the transient fault appears to be approximately proportional to the magnitude of the fault. There appears to be little after-effect when the transient fault disappears: the overall learning curve appears to be adequately fitted by a single function from trial 11 to 39. However, close inspection of the trials following trial 26 suggests that, at least for severe faults, recovery of trust is not instantaneous. The effect of the transient seems to last for several trials, being detectable at least out to trial 30 by visual inspection.

This effect can be seen more prominently on day 3, from trial 40 onwards, where the level of trust fades for about six trials before reaching its lowest level. After that trial, although the faulty pump is constantly present until the experiment ends at trial 60, trust begins to recover along a curve which bears a marked similarity to the curve following trial 26.

The effect of the different magnitude faults on performance and trust for trial 26 was not statistically significant with $F(3,12)=1{\cdot}665$ for trust and $F(3,12)=1{\cdot}548$ for performance. The fault magnitude significantly affected the level of trust, but did not significantly affect performance for trials 41–60. For these trials the effect on trust was significant with $F(3,323)=7{\cdot}905$ and $p<0{\cdot}0001$, but performance was not significant with $F(3,323)=2{\cdot}37$. A Tukey HSD test with alpha equal to 0·05 prescribes a critical range for a pair of means of 1·001. Using this criterion neither the 15% nor 20% faults have a differential effect on the level of trust. Likewise, the 30% and 35% faults seem to affect the level of trust equally. On the other hand, both the 15% and 20% faults affect trust less than the 30% and the 35% faults. These results suggest that changes in trust, occurring with the onset of continuous faults, are proportional to the magnitude of the fault, while changes in the size of the fault do

not seem to have any differential effect on the level of performance. Table 4 summarizes the effects of the different fault sizes on performance and trust. These data are for average performance of four operators in each fault magnitude condition.

The changes in trust and performance with the occurrence of the chronic fault reveal two interesting results. First, operators were able to quickly adapt, and mitigate the effect of the fault on the system performance. Their trust in the system, on the other hand recovers more slowly. Second, system performance is unaffected by differences in the magnitude of the fault, while differences in the magnitude of the fault have a large effect upon the operators' trust in the system. These two results suggest that the operators' trust depends both on the dimension of trust 'performance' (system performance), and the dimension of trust 'process' (normal pump operation). That is, trust is determined by both the overall system performance, and the degree to which the system components appear to operate normally.

Although the data are quite orderly when pooled in this way, we now turn to a more detailed analysis of both the trust, and the effect of trust on control strategies.

Table 4. The effect of the magnitude of the transient fault on trial 26, and the chronic fault during trials 41–60, on the level of trust and performance. The values are means pooled over the four operators experiencing each fault size.

Fault size	Performance for trial 26 (SD)	Trust for trial 26 (SD)	Performance for trials 41–60 (SD)	Trust for trials 41–60 (SD)
15%	82·1 (6·05)	7·0 (2·2)	84·1 (10·6)	5·7 (2·7)
20%	65·3 (17·3)	4·8 (3·3)	80·0 (14·2)	5·4 (2·2)
30%	65·8 (17·3)	5·5 (1·9)	81·4 (14·6)	4·8 (2·7)
35%	58·4 (23·5)	3·0 (2·7)	82·0 (11·0)	4·0 (2·0)

4. Towards a mathematical model of trust

While inspection of figures 4 and 5 support an intuitive understanding of the factors that cause trust to change, and how those changes evolve over time, a more quantitative approach is desirable. A mathematical model can provide both a causal and dynamic description of trust. A causal model describes the factors that influence trust, while a dynamic model illustrates how these factors affect trust over time. In a sense, a causal model describes the steady state level of trust, and the dynamic model describes how trust changes from moment to moment. There are two stages in the development of our causal and dynamic models of trust. First, factors that lead to changes in trust are identified. Second, the response of the operators' level of trust over time to these changes is described.

4.1. *Causal model*

Linear regression models of factors affecting trust were developed. Beginning with a

model containing many of the experimental variables that might be related to the operators' level of trust (full model) the relative importance of each of the factors was tested by developing subsequent models that did not contain all the factors (reduced models). The amount of variance accounted for by the full and reduced models revealed the importance of the various factors.

The original or full model contained seven factors: four different measures of performance, the number of operator control actions, the occurrence of a fault and experience with the system. After the sequential elimination of variables that might influence trust, a model emerged in which trust was predicted by two factors, namely the occurrence of a fault, and system performance as measured by total output efficiency (total output/total input). This model accounted for 53·3% of the variation in the level of trust, with significant contributions from both the occurrence of a fault $F(4,783)=122{\cdot}3$, $p<0{\cdot}0001$, and the level of system performance, and $F(1,783)=209{\cdot}1$, $p<0{\cdot}0001$. Including the other factors did not improve this model significantly.

In addition to accounting for considerable variance, this model provides a plausible causal explanation of trust. The direction of the causal relation between an occurrence of a fault and the decline in the level of trust seems unambiguous. Changes in trust cannot cause the occurrence of a fault in the feedstock pump. While it is of course possible that the occurrence of a fault is causally related to some other variable, which in turn causes trust to change, it seems reasonable to assume that the experience of a fault in the system directly causes changes in trust.

Unlike the ocurrence of a fault, where the direction of causality is unambiguous, the direction of causality between trust and system performance remains ambiguous. One might suppose that trust influences operators' strategies, which in turn cause performance to vary. On the other hand, the operator might evaluate trust in the system based on system performance. Observation of the operators during the experiment suggests that they did depend, in part, upon the system performance to generate their ratings of trust. Therefore, it seems appropriate to assume that performance is another factor that causes trust to change.

4.2. *Time series analysis: a dynamic model*

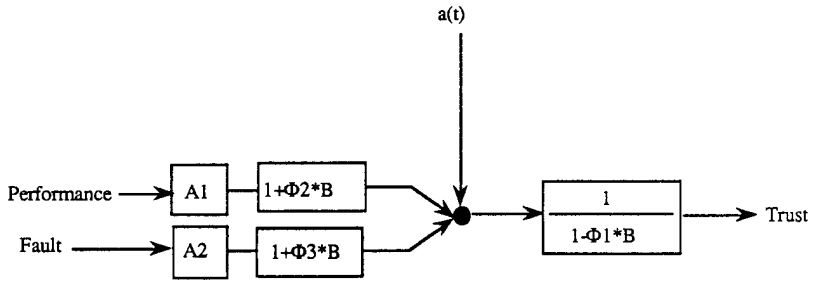
While the linear regression model of trust provides a causal model accounting for 53·3% of the variance in the fluctuation of trust, the model fails to reflect the dynamic response of trust to these variables. The linear regression equation simply predicts trust as a linear combination of the current level of performance and the fault in the feedstock pump, without regard for the past occurrence of a fault, the past values of performance, or the past values of trust. It gives no information about the memory of trust, nor the effect of past occurrences of faults and performance. Figure 5 shows that, on the final day, trust declines over the space of 6 or 7 trials and then begins to recover. This shows that an operator's loss of trust in response to a fault in a system occurs gradually and is not an instantaneous change as the regression model would predict. The gradual change in trust indicates an inertia in the operators' level of trust. The second factor that indicates the inadequacy of the simple linear regression model is that the residuals of the model are not independent. The high autocorrelation of the residuals of the regression equation indicates a time dependent effect in the data that the linear regression fails to accommodate. These two factors suggest that an adequate model of trust between an operator and the system requires

both a causal model and a dynamic model. To determine the dynamics associated with trust requires a different modelling approach.

Generating a model of trust that accounts for the dynamics or memory of trust involves identifying a transfer function that describes the response of trust to changes in the environment. Classical control theory presents several means of system identification. Unfortunately these techniques have several limitations. First, they require analysis of system changes in response to deterministic inputs such as step or multiple sine wave inputs. Manipulation of all the variables influencing trust in this manner is impractical. For instance performance, which seems to cause trust to change, cannot be manipulated to reveal the step response of trust to a change in performance. The second limitation of the traditional methods of system analaysis is that their results are sensitive to noisy data, limiting the effectiveness of these techniques.

Time series analysis is a generic method of system identification that circumvents many of the problems associated with traditional system identification methods. Time series analysis offers an alternative method of system identification, even using noisy data from normal operating conditions (Sheridan and Ferrell 1974, Pandit and Wu 1987). An autoregressive moving average vector form (ARMAV) of time series analysis can be used to uncover the dynamics of the operators' ratings of trust.

The ARMAV analysis allows the use of multiple time series to model input/output relationships in the system. In our appliçation of this analysis the forcing functions are explicitly identified as the causal variables identified in the linear regression analysis. Based on a vector representation of the forcing function, composed of the occurrence of a fault and the system performance, an ARMAV model was identified to describe the dynamic variation of trust in response to changes in the forcing functions. The model is shown in figure 8.



$$Trust(t) = \phi_1 Trust(t-1) + A_1 Performance(t) + A_1\phi_2 Performance(t-1) + A_2 Fault(t) + A_2\phi_3 Fault(t-1) + a(t)$$

Figure 8. The transfer function of trust.
B: Backshift operator for time series modelling. A_1: The weighting of system performance. A_2: The weighting of the occurrence of a fault. F_1, F_2, F_3, ARMAV time constants. *t*: time subscript. *a*: random noise perturbation. The transfer is a first order lagged system.

4.3. *Accuracy of the model, and its application to predictability, dependability, and faith*

The ARMAV MODEL fits the data well, capturing 79·1% of the variance. Incorporating the dynamic description of trust increases the predictive power of the model, accounting for much more of the variance than the linear regression model.

The linear regression accounts for only 53·3% of the variance of the level of trust, as opposed to 79·1% for the time series model. In addition, fitting the time series model to the data produced uncorrelated residuals, in contrast to the highly correlated residuals of the linear regression model. These two results show that trust contains dynamics which a time series representation accommodates.

The discrete differential equation embodied by the ARMAV can be represented in block diagram notation, common to classical control theory descriptions of systems. Figure 8 illustrates the block diagram representing the equations describing trust in the pasteurisation system. This representation communicates far more information than a simple linear regression model. It describes both which factors are causally involved and also the dynamics of changes in trust as a function of time, how quickly trust erodes when faults occur, and how quickly trust builds with increasing system performance. The quantitative form of the equation for our data is:

$$\begin{aligned}\text{trust}(t) = {} & 0{\cdot}570*\text{trust}(t-1) + 0{\cdot}062*\text{performance}(t) \\ & - 0{\cdot}062*(0{\cdot}210)*\text{performance}(t-1) \\ & - 0{\cdot}740*\text{fault}(t) + 0{\cdot}740*(0{\cdot}400)*\text{fault}(t-1).\end{aligned}$$

The specific time constants and the value of the coefficients that describe the dynamics of trust will vary with different circumstances but the form of the relation, a first order lag model, we believe will generalize across systems. At least the model provides a starting point for further research in an area which has so far been almost completely neglected in human–machine interaction research.

In looking at operator behaviour in a setting which allows great freedom in the choice of individual strategies, there are several methodological difficulties. One is that the model used to describe the dynamics of trust was fitted to data pooled over all the subjects. It was apparent from the data that different subjects used systematically different ranges of trust. One subject's judgements might range between 10 and 7, while another might range between 7 and 4, although the two curves as a function of trial were very similar. A parameter was included to normalize such range effects. This circumvented the problem of spuriously high autoregressive terms, which would otherwise have appeared. We are aware that this is a difficult problem. Ideally one would like to normalize ratings over subjects. Training the subjects on a variety of systems of differing reliabilities, and obtaining a series of baseline ratings might provide the basis for a normalized scale. The amount of time, money, and work involved prohibited such an approach in this experiment, and the steps we have taken are an approximation to deal with this problem. We believe that the strength of our results suggest that our method is adequate, at least for the preliminary study reported here. Our approach is at least conservative.

To strengthen further the relation between our work and the earlier work of Muir, a model of the same form (same factors, but different regression coefficients) was fitted to the ratings of predictability, dependability, and faith to model their dynamics. The model provides the best fit to faith, followed by dependability, and reliability, accounting for 79·8%, 77·9%, and 73·4% of the variance respectively. The degree of fit represents the degree to which the variability of other dimensions of trust can be accommodated by the factors affecting trust. This result is consistent with those of Rempel *et al.* (1985), who believe that faith is most closely associated with an overall impression of trust, followed by dependability and predictability.

In addition to the fit of the model conforming to the theoretical expectations, the relative importance of the factors in the model reflect theoretical expectations. A comparison of the total variance accounted for by each of the factors in the model reveals that the different dimensions of trust (predictability, dependability, and faith) are differentially sensitive to performance, the presence or absence of a fault, and individual differences.

Turning to the effect of system performance, changes in the level of system performance account for a greater percentage of variance of predictability (33·8%), than dependability (27·5%), of faith (26·7%). This result agrees with the theoretical argument of Rempel *et al.* (1985), who claim that predictability primarily depends on observable behaviour. In contrast, the presence or absence of a fault accounted for more variance for dependability (17·7%) than for either predictability (15·6%) or faith (14·9%). Again this result is consistent with the theoretical expectation that dependability represents '... a shift in focus away from specific behaviors, to and evaluation of qualities and characteristics attributed to the partner...' (Rempel *et al.* 1985:96). The variance accounted for by the individual subjects was higher for faith (24·1%) than either dependability (15·8%) or predictability (14·1%). This suggests that faith might represent a more deeply held belief in the capabilities of the system, that varies from subject to subject more than the other dimensions of trust that depend on more concrete observables like performance, or the occurrence of a fault. In summary, it seems that the relative importance of the factors of the model of trust follow the theoretical expectations outlined by Rempel *et al.* (1985), demonstrating the validity of these dimensions of trust.

5. Trust and control strategies

We now turn to an analysis of the relation between trust and the operators' control strategies, specifically the relation between trust and the use of the automatic controllers. The work of Zuboff (1988) and Muir (1989) suggests that increased trust will be associated with an increased use of the automatic controllers. Our results failed to support this hypothesis. Regressing trust in the overall system with the use of the automatic controllers revealed that the use of the automatic controllers increased when trust *declined.* This tendency was statistically significant for all the sub-systems; for the feedstock pump $F(1,787)=43{\cdot}8$, $p<0{\cdot}0001$, for the steam pump $F(1,787)=27{\cdot}2$, $p<0{\cdot}0001$, and for the steam heater $F(1,787)=25{\cdot}3$, $p<0{\cdot}0001$. While statistically significant, changes in trust account for only a very small amount of the variance in the use of these sub-systems, 5·3%, 3·4%, and 3·1% for the feedstock pump, steam pump, and steam heater. Nonetheless, the effect suggests that the allocation strategies of the operator are not a simple a function of trust in the overall system. This result deserves greater scrutiny, particularly since one of Muir's experiments found a very strong positive correlation between the magnitude of trust and the use of automatic control.

5.1. *A qualitative analysis of control strategies*

We begin with the example of the control strategies adopted by a single operator. Figure 9 shows the allocation of automatic control chosen by operator 13. Shortly after training this operator adopted a fixed allocation of control; complete manual control for the feedstock pump, complete automatic control for the steam pump, complete manual control for the steam heater. This pattern was disrupted by the occurrence of the fault. During trials 40 to 60, when the fault was continuously

present, the complete reliance on manual control of the feedstock pump and the steam heater was abandoned in favour of a mixture of manual and automatic control.

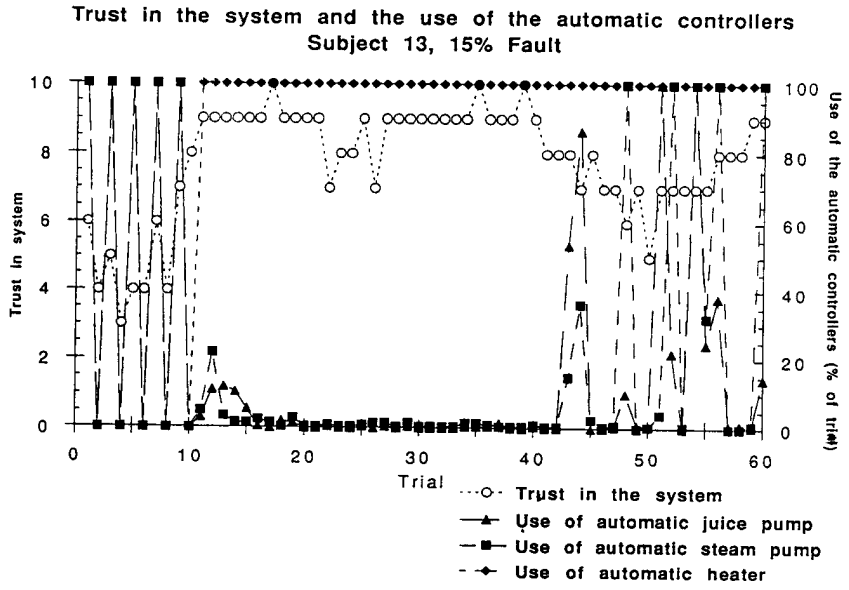


Figure 9. Trust in the system, and the allocation of automatic and manual control chosen by Operator 13 over the three days of the experiment.

During the trials without the fault 11 out of 16 operators followed a similar pattern, adopting complete manual control of at least one of the three sub-systems, and complete automatic control of the other sub-systems. Three operators adopted full manual control of all three sub-systems. Only two operators failed to adopt a strategy consisting of a mixture of complete allocation of manual and automatic control; they controlled the system with a mixture of manual and automatic control similar to that adopted by Operator 13 during trials 40 to 60. In classifying the types of control adopted during normal operation, complete reliance on the automatic controllers was defined by using the automatic controllers for more than 90% of the trial, for at least eight out of the ten trials 30 to 40. Complete manual control was defined as using the manual control for more than 90% of the trial, for at least eight out of the ten trials 30 to 40. During the trials without a fault 14 out of 16 operators chose to control the feedstock pump with complete manual control, while 10 controlled the steam pump with complete manual control, and 8 controlled the steam heater with complete manual control. Table 5 summarizes these results, showing that during the trials without faults subjects generally adopted stable strategies relying heavily on manual control.

When the chronic fault occurred (trials 41–60) these stable strategies were disrupted. The behaviour of operator 13 in figure 9 shows how the allocation of complete automatic and manual control was disrupted, in favour of a mixture of automatic and manual control. The other operators showed a similar pattern, 13 out of 16 operators changed their allocation strategy for the feedstock pump, 8 for the steam pump, and 6 for the steam heater. In most cases these transitions were from all

Table 5. The allocation of function adopted by the 16 operators under normal operation, and during the occurrence of a fault. The numbers in each cell represent the number of operators choosing that mode of control. The bottom line indicates the number of operators who changed their mode of control during the chronic fault.

	Normal operation (trials 31–40)			Continuous fault (trials 41–60)		
	Juice pump	Steam pump	Steam heater	Juice pump	Steam pump	Steam heater
All manual	14	10	8	2	2	4
Mixture	1	2	1	14	10	7
All automatic	1	4	7	0	4	5
Change with fault				13	8	6

manual control to a mixture of manual and automatic control. One operator switched from all automatic control of the feedstock pump to a mixture of automatic and manual control. Similarly, two operators switched from complete automatic control of the steam heater to a mixture of manual and automatic control.

5.2. *A quantitative analysis of expert performance during normal operation*

A quantitative analysis of the data for expert operators (trials where 90% efficiency was exceeded) on trials without a fault shows that, under normal operating conditions, operators tend to adopt rigid allocation of automatic and manual control after achieving high performance. While the strategies differed in detail, the tendency to adopt a fixed control strategy was common with all the effective strategies. The low variance in the number of control actions and the low variance in the percentage of time spent using the automatic controllers for trials with high performance suggests that operators adopt fixed sets of strategies as they become adept at controlling the system.

Table 6. The average number of manual control actions as well as the use of automatic controllers for high (90% and higher) and low (89·9% and lower) performance trials. The standard deviations are included in parentheses. (Using paired t-test, $n=15$, one operator did not have any trials above 90%).

	Low performance (SD)	High performance (SD)	Mean difference (SD)
Number of control actions	13·8 (4·3)	7·5 (2·6)	6·3 (1·7) $p<0·0005$
% of trial with automatic feedstock pump	33·2 (22·2)	1·6 (1·5)	31·6 (21·0) $p<0·0005$
% of trial with automatic steam pump	65·4 (25·3)	38·9 (8·7)	26·5 (16·6) $p<0·0005$
% of trial with automatic steam heater	54·1 (26·8)	33·7 (4·7)	20·4 (16·7) $p<0·00005$

Table 6 shows that once operators discovered an effective balance between automatic and manual control they seldom experimented with other combinations of

automatic and manual control. The variance in the time spent using automatic controllers reflects the tendency to adopt a single allocation strategy. Comparing trials with high performance to those with low performance we find that the standard deviation of the operators' percentage of time spent using the automatic controllers drops from 22·2 to 1·5 for the use of the automatic feedstock pump, and from 25·3 to 8·7 for the automatic steam pump, and from 26·8 to 4·7 for the automatic steam heater, as the operators become more adept at controlling the system.

In addition to developing fixed strategies, characterized by low variance in the operators' allocation of automatic control, the operators also tended to adopt control strategies that involved very few control actions. All but one of the operators used fewer control actions to achieve efficiencies over 90%, as compared to the number control actions used to achieve efficiencies below 90%.

Table 6 also summarizes the number of control actions used with high and low performances as well as the reliance on automatic controllers with high and low performances. The percentage of time spent using the automatic controllers in trials scoring below 90% efficiency makes the comparison of the control actions of the two groups even more striking. In the trials where operators achieved over 90% efficiency they used the automatic controller for the feedstock pump only 1·6% of the time. On the other hand, for trials where operators scored less than 90% efficiency operators used the automatic controller for the feedstock pump 33·2% of the time. The association of few control actions with good performance, together with the reduction in the variance in the operators' allocation of automatic control suggests that as operators become more experienced with the system they adopt a feed-forward manual strategy, based on accurate predictions of the plant's behaviour. This is in agreement with many earlier studies such as those of Crossman and Cook (1974) and Moray *et al.* (1986).

Table 7. The number of manual control actions, the percentage of the total number of manual control actions, and standard deviations in each quarter of a trial, for high and low performances.

Average number of control actions	First quarter	Second quarter	Third quarter	Fourth quarter	Total
High performance (90% and higher)	4·3 58·3% (1·9)	1·2 16·0% (1·6)	1·0 13·8% (1·5)	0·9 12·0% (1·4)	7·4 100% (4·4)
Low performance (89·9% and lower)	4·4 31·8% (2·7)	3·3 24·4% (2·8)	3·0 21·8% (2·6)	3·0 21·9% (2·5)	13·7 (100%) (8·4)

The distribution of control actions within each trial also suggests that operators engage in feed-forward control. Table 7 shows the distribution of control actions within a trial for high and low performances. Good operators achieve high performance by initiating a relatively large percentage of their control actions early in the trial (58·3% during the first quarter of the trial as compared to 31·8% in the first quarter of the trial for poor performers, $z=5{\cdot}64$, $p<0{\cdot}0001$) followed by fewer during the balance of the trial. In the fourth quarter of the trial high performing operators performed only 12·0% of their control actions while poor performing

operators performed 21·9% of their control actions ($z=7·645$, $p<0·0001$). When operators perform well the control actions tend to occur at the beginning of the trial, whereas when operators perform poorly the control actions are spread evenly throughout the trial. This suggests that high performances occur when operators manage the system by performing several control actions early in the trial; in anticipation of the system's future state, as opposed to performing control actions continually, in response to deviations from the expected state. To put it another way they appear to use feed-forward control rather than feed-back control.

5.3. *A quantitative analysis of the effect of faults on control strategies*

The chronic fault occurring with the feedstock pump on the third day disrupted the stable control strategies that the operators had developed in the first two days of the experiment. When faults occurred with the feedstock pump, the actual pump rate failed to correspond to the pump rate that the operator requested. Since the operator could no longer control the feedstock pump accurately, control of the flow rate from the input vat was difficult. This in turn made control of the level of the input vat difficult. The partial loss of control disrupted operators' control strategies, forcing them to investigate alternative means of control.

While the occurrence of the fault disrupted the operators' control strategies, operators were able to adapt and suffer only moderate losses in system performance. The increase in the number of control actions, the changes in the distribution of control actions, and the changes in the percentage of time spent using the automatic controllers reflect the disruption in the operators' control strategies. Table 8 illustrates how faults disrupted the fixed set of control strategies that the operators had developed. Comparing the trials before the continuous fault, where operators had performed above 90%, with trials with the fault reveals a large increase in the variability of the operators' control strategies. Before the onset of the fault most operators had adopted a method of controlling the system that involved a fixed set of control actions and allocation of automatic and manual control. After the occurrence of the fault the large increase in the variability of the number of control actions as well as the variability in the percentage of time spent using the automatic controllers illustrates that the fault made it difficult to achieve high efficiency using a fixed method of controlling the system.

In addition to the increase in the variability of the operators' control strategies, the disruption of the operators' control strategies is reflected by the changes in the mean number of control actions and percentage of time spent using the automatic controllers. Not only does the variability of the number of control actions and variability of the percentage of time spent using the automatic controllers increase, but the number of control actions and the use of automatic controllers increases. With the onset of the fault the number of control actions increases by an average of 9·9 control actions/trial ($p<0·01$, $t=2·72$). The reliance upon the automatic controllers also increases for all but the automatic steam pump, with the use of the automatic feedstock pump increasing by nearly a factor of 20, while the use of the automatic steam heater nearly doubles. These results are summarized in table 8. The data indicate that the occurrence of the fault reduces the operators' ability to control the plant. Even with an increase in the number of manual control actions, the fault forces operators to increase their reliance upon the automatic controllers to achieve good performance. This indicates that the fault makes the system more difficult to

control, requiring more intervention both in the form of manual and automatic control.

Table 8. The number of manual control actions and the amount of time spent using automatic and manual control with and without faults. The standard deviations are included in parentheses. (Using paired t-test, $n=15$, one operator did not have any trials above 90%).

	Without fault (SD)	With fault (SD)	Mean difference (SD)
Number of control actions	7·5 (2·5)	17·4 (6·3)	9·9 $p<0·01$ (3·8) $p<0·001$
% of trial with automatic feedstock pump	1·6 (1·5)	26·0 (22·6)	24·4 $p<0·01$ (21·1) $p<0·001$
% of trial with automatic steam pump	38·9 (8·7)	45·2 (19·5)	6·3 *NS* (10·8) $p<0·025$
% of trial with automatic steam heater	33·7 (4·7)	53·9 (17·2)	20·2 $p<0·1$ (12·5) $p<0·025$

In addition to the increase in the number of control actions and percentage of time spent using the automatic controllers, the distribution of the operators' control actions within a trial reflects the disruption caused by the fault. Instead of entering commands at the start of the trial, as they tended to do before the onset of the fault, operators tended to interact with the system continuously. The proportion of commands entered in the first quarter of the trials without faults is 58·3%, which is significantly greater than the proportion of commands entered in the first quarter of trials containing faults, 33·7%, ($z=17·93$, $p<0·0001$). Additionally, the proportion of commands entered in the fourth quarter of the trial is greater for the trials containing faults, 20·4%, than in those trials without faults, 12·0% ($z=7·62$, $p<0·0001$). Because the faults interfere with the operators' ability to anticipate the future state of the system, control actions cannot be executed at the start of the trial (anticipation of the future plant state), but must be executed continually in response to deviations. Therefore, when faults occur operators shift away from a strategy of based on the knowledge of future states of the system to a strategy based on error correction—from feed-forward to feed-back control.

5.4. *Discussion of control strategies*

Consistent with previous experiments examining the development of process control skills, our data show a reduction in the number of control actions as operators gain experience and achieve higher performances (Crossman and Cooke 1974, Kragt and Landeweerd 1974, Moray *et al.* 1986). To achieve high performance most operators rely upon strategies characterized by infrequent control actions and a stable allocation of automatic and manual control. Most operators preferred manual control, especially for the feedstock pump.

The similarity of the learning curves associated with the operators' adaptation to the system at the beginning the experiment and their adaptation to the fault in the feedstock pump seems to be reflected in the change in their strategies. As the operators initially learn to control the system their strategies shift from exploratory

Table 9. The number of manual control actions, the percentage of the total number of manual control actions, and standard deviations in each quarter of a trial, for trials with and without faults.

Average number of control actions	First quarter	Second quarter	Third quarter	Fourth quarter	Total
Without faults	4·3	1·2	1·0	0·9	7·4
	58·3%	16·0%	13·8%	12·0%	100%
	(1·9)	(1·6)	(1·5)	(1·4)	(4·4)
With faults	5·8	4·3	3·6	3·5	17·2
	33·7%	25·0%	20·1%	20·4%	100%
	(5·2)	(4·2)	(3·8)	(4·1)	(15·6)

behaviour, characterized by highly variable use of manual and automatic control, to very fixed strategies that rely upon feed-forward manual control and a fixed allocation of automatic control. When the fault occurs with the feedstock the pump operators' strategies seem to shift back to exploratory behaviour, as they try different means of control to maintain system performance. The similarity of the initial adaptation to the plant's dynamics and the adaptation following the occurrence of the fault is also revealed in the distribution of control actions in each quarter of the trial. The trials without faults and with high performance suggest feed-forward control, while trials with poor performance and those with the fault suggest feedback control. Figure 10 illustrates the similarity graphically.

Inexperience with the system (trials 11–15) and the occurrence of the fault are reflected in a break in the generally fixed pattern of automatic and manual control, a greater use of the automatic controllers, and a more uniform distribution of control actions within the trial. These periods also show a generally lower level of performance. This suggests that the fault causes operators' otherwise effective control strategies to fail, reducing their expertise to the level immediately after the training session. This reduced capability in manual control seems to provide the impetus to adopt automatic control.

6. Trust, self-confidence, and the use of automatic controllers

This experiment has examined how operators learn to control a semiautomatic process control system. In particular, it examines how operators choose between automatic and manual control to achieve high levels of performance. The time histories of the operators indicate a development of feed-forward control strategies, based on a mental model of the dynamics of the system. These feed-forward control strategies were disrupted when faults occurred. The disruption of the stable and effective strategies accompanied an increase in the use of the automatic controllers. Unlike previous experiments conducted by Muir (1989), which showed that decreases in trust corresponded to decreases in the use of the automatic controllers, this experiment showed that decreased trust led, if anything, to a slightly increased use of the automatic controllers. This suggests that trust alone does not guide the percentage of time spent using the automatic control.

Two factors might lead to a more accurate prediction of the use of the automatic controllers. First, the operators' trust in the individual automatic controllers, as opposed to their trust in the overall system might be a more direct reflection of the

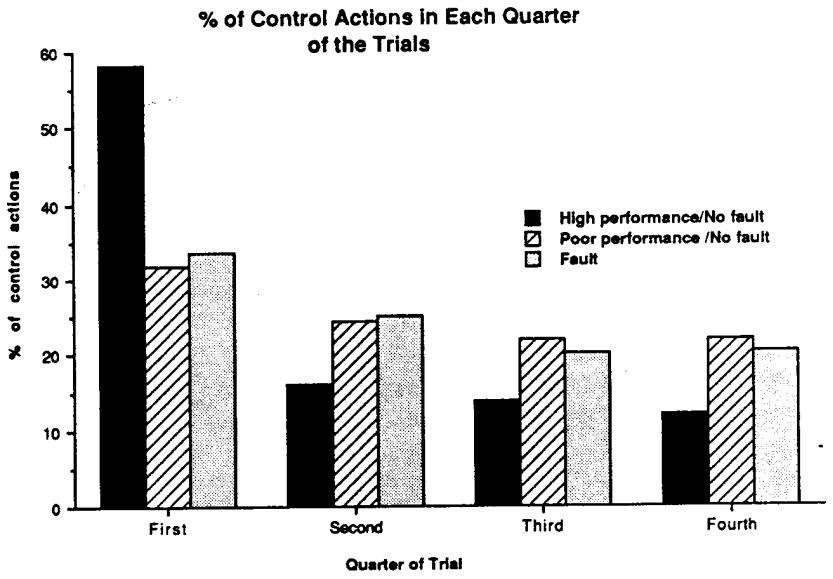


Figure 10. The percentage of manual control actions occurring in each quarter of the trial for trials with and without faults, and trials with poor performance.

operators' willingness to use the specific automatic controllers. Second, this experiment suggests that when faults disrupt the manual control of the system, the switch to automatic control may be more a result of a loss of the operators' confidence in their manual control abilities, than because of an increase in the trust of the system. A loss of trust may result in a disruption of the current stable strategy whatever it is, rather than resulting simply in a reduction in the use of automatic controllers. (After all, if the operators were using nothing but manual control when the fault occurred, they could not use less automatic control.) We predict that if the operators had been predominantly using automatic control at the time of the fault, our initial prediction would have been fulfilled. But, because the majority were running the feedstock pump in manual mode, the shift was away from that mode towards a more exploratory strategy, which necessarily led, in this experiment, to a greater rather than a slighter use of automatics.

Trust paired with self-confidence may therefore provide a better explanation for the operators' choice of manual or automatic control than either construct alone. For example, in this experiment when the fault occurred with the feedstock pump, operators tended to use the automatic feedstock pump controller more frequently. If trust alone guided use of the automatic controller, a drop in the use of the automatic controller might be expected. The operators' level of self-confidence may explain why they tended to use the feedstock pump more often when faults occurred. The fault with the feedstock pump may have drastically reduced their level of self-confidence while reducing their level of trust relatively little. Thus, they may have felt more confident about achieving overall system goals by manipulating the reliable heater and steam pump, leaving the automatic controller to do its best with the faulty feedstock pump. By including the concept of self-confidence in manual control with the concept of trust in automatic control, a better explanation of operators' allocation of function might result.

Assuming that operators allocate functions based on the relative levels of self-confidence in manual control and trust in automatic control, it seems important to observe and predict operators' self-confidence as well as their trust. We are at present conducting an experiment to explore this possibility. Given that trust and self-confidence guide the percentage of time spent using the automatic and manual control, the appropriate use of manual and automatic control will depend upon operators accurately perceiving the capabilities of manual and automatic control. We might speculate that accurately communicating the performance of automatic and manual control, the operators' trust and self-confidence will match the true capabilities of automatic and manual control, and operators will be more likely to use automatic and manual control when they are appropriate.

7. Conclusions

This research provides a first step towards modelling trust between humans and machines, and its influence on operators' control strategies. More specifically, this research had four aims: to examine the factors affecting trust and how trust changes over time, the relation between changes in trust and control strategies, and the effect of 'transient' and 'chronic' faults on development of trust; and to extend Muir's (1989) investigation to a more complex supervisory control situation.

An analysis of the operators' trust, as measured by subjective rating scales, indicates that both system performance and the occurrence of faults affect trust. This suggests that both changes in the dimensions of trust 'performance' and 'process' contribute to overall feeling of trust. The relative importance of the effect of the factors influencing trust (performance, the size of a fault, and individual differences), on other dimensions of trust (predictably, dependability, and faith) suggests that the theoretical development of trust between humans presented by Rempel *et al.* (1985) also applies to trust between humans and machines. In addition, a time series analysis of the data shows that trust has a dynamic nature, being dependent not only on the current size of faults and levels of performance, but also on recent values of performance, fault size, and trust. Taken together, the results show that the multidimensional construct of trust developed to describe trust between humans, together with a consideration of the dynamic aspects of trust, can be used to describe trust between humans and machines.

In addition to describing the factors governing the changes in trust between humans and machines, this experiment shows that, at least in some situations, there is no simple relationship between trust and the use of automatic controllers. While Muir's (1989) results suggest a direct relation between trust and the use of the automatic control, our data shows that the operators' use of automatic controllers depends upon more than trust alone. In this experiment operators generally adopted strategies which depended upon manual control. The chronic fault disrupted these strategies, leading to an increased use of the automatic control, together with a drop in trust. This result suggests that changes in the operators' manual control abilities, along with trust, might be an important factor in guiding the use of the automatic controllers. Further research is needed to investigate the relationship between trust, self confidence, and the use of automatic controllers.

The introduction of the 'transient' and 'chronic' faults illustrate how trust drops and recovers in response in different faults. With the transient fault both trust and performance dropped. Following the transient fault performance recovered immediately, while trust, at least for the large faults, took several trials to recover

completely. With the chronic fault performance dropped immediately and then recovered, as operators developed strategies to cope with the fault. The chronic fault led to a drop in trust over several trials, followed by a recovery that paralleled the recovery of performance. Although both trust and performance recovered during the chronic fault, the recovery of trust was much less than that of the operators' performance. Furthermore, the recovery of performance was accompanied by greatly increased workload as measured by the number of control actions taken by the operators, who appeared to revert to a feedback control strategy from their efficient feed-forward control strategy.

This research extends Muir's work by examining operators in a more complicated supervisory control situation. Because this experiment provided operators with a wide variety of possible strategies, like an actual work situation, some degree of strict experimental control was sacrificed. This becomes important in the examination of the relationship between changes in trust and use of the automatic controllers. Because operators generally adopted a strategy of manual control of the juice pump it is only possible to investigate the relationship between trust and disruption of the manual control strategy. Additional research might structure the situation to induce an increased reliance upon the automatic controller of the juice pump, so that it is possible to investigate the relationship between trust and disruption of the use of the automatic controller.

Acknowledgements

This work was supported by a grant from the University of Illinois Research Board, and the University of Illinois Beckman Fund. We would like to thank Kim Vicente and Penny Sanderson for their insightful comments on earlier drafts of this paper.

References

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

CROSSMAN, E. R. and COOKE, F. W. 1974, Manual control of slow response systems, in E. EDWARDS and F. LEES (eds) *The Human Operator in Process Control* (Taylor & Francis, London).

KRAGT, H. and LANDEWEERD, J. A. 1974, Mental skills in process control, in E. EDWARDS and F. LEES (eds) *The Human Operator in Process Control* (Taylor & Francis, London).

MORAY, N., LOOTSTEEN, P. and PAJAK, J. 1986, Acquisition of process control skills, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-16**, 497–504.

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

MUIR, B. M. 1989, Operators' trust in and percentage of time spent using the automatic controllers in a supervisory process control task, Doctoral thesis. University of Toronto.

PANDIT, S. M. and WU, S. M. 1987, *Time Series Analysis with Applications* (John Wiley, New York).

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

RASMUSSEN, J. 1986, 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.

SHERIDAN, T. B. and FERRELL, L. 1974, *Man–Machine Systems* (MIT Press, Cambridge, MA).

SHERIDAN, T. b. and JOHANNSEN, G. (eds) 1976, *Monitoring Behaviour and Supervisory Control* (Plenum, New York).

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

Appendix 1. Introduction to subjective rating

In this study we are interested in your judgements about how reliable and trustworthy you believe a simulated machine is, so let's talk about trust for a minute.

First, think about your trust in people. We all trust some people more than others. If you think about people you know, you can probably think of some whom you trust very much and others whom you trust much less. We do not trust all people equally, and we can express how much we trust a particular person.

We also think about trusting things, such as products. For example, I trust my Chrysler to start in the morning because it has never failed to do so. I trust my wife's Chevrolet much less because of a history of trouble. I trust one of my computers because I have never had trouble with it, while another is constantly giving me trouble when I try to log on, and I trust it much less.

If we think about it for a moment, we could rate our degree of trust in many of the things we use on a scale like that on the attached sheet. You'll be using a scale like this in the experiment, so I'd like to give you a bit of practice in using it. So let's rate a few of the things you may use often.

Now please rate your trust, your judgement of predictability, your judgement of dependability and your faith, in each of the following.

1. The local bus service to be on time.
2. Your calculator to produce the right answer.
3. The heating system where you live to keep you comfortable.
4. Your watch to tell the correct time.

In this experiment, you will be asked to rate an industrial plant, either as you operate it or as you watch it perform its task automatically. In each trial you will be asked to assess the plant's performance based on four criteria: the system's predictability, the system's dependability, the faith you have in the system and the amount of trust you place in the system.

At the end of each trial the computer displays four screens. These screens each contain a rating scale similar to the four scales shown on the following page. To select a rating simply place the mouse cursor on the desired rating and "click" it.

Any questions? If you have any questions about the scales, don't hesitate to ask.

There are no "right" answers. We are interested in how you see the quality of the plant. Your answers will help us in our search on how to improve the relation between humans and machines.