WHEN TO ACT? MANAGING TIME-ACCURACY TRADE-OFFS IN A DYNAMIC BELIEF UPDATING TASK

Michael Hildebrandt
Department of Computer Science
University of York
York YO10 5DD, UK
hilde@cs.york.ac.uk

Joachim Meyer
Dept. of Industrial Engineering and Management
Ben Gurion University of the Negev
Beer Sheva 84105, Israel
joachim@bgumail.bgu.ac.il

Diagnostic decisions in dynamic environments often require trade-offs between decision accuracy and timeliness. The longer a diagnostic decision is postponed, the more the accuracy of the decision may increase, while at the same time the probability of successfully executing remedial actions decreases. Kerstholt (1994) reports that in a task where a continuous process had to be monitored, subjects' reliance on a judgment-oriented strategy (requesting additional information before making a decision) frequently led to late decisions. In this study, we were interested if similar effects appear when the motivation to postpone the decision was induced by the prospect of an alarm appearing later in the trial. A normative model based on Bayesian belief updating was constructed to determine optimal strategies under the conditions of the independent variables alarm timing (early, late) and alarm reliability (0.7, 0.9). Results are in partial agreement with earlier studies by showing evidence of a judgment-oriented strategy in the low-reliability condition. However, in the high-reliability condition, a high proportion of early decision errors, consistent with an action-oriented strategy favoring decision timeliness over accuracy, occurred.

INTRODUCTION

The pace and complexity of many of today's work domains - such as air traffic control, transportation, or manufacturing - requires operators to manage multiple tasks and to adapt to a dynamic environment. Analysis and design of such systems therefore requires a detailed understanding of the temporal properties of the physical system, the task, the environment and the agents. While some temporal problems such as multi-tasking and decision-making under time pressure have received much attention, other aspects are less well understood. These include the role of time perception in decision-making and control, the control of time-lagged systems, sequence errors (e.g. omission, commission, revision or repetition), duration errors such as temporal overshoot or undershoot, duration neglect in judgment and decision-making, interruption scheduling, human scheduling performance, and temporal awareness (see De Keyser, 1995, for a review).

When-to-act problems

This study is concerned with a particular aspect of time in decision-making, namely biases in the management of accuracy-timeliness trade-offs. For many diagnostic tasks, the quality of diagnostic decisions increases over time, while the probability of successfully executing the action decreases. For instance, the more pronounced a patient's symptoms become, the more certain we can be that the patient is suffering from a particular disease. At the same time, the longer the diagnostic decision is postponed, the more difficult it may become to treat the disease. Kerstholt (1994) used a similar cover story in an experimental study and found that in conditions of high time pressure (rapid deterioration of the controlled process), participants' use of judgment-oriented strategies (requesting additional information about the cause of a problem) led to an increase in system failures. This finding suggests that decision-makers may be biased towards improving accuracy at the expense of timeliness. In the current experiment, the incentive to postpone a decision was induced by an alarm that appeared at some point during the trial and would help to improve the quality of the decision. Based on Kerstholt's (1994) findings, over-reliance on the alarm in situations where a decision should be taken immediately was expected, so that decisions would frequently be late.

METHOD

Task

Participants had to make binary choices between two possible routes for aircraft approaching a sector boundary. The 'direct' route was to be chosen if the adjacent sector was free of turbulence, otherwise aircraft were to be sent on the 'detour'. Two sources of information were available:
System 1 provided a numerical value between 0 and 10, sampled from a signal-noise distribution (see below). The higher the value, the more likely was the presence of turbulence. Participants had to learn over the course of the experiments which values were indicative of turbulence and which ones were not. System-1 information was available at the start of each trial and remained visible during the trial.

System 2 was an alarm that returned a binary recommendation ("turbulence", "no turbulence"). The alarm would appear at some point during each trial.

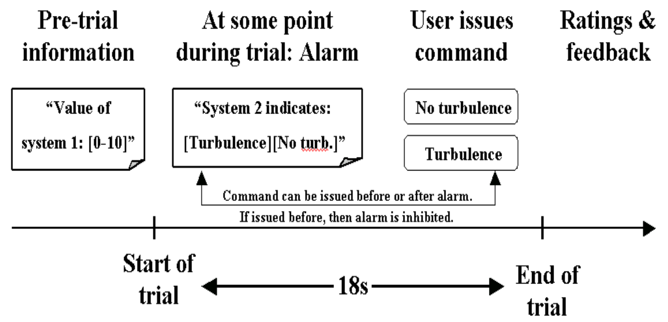


Figure 1. Structure of a trial

Figure 1 shows the structure of a trial. Participants could issue the command at any point during the trial, before or after the occurrence of an alarm. Once made, the decision could not be revised. Also if the decision were made before the occurrence of the alarm, the display of the alarm would be suppressed. Each 18-second trial started with an aircraft entering the sector and ended with the aircraft leaving the sector. Participants received notification about these events and were informed that aircraft passed the sector at a constant speed, but did not have a visual cue (progress marker) indicating the position of the aircraft in the sector. Thus participants had to rely on their own perception of time to estimate the position of the aircraft. This judgment was important as the probability of the aircraft successfully executing the command decreased over time. In other words, the longer participants delayed issuing the command, the less likely it was that the aircraft would execute the command. Overall success on the trial required the decision to be correct *and* timely. There was no binary cut-off point for decision timeliness, but instead the probability of successful execution decreased linearly over the trial (details below).

Reward structure and ratings. The payoff for a correct decision was not fixed; instead participants could choose to invest 0, 5, 10, 15 or 20 credits in the decision. If the decision was both timely and correct, the amount was added to their balance, otherwise it was deducted. If no decision was made by the end of the trial, 20 credits were deducted. This scheme allowed participants to manage the risk of losing or gaining money by investing a large amount if they were confident of success in the trial, and fewer credits if they expected failure.

The following ratings were collected after each trial: Confidence that the correct decision was made; confidence that the decision was made in time; judgment of when participants thought they made the decision in this trial.

Participants were informed about the outcome of the trial (overall success, accuracy, timeliness, credits earned this trial, total credits earned). Participants worked through 100 trials (50 turbulence, 50 no turbulence), with additional feedback questionnaires after every 20 trials. Participants were 40 student volunteers who received financial rewards of £5-10 depending on their performance.

Independent variables

The aim of this experiment was to create conditions where it was normatively correct to *wait* for an alarm (even if the probability of executing the decision decreased), and others where participants should *not wait* for additional information. The choice between these options depended on the difficulty of the decision (defined as the ambiguity of pre-trial information), and the experimental conditions generated by the manipulation of the alarm timing (early, late) and reliability (high, low; both factors between-subjects).

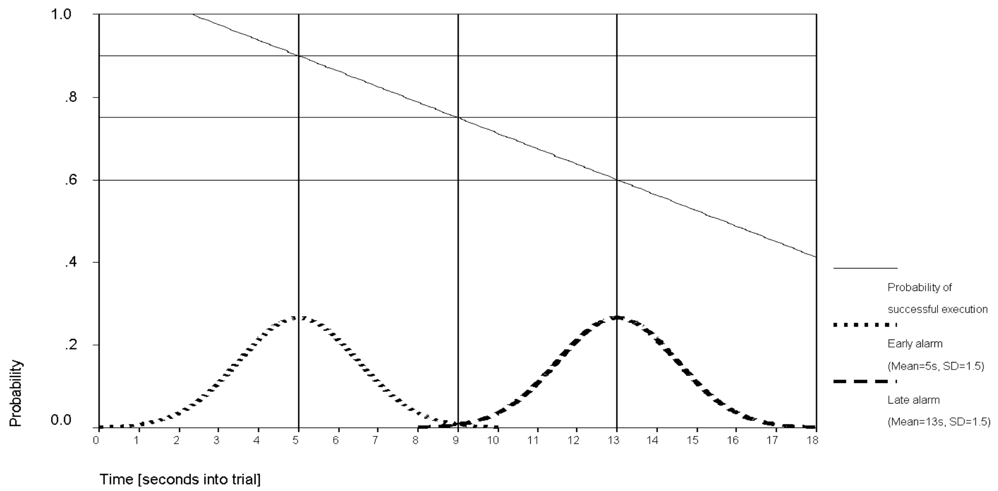


Figure 2. Probability of early and late alarm, and action success, over time

Alarm timing. Participants either received an early alarm (sampled from a Gaussian distribution with a mean at 5 seconds and a standard deviation of 1.5), or a late alarm (mean 13s, SD 1.5). The probability of successfully executing the command decreased across the trial according to the linear function $f(x)=1.0875-x/18*0.675$, so that a decision made before 2.3 seconds was *always* executed, at 5 seconds the probability was 90%, at 13 seconds 60%, and at the end of the trial at 18s, the probability was 41% (Figure 2).

Alarm reliability. The second independent variable was the reliability of the alarm ("system 2"). The highly reliabile alarm correctly identified the presence or absence of turbulence in 90% of cases and incorrectly in 10% (low reliability: 70% and 30%). To generate the information provided by system 1, a signal detection paradigm was used, with the value for turbulence trials sampled from a Gaussian distribution with a mean at 5 and a SD of 1 (no turbulence: mean=4, SD=1).

Normative model: Bayesian updating

To decide whether a system-1 value ("x") was indicative of turbulence, participants had to learn to discriminate the two distributions. With the payoff matrix symmetrical and p(turbulence)=p(¬turbulence)=0.5, participants should assume the presence of turbulence for any value above 4.5. At this indifference point, the conditional probability of turbulence given x is 0.5:

$$p(turb|x)=p(x|turb)*p(turb)/p(x)$$

The probability of turbulence given an alarm can be calculated as

$$p(turb|alarm)=p(alarm|turb)*p(turb)/p(alarm)$$

This equation returns 0.9 for the high and 0.7 for the low reliability group.

With the diagnostic value of the two information sources calculated individually, we can compute their combined diagnostic value using the Bayesian updating equation:

$$p(turb|x\cap alarm)=p(turb)*p(x|turb)*p(alarm|turb)/p(x\cap alarm)$$

Figure 3 shows the criterion shift in the cumulative probability function provided by the high and low reliability alarm. Whereas without an alarm, participants should respond 'turbulence' for any value above 4.5, this criterion shifts to 2.3 and 3.65 with a high and low reliability alarm, respectively. Without an alarm, the probability of making the correct decision is lowest (50%) half way between the means of the signal and noise distribution (x=4.5). At this indifference point, the occurrence of an alarm increases the probability of the presence of turbulence to 0.7 (low reliability) or 0.9 (high reliability). Notice that when the alarm indicates no turbulence, the cumulative probability function p(¬turb|x∩noTurbAlarm) is a projection of p(turb|x∩turbAlarm) along the ordinate at x=4.5.

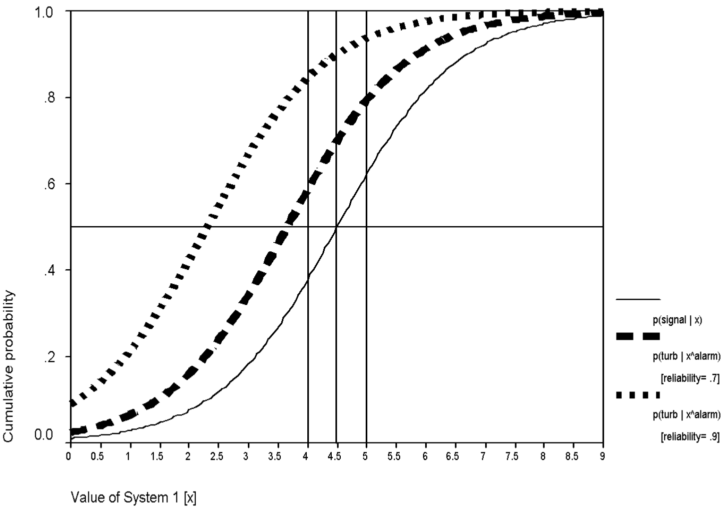


Figure 3. Cumulative probability function p(turb|x) and (turb|x∩alarm)

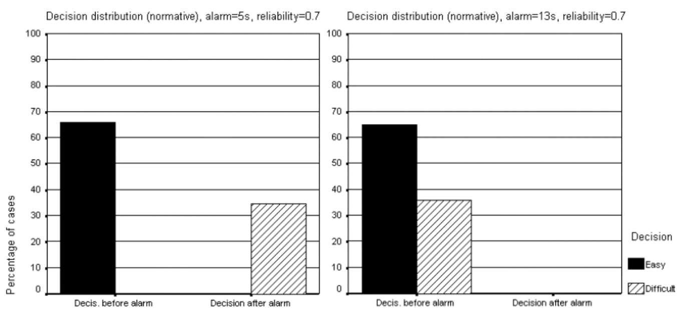
Optimal strategies. To decide whether to wait for an alarm, the gain in decision accuracy has to be traded off against the decrease in execution probability. As success in a trial is defined as p(success)=p(decisionCorrect)*p(decisionIn Time), and assuming decisions made on the basis of system 1 alone are taken at the start of the trial (where p(decisionIn Time) = 1), the normative strategy is to wait for the alarm if

$$p(turb|x)*p(x|turb) < p(decisionInTime[time])*(p(turb|x\cap turbAlarm)*p(x|turb)+(p(noTurb|x\cap noTurbAlarm)*p(x|noTurb)-p(turb|x\cap turbAlarm)*p(x|turb))/2)$$

Based on this model, temporal strategies can be derived for the different experimental conditions so that the overall probability of success on a trial, p(correct decision) * p(decision in time), is maximized:

- *Early alarm, high reliability:* Wait for alarm for system-1 values between 3.3 and 5.7, otherwise decide immediately
- *Early alarm, low reliability:* Wait for alarm for system-1 values between 4.0 and 5.0, otherwise decide immediately
- *Late alarm, high reliability:* Wait for alarm for system-1 values between 4.35 and 4.65, otherwise decide immediately
- *Late alarm, low reliability:* Always decide immediately

It becomes clear from these strategies that the decision to wait for an alarm relies critically on the value of the a-priori information ("system 1"). If this information is sufficiently predictive of the presence of turbulence, then waiting for the alarm provides little additional benefit, but instead reduces the overall probability of success by the decrease in the probability of successful command execution over time. However, the closer the system-1 value is to the indifference point (x=4.5), the more difficult it becomes to make the correct decision based on system-1 information alone, and the greater the gain in decision accuracy from waiting for the alarm becomes. As decision difficulty plays such an important role in determining the correct temporal strategy, it was included in the analysis as a within-subjects factor. *Difficult decisions* are defined as any system-1 value between 4 and 5 (the means of the signal and noise distribution), with values below 4 and above 5 classed as *easy decisions*. Figure 4 shows the expected number of early and late decisions (i.e. before or after the alarm) in the different timing and reliability conditions and split by decision difficulty.



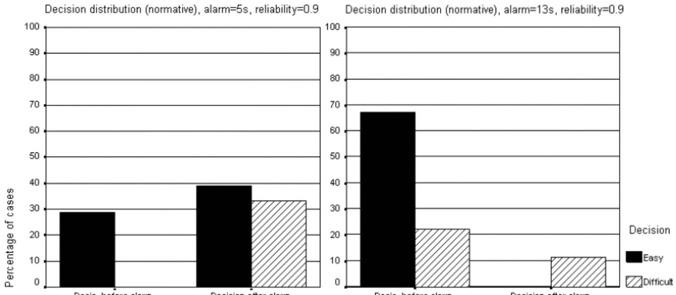


Figure 4. Normative distribution of pre- / post-alarm decisions by alarm time, reliability and decision difficulty

RESULTS

In this paper, we present selected results about the effects of alarm timing, alarm reliability and signal difficulty on decision accuracy, decision timeliness and temporal strategies. A paper reporting the full set of data, including feedback ratings and self-reported strategies, is in preparation.

In this section, we will first present data about overall success before moving on to more detailed analysis of decision accuracy and timeliness. The latter results are analyzed using repeated-measures analysis of variance (ANOVA) with between-subjects factors alarm timing and reliability, and within-subjects factors signal difficulty (easy: signal value between 4-5, difficult: signal value < 4 and > 5) and decision timing (decision before alarm, decision after alarm). In all cases, the first 20 trials were excluded from the analysis as practice trials.

Overall credits earned. Figure 5 shows the mean amount of credits earned during the experiment. ANOVA of this dependent variable shows a highly significant effect of alarm timing ($F_{(1,36)}$=19.6, p<.001) and reliability ($F_{(1,36)}$=19.3, p<.001), but no significant interaction.

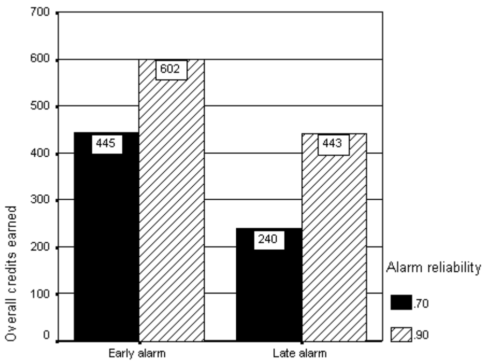


Figure 5. Credits earned by alarm timing and reliability

Trial success and investment. The amount of credits earned at the end of the experiment depended on the success of the command (correct and timely), and the amount of credits invested in the decision (the invested amount was added in case of success and subtracted in case of failure). ANOVA of trial success showed a higher rate of overall success for pre-alarm decisions ($F_{(1,22)}$=6.4, p<.05). This factor interacted with alarm timing ($F_{(1,22)}$=7.3, p<.05; Fig. 6) and decision difficulty ($F_{(1,22)}$=9.5, p<.01). Both independent variables (alarm timing, reliability) had a significant main effect ($F_{(1,22)}$=26.4, p<.001; $F_{(1,22)}$=6.6, p<.05).

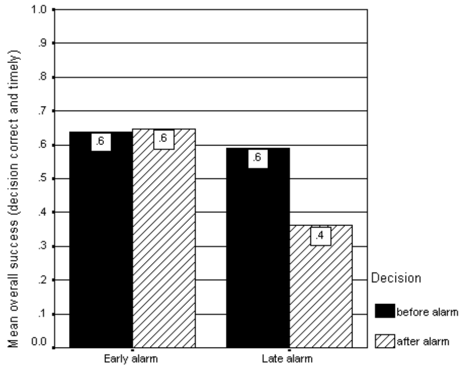


Figure 6. Effect of decision timing and alarm timing on overall success

ANOVA of the amount of credits invested in each decision showed a significant 3-way interaction (Fig. 7) between alarm timing (5s, 13s), signal difficulty, and decision timing (before, after alarm), so that an easy decision, an early alarm and decisions made prior to the alarm were associated with the highest investment ($F_{(1,22)}$=5.2, p<.05). Alarm timing, but not reliability, had a significant effect ($F_{(1,22)}$=14.8, p<.001).

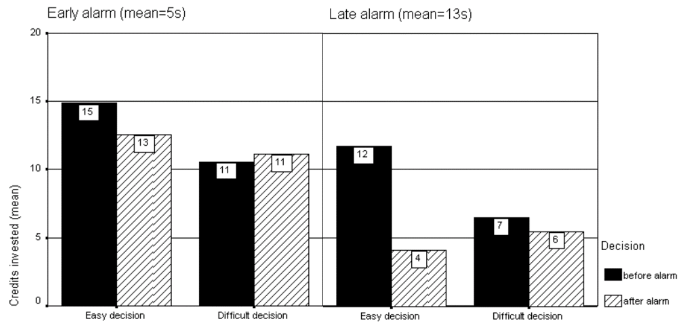


Figure 7. Mean amount of credits invested in a decision by decision difficulty, decision timing and alarm timing

Decision accuracy and timeliness. Results show a highly significant interaction between decision difficulty and decision timing on the correctness of the decision ($F_{(1,22)}$=18.9, p<.001).

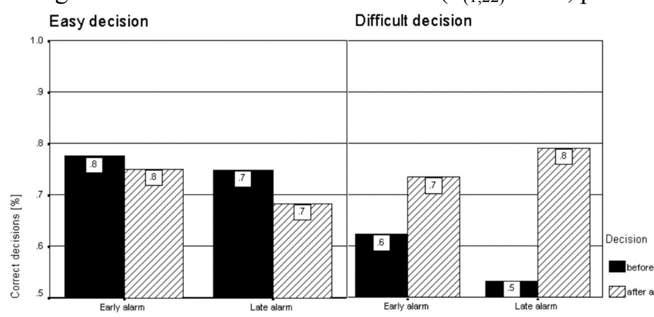


Figure 8. Effect of alarm timing, decision difficulty and decision timing on decision correctness

The marginally significant 3-way interaction of these factors with alarm time (Fig. 8) shows that this effect is mainly due to a marked difference in decision correctness before and after the occurrence of an alarm for difficult decisions in the late alarm condition ($F_{(1,22)}$=3.2, p<.09). Alarm reliability, but not alarm timing, had a highly significant effect ($F_{(1,22)}$=12.8, p<.005). Predictably, a significant main effect of alarm timing ($F_{(1,22)}$=45.4, p<.001) on action success and an interaction of this factor with decision timing (before or after alarm, $F_{(1,22)}$=24.5, p<.001) was found.

Temporal errors. Comparison of the normative decision point for each trial with the actual timing of a participant's decision yielded a classification of decisions as correct early, correct late, too early and too late. Fig. 9 shows the distribution of temporal errors across the experimental conditions. Not surprisingly, the late alarm condition, where it was normatively correct to decide before the alarm in the vast majority of cases (c.f. Fig. 4), showed only few temporal errors. In the early alarm conditions, however, 30-40% of decisions were not made at the normatively correct point in time. Interestingly, the distribution of error types (too early / too late) reverses between the two reliability groups, so that in the low reliability condition, the majority of errors are late decisions (consistent with a judgment-oriented strategy), while in the high-reliability condition, the predominance of early decision errors suggests an over-reliance on an action-oriented strategy, favoring decision timeliness over accuracy.

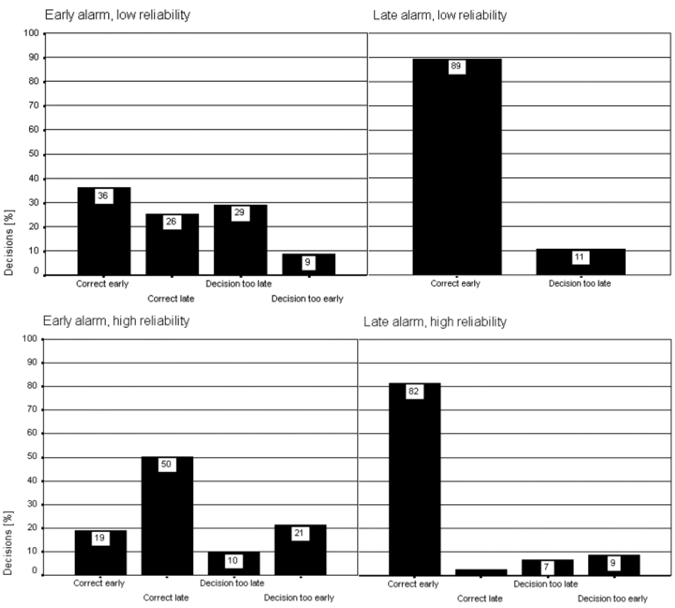


Figure 9. Temporal errors by alarm timing and reliability

DISCUSSION

This experiment required participants to manage the trade-off between decision timeliness and accuracy. The overall benefit of earlier and more reliable alarms (Fig. 5) can be attributed to the positive effect of high alarm reliability on decision accuracy, and to higher levels of action success and higher average investments for early alarms. However, the high rate of temporal errors observed in the early alarm group suggests that performance might have been improved further by adopting a more accurate temporal decision strategy.

Both the quantitative data and written feedback show that participants were sensitive to the parameters involved in the decision. The following quote from a participant in the early alarm, low reliability group illustrates their reasoning:

> *Any value in system 1 below about 3.5 or above about 5.5 I would immediately select direct or detour respectively, and bank 20 credits. Values within that range I would usually wait for system 2 and base my answer on that, however if I counted in my head that I had waited more than about 7 or 8 sec and system 2 hadn't given a result I would say direct if less than 4.5 or detour if more, and only back 5 or ten credits.*

More research is necessary to determine the factors that seem to induce the pattern of early decision errors in the high reliability group and late decision errors in the low reliability group. These factors may include the variability of alarm timing, misperception of elapsed time, misjudgement of the change in action probability, or over reliance on one of the information sources due to misjudgement of its predictive value. Visual display features such as progress indicators may support more accurate decision timing.

CONCLUSION

Control decisions and alarm responses in dynamic real-time systems are not isolated, one-off events, but are embedded in a process of information seeking and receiving, belief updating, and action. A common problem in such situations is the when-to-act dilemma, where a decision can be taken on current information, or alternatively further information could be sampled. Unless the temporal properties of the task, the environment, the available information sources, and the temporal control heuristics of the operator, are analysed in greater detail, this important class of temporal error phenomena may be neglected in system analysis and design. Formal modelling of the decision plays an important role in this process by providing a standard against which performance can be compared. This line of research may provide useful insights into the role of alarms by focusing not only on an operator's *reaction* to an alarm, but on the *prospect* of receiving the alarm, and the reasoning involved in deciding whether to wait for it or not.

ACKNOWLEDGEMENTS

This research was funded by the UK Dependability Interdisciplinary Research Collaboration and a Minerva short-term research grant.

REFERENCES

De Keyser, V. (1995). Time in ergonomics research. *Ergonomics, 38*, 1639-1660.

Kerstholt, J.H. (1994). The effect of time pressure on decision-making behaviour in a dynamic task environment. *Acta Psychologica, 86*, 89-104.