Overcoming the cry-wolf effect: Operators' checking behaviour in response to different levels of imperfect alarms in a dual-task paradigm

Nina Gérard and Dietrich Manzey

Key words: Alarm Reliability, Compliance, Process Control, Checking Behaviour, Dual Task

Abstract

Behavioural responses to alarms are essentially influenced by the perceived validity of the alarm system, i.e. the number of misses and false alarms. Recently, especially the problem of false alarms and the associated so-called "cry-wolf effect" has become more and more popular. These effects have mainly been investigated in studies using experimental paradigms that forced the operator to choose between reacting to the alarm immediately or ignoring it. The present laboratory experiment investigated the effect of different base rates of critical incidents on the reliance and compliance of operators. An operator workplace in a chemical plant was simulated that requested subjects to complete two tasks simultaneously and that gave participants the possibility to check the raw data behind the alarm. Five levels of imperfect alarms, operationally defined as number of misses and false alarms, have been investigated. Analyses of the behavioural data showed no evidence of a cry-wolf effect. In contrast, operators seem to be aware of conditional probabilities and adapt their behaviour accordingly, showing a responsible way of handling alarms.

Introduction

The increasing introduction of automation in operator workplaces in different domains like power plants, production, cockpits and air traffic control rooms have led to a shift in operators' tasks. Tasks have metamorphosed from active intervening over primary control to shared control with the automation. The main goal of this redistribution of tasks is to enable the operator to allocate his resources to other task. Despite this attempt to decrease operators' workload by this concept of human and machine "team work", there is still no perfect process without accidents or incidents. Even if the term team work is used frequently in this context, intervention by the human operator results nowadays primarily from distrust in the automated system provoked by error-prone automation monitoring detected by the human ("supervisory control"; Sheridan, 1992). There are four possible outcomes including two system errors resulting of the actual state of the world and the dichotomous diagnosis of the alarm system (alarm / no alarm; see table 1).

Table 1: State-alarm-contingency

	Alarm	No Alarm
Failure	Hit	Miss
System OK	False Alarm	Correct Rejection

The proportion of hits out of all critical system states is called the hit-rate of a system. The proportion of false alarms out of all failure-free system states is called the false alarm-rate of a system. The information processing of raw data by the combined human-alarm system can be described in terms of a two-stage decision making process based on signal detection theory (SDT). The alarm system has a certain sensitivity (i.e. a reaction threshold that is determined by technical constraints) that leads to the emission of a specific amount of (true and false)

alarms. In a second decision making level, the human operator decides how to react to this diagnosis of the alarm system, depending on his/her trust in the system.

The perceived reliability of an automated system seems to be the main factor influencing human trust in automation (Bliss, Jeans & Prioux, 1996). Meyer distinguishes between two behavioural aspects of trust: compliance that refers to the immediate and required reaction of the operator to an alarm, and reliance that means that the operator refrains from such an action when the alarm system signals no critical state. High-reliable alarm systems lead to highly compliant operators, whereas systems with a low reliability lead to non-compliant behaviour. For example, imperfect alarms with a high false alarm rate tend to be ignored by the operator (decreased compliance). This phenomenon is also known as the cry-wolf effect (Breznitz, 1983). On the other hand, misses are supposed to decrease reliance. Recently, research has focused on false alarms since studies found an effect of false alarms on compliance and reliance suggesting that false alarms undermine trust in the system more generally, affecting not only reliance, but also compliance (Dixon, Wickens & McCarley, 2007). These effects have mainly been investigated in studies using experimental paradigms that forced the operator to choose between reacting to the alarm immediately or ignoring it (Dixon, Wickens & McCarley, 2007; Meyer, Feinshreiber & Parmet, 2003). Generally, one could argue that alarm systems are only installed if they have a satisfactory reliability (to serve their purpose) but this assertion does not take into account failure base rates. Base rates describe the rate of occurrence of critical incidences in a process in practice. These incident base rates are commonly very low, especially in high risk environments like in nuclear power plants or in cockpits. However, base rates have a crucial influence on the so-called a posteriori probability. The a posteriori probability is a measure of reliability that Meyer (2002) describes as the probability that there is really a critical event given an alarm ("positive predictive value"; PPV). The PPV is the percentage of hits out of all alarms generated by the system. The PPV rises with increasing base rates, as the proportion of hits compared to false alarms increases. This can be explained by the fact that there are many critical incidences to be detected with high base rates whereas there are basically only few failure-free events that could cause a false alarm (even if the alarm system has a low sensitivity!). As a consequence, the base rate can undermine the reliability of an alarm system without any change in its sensitivity, that is, its hit- and false alarm-rate (Parasuraman, Hancock & Olofinboba, 1997). With low base rates resulting in a PPV at a medium level a rational operator's behaviour should be the cross-check of the raw data of the process. At a very low PPV, checking behaviour should be considered as redundant because most alarms represent false alarms. Similar, at very high PPVs checking is dispensable as alarms are mainly hits. But are operators sensitive to conditional probabilities at all? Are a posteriori probabilities the factor determining their behaviour? Or do humans adapt to hit- and false alarm-rates? We hypothesized the checking behaviour to follow an inverted u-shape in function of different levels of a posteriori probabilities. A study was conducted that required participants to complete two tasks simultaneously in a simulation of a control room workplace. In addition to the two extreme response patterns of ignoring or complying with an alarm, the monitoring task gave participants the possibility to check the raw data behind the alarm ("informed compliance") and then, in a second step, to decide whether to comply with it or ignore the alarm.

Methods

Participants

80 paid volunteers participated in the experiment. Their ages ranged from 19 to 52 years (M = 27,24; SD = 5,87). 16 females and 16 males were allocated to each of the five conditions.

The Task Environment

A PC-based laboratory multi-task environment was used for the investigation (Multi-Task Operator Performance Simulation 2, "M-TOPS 2"). M-TOPS 2 is a simulation of an operator workplace in the control room of a chemical plant. The user interface is shown in figure 1.

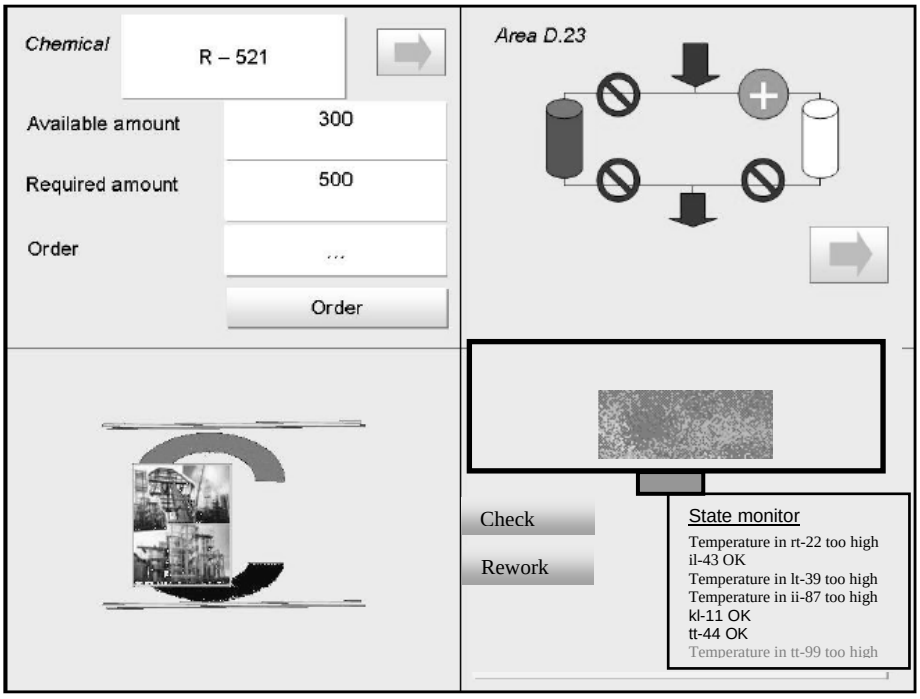

Fig. 1: User interface of M-TOPS 2

Ordering Task (top left). Participants were required to subtract the amount of stocked chemicals ("available amount") of the amount of needed chemicals ("required amount") to keep the process going. They then had to enter the result, send the command by clicking on the corresponding button ("order") and finally click on the arrow in the upper left to get a new trial after a delay of 3 seconds. If the participant had not sent an order within 15 seconds, a new trial was displayed.

Refilling Task (top right). This task represents a tank refilling task but had to be ignored by the participants (for detailed description of this task, see Domeinski, Wagner, Schöbel & Manzey, 2007).

Monitoring Task (lower right). The task in the lower right quadrant represents the monitoring task. Participants saw a fuzzy image of the content of a reaction chamber. A coloured alarm bar under the monitoring screen, signalled if the chemical product in the container was intact (green light) or if the temperature was too high (red light). A state monitor gave a more detailed diagnosis with the exact label of the actual container (e.g. "Temperature in container rt-22 too high" or "Container ll-41 OK"). If the participant trusted the automated system, he/she should click on the button "rework" with a red light and ignore the container when the alarm system displayed a green light. In contrast, if he did not trust the system, the participant had the (time-consuming) possibility to cross-check the raw data behind the alarm by clicking on the button "check", choosing the exact container label and thereby starting a measure that showed the actual heat distribution in the container (the raw data).

The alarm system had a hit rate of 80% and a false alarm rate of 40%. However, base rates changed over five different levels (5%, 18%, 33%, 54% and 81%) leading to five different levels of PPV: 10%, 30%, 50%, 70% and 90%. As can be seen, even if system parameters are

not changed, the reliability in terms of a posteriori probability varies over a broad range depending on base rates.

Design

A mixed design (5 x 2) was used with five levels of a posteriori probability (PPV) as the between-subjects factor (PPV: 10%, 30%, 50%, 70% and 90%). The within-subjects factor was represented by a time-on-task factor, i.e. performance in the two tasks was assessed for two successive blocks lasting 800 seconds each. The proportion of hits (out of all alarm trials) by participants (i.e. their PPV) and their proportion of checking behaviour, reworking behaviour and ignoring behaviour (out of all trials per block) served as dependent variables in the monitoring task. The absolute amount of correct orders sent within each experimental block served as a measure for performance in the ordering task. Participants' trust in the alarm system after their first interaction with it (consisting of 100 trials) was operationalized by estimates of the system's proportional hits, false alarms, misses and correct rejection out of the 100 trials.

Procedure

After a detailed instruction for the two tasks to be performed in the M-TOPS 2 environment including short practice trials for the ordering task (120 seconds) and the monitoring task (180 seconds), participants worked on the monitoring task again, this time as long as 100 containers would pass the screen. To experience the alarm system's reliability, they received an acoustic feedback following each action they chose after the diagnostic support of the alarm system. By means of the acoustic feedback participants had the occasion to learn to know if the system was right or wrong for every single trial and to build an overall impression of its actual reliability. After all the 100 containers in this experience block had passed, participants were requested to estimate the proportion of hits, false alarms, correct rejections and misses the alarm system had produced in the 100 trials and were then informed about their real distribution. This procedure made sure that participants experienced the system's reliability themselves and at the same time that they adapted their expectation of the system's performance to the correct level of reliability. After the experience block, feedback was removed for the following two experimental blocks that lasted 800 seconds each (the time 100 containers would need to run through the screen without any interventions). After each experimental block, a detailed feedback was given showing the points gained in the ordering task and in the monitoring task. The feedback for the monitoring task showed the actual state of the system, the alarm system's reaction and the reaction of the participant resulting in a gain of 2 points for a correct response (hit and correct rejection) and a loss of 2 points for an incorrect response (miss or false alarm) by the participant. At the end of the session all participants completed a short survey containing some questions on demographic background, were debriefed and paid for their participation.

Results

Performance Measures

Monitoring task. Multivariate Analyses of Variance revealed no main effect of level of a posteriori probability on participants' PPV, $F(4,75) = 1,22$, n.s. Participants complied with false alarms at a very low rate throughout all five conditions (mean PPV across conditions: M = 95, SD = 09). There was neither a significant main effect of block, nor a significant interaction.

Ordering task. No significant differences were found in the frequency of correct orders sent by participants in the different a posteriori conditions. However, we found a significant main effect of block, $F(1,75) = 850,32$, $p < .001$ in the sense that participants performed better in the second block than in the first.

Behavioural Categories in the Monitoring Task

Reworking. Multivariate Analyses of Variance revealed a significant main effect of level of a posteriori probability on the proportion of participants' reworking behaviour for alarm trials out of all the three behavioural categories (rework, ignore and check), $F(4,75) = 29{,}36$ $p < .001$ (see figure 2). There was neither a significant main effect of block, nor a significant interaction.

Ignoring. The manipulation of a posteriori probabilities led to no significant effect on the proportion of participants' ignoring behaviour (see figure 2). There was a significant main effect of block, $F(1,75) = 6{,}87$, $p = .002$, but no significant interaction.

Checking. The proportion of checking behaviour was significantly influenced by a posteriori probability on alarm trials, $F(4,75) = 14{,}38$, $p < .001$ (see figure 2). There was a significant main effect of block, $F(1,75) = 3{,}53$, $p = .044$, but no significant interaction.

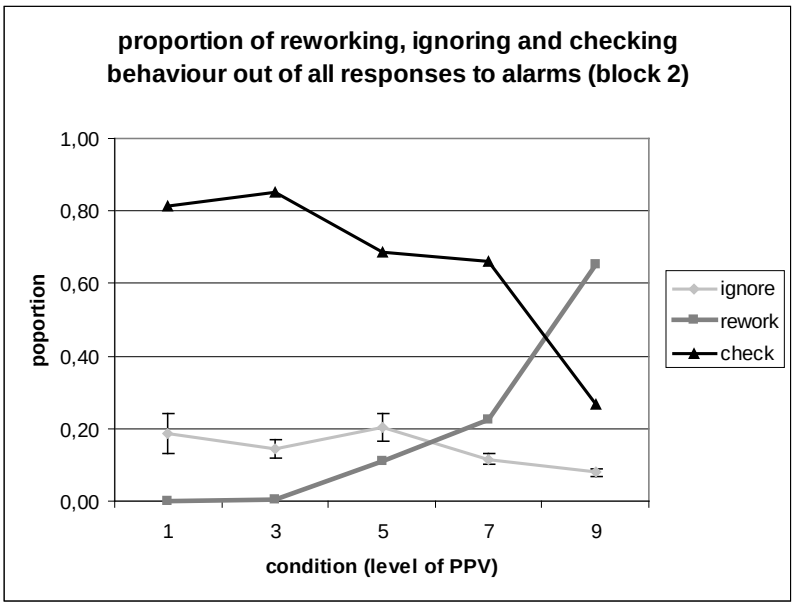

Fig. 2: means of behavioural proportions across conditions (block 2)

Estimation of the true system's reliability

Descriptive statistics showed that participants perceived the PPV to be low for the 10% condition ($M = 16$, $SD = 9$), gave a medium judgment for the 30% ($M = 57$, $SD = 18$), 50% ($M = 59$, $SD = 1$) and 70% condition ($M = 65$, $SD = 11$) and an elevated judgment for the 90% condition ($M = 83$, $SD = 11$).

Discussion

In the presented study, participants worked in a simulated dual-task operator workplace, aided by an alarm system in the monitoring task. This alarm system had a constant hit-rate and false-alarm rate across all conditions, but varied in its positive predictive value caused by different underlying failure base rates.

Participants showed a responsible way of handling alarms of different a posteriori probabilities. With very low PPV, they checked the raw data behind the alarm instead of ignoring it. This checking behaviour decreased only slightly across conditions and then showed a steeper slope at the 90% condition. It seems that participants did not adjust their behaviour to the hit- and false alarm-rate (that were kept constant over all conditions) but that they behaved according to some kind of probability matching of the underlying a posteriori probabilities. Contrarily to our expectations, participants showed no evidence of a cry-wolf effect with low PPV but

behaved more carefully in dealing with low alarm reliabilities. These findings suggest that the cry-wolf effect might be due to experimental settings rather than to the incapability of humans to judge conditional probabilities.

A limitation of the study's generalizability is the balanced payoff-matrix. Participants lost the same number of points for a miss and for a false alarm and got the same (positive) number of points for a hit and a correct rejection. In many domains (e.g. in nuclear power plants or in aviation), a miss may have much worse consequences than would have a false alarm. Nevertheless, there are other automation-dominated fields like production, where a false alarm means a loss of precious time (and money) and where a miss would only result in an imperfect product that would be excluded from further use. Furthermore, this paper dealt mainly with human's perception of and reaction to complex statistic concepts like the a posteriori probability. Before we can give guidelines to designers how to construct an optimal alarm system, we have to acquire a detailed knowledge of human reactions to different patterns of alarm characteristics.

The results of this study imply that

1. base rates of specific processes should be taken into account as they influence the a posteriori probability of alarms and because operators seem to be sensitive to these probabilities
2. work places should always give the possibility to check the raw data to the operator as this behaviour seems to counteract the cry-wolf effect

It is evident that these recommendations cannot be followed without limitations: the sensitivity of an alarm system is always determined by technical constraints. However, designers can manipulate its criterion setting, which means that they can set a high threshold, so that the alarm system gives less alarms resulting in more misses and less false alarms or they can set it at a lower threshold which would lead to more alarms implying less misses and more false alarms. In choosing an appropriate response criterion, designers should always take into account the underlying base rates of the process, not only the system's sensitivity. The higher the base rates are, the higher will be the system's a posteriori probability and the scope for the criterion setting.

References

Bliss, J. P., Jeans, S. M. & Prioux, H. J.(1996) .Dual-task performance as a function of individual alarm validity and alarm system reliability information. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, 1237-1241.

Breznitz, S. (1983). *Cry-wolf: the psychology of false alarms.* Hillsdale, NJ: Erlbaum.

Dixon, S. R., Wickens, C. D. & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?. *Human Factors, 49*, 564-572.

Domeinski, J., Wagner, R., Schöbel, M., & Manzey, D. (2007). Human redundancy in automation monitoring: Effects of social loafing and social compensation. *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, 587-591.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*, 196-204.

Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors*, *44*, 343-354.

Parasuraman, R., Hancock, P. A. & Olofinboba, O. (1997). Alarm Effectiveness in Driver-Centered Collision-Warning Systems. *Ergonomics, 40*(3), 390-399.

Sheridan, T. (1992). Telerobotics, Automation and Human Supervisory Control. Cambridge, MA: MIT Press.