Is Operators' Compliance with Alarm Systems a Product of Rational Consideration?

Rebecca Wiczorek and Dietrich Manzey,
Berlin Institute of Technology,
Berlin, Germany

Most theories about operators' responses to alarm systems suggest that the operators' behavior is guided by their trust towards the system which in turn results from the subjective perception of system properties, namely the perceived reliability of the alarm system. However, some doubts about that assumption have arisen as recent research has not proven the mediating effect of trust. The purpose of this research was to examine the relationship between alarm system properties, trust, and behavior. The alarm reliability was varied while keeping the other system properties constant. It was found that participants' response-rates to alarms were predicted by their perceived alarm reliabilities. However, no mediation by trust could be established. These results suggest that operators' behavior is not always guided by their trust towards the system. Under specific circumstances their compliance rather depends on rational consideration regarding the most efficient strategy.

INTRODUCTION

Operators' responses to alarms can be understood as the result of a decision-making process under uncertainty (Meyer, 2004). This holds in particular for situations where operators do not have the opportunity to cross-check the validity of an alarm against other information sources (e.g. system raw data). Thus, operators need a certain rational criterion as basis for their decision to comply or not comply with an alarm. It has been argued that operators' responses to alarms are mainly related to their trust in the system (Lee & See, 2004; Lees & Lee, 2007). According to several theories trust in automations like alarm systems is determined by different aspects, e.g. performance, process and purpose (Lee & Moray, 1992) or predictability, technical competence and operators' expectation towards the automation (Muir, 1994).

The trust towards the alarm system in turn is assumed to be determined by the perceived reliability of an alarm system (e.g. Kantowitz, Hanowski & Kantowitz 1994; Madhavan Wiegmann & Lacson, 2006; Stanton, Ragsdale & Bustamante, 2009). In terms of signal-detection theory (SDT) the reliability of an alarm system can be defined as the percentage of correctly identified events, i.e. the relative proportion of hits and correct rejections out of all possible events. However, deciding on whether or not to trust in an alarm and to comply with an alarm, another system property seems to be more important, i.e. the "positive predictive value" (PPV). The PPV of an alarm system is defined as the probability that an alarm truly indicates a dangerous condition (Getty, Swets, Pickett & Gonthier, 1995). The PPV can be seen as the specific *alarm* reliability and therefore proposed to affect the willingness to respond to a given alarm to a greater extent than does the overall system's reliability. It exists also the analogues concept of *non alarm* reliability or "negative predictive value" (NPV), i.e. the probability that the absence of an alarm truly indicates a non dangerous situation (Meyer & Bitan, 2002). That should affect on its part the operators' reactions to the absence of alarms. However, the objective PPV and NPV can differ considerably from the subjectively perceived one as earlier research has shown that true system reliabilities often get underestimated by operators (Wiegmann, 2002).

Many different experiments have shown a link between system properties and operator's resulting behavior (e.g. Bliss, Gilson & Deaton, 1995; Getty et al., 1995; Meyer, 2001; Bustamante & Bliss, 2004; Gérard & Manzey, 2009; Rice, 2009). Specifically, a high number of false alarms (i.e. low alarm reliability) has been found to be the main reason for a decrease in compliance with alarms (Rice, 2009). This effect is also known as the "cry-wolf" phenomenon (Breznik, 1983) that can result in longer response times to alarms (Getty et al., 1995) as well as in lower response frequencies (Bliss et al., 1995). These findings are particularly relevant since a low PPV can result even in alarm systems with a high sensitivity (d') when the base rate of critical events is low (Parasuraman, Hancock & Olofinboba, 1997).

Referring to common theories of trust Bliss (2003) argues that the ignorance of alarms might be taken as a perspicuous indicator for alarm mistrust. Furthermore it has been supposed that the relationship between system properties and operators' responses is not a direct one but mediated by the trust of operators in the alarm systems (e.g. Lees & Lee, 2007; Madhavan et al., 2006).

However, the role trust plays in this context is not as clear. Whereas some studies have provided evidence for a close relationship of trust in alarms and operators' responses (e.g. Madhavan et al., 2006; Lees & Lee, 2007), others have challenged this assumption. For example Bustamante (2009) used structural equation models in order to analyze whether the influence of alarm characteristics on operators' responses was mediated by trust. In his study he did not find much evidence for this effect. This suggests that operators' behavior is based mainly on rational considerations rather than on trust. Certain behaviors can provide a benefit for the operators'

work. To ignore alarms that are expected to be false offers more time for concurrent tasks. This reduced compliance may not be due to mistrust, but is the consequence of rational consideration and may represent a logical solution given the situational context. Meyer's (2004) theory of the "expected value" describes one possible approach regarding operators' considerations.

The purpose of this research was to examine to what extent the operators' behavior towards alarms is predicted by their perceived alarm reliability and whether this effect is mediated by their subjective trust towards the system. In order to provide situational context for the described "cry-wolf" effect an alarm system was used that did not allow a cross-check of alarm trial validity towards other available system information.

METHOD

Participants

56 participants (27 females, 29 males, mean age: 26.98 years) were randomly assigned to one of five conditions. The group sizes ranged from ten to twelve participants. The participants received a basic payment of € 7 for their participation and a bonus payment of maximal € 8 depending on their performance during the experiment. On average, they received € 13.10.

Task Environment

The PC-based laboratory environment M-TOPS 2 (Multi-Task Operator Performance Simulation 2) was used for the experiment. The different tasks simulated within this paradigm require cognitive demands which are typical for operators' work in chemical plants. A maximum of three tasks has to be performed simultaneously, i.e. an Ordering Task, a Monitoring Task and a Refilling Task. The interface of M-TOPS 2 is shown in Figure 1. For the current experiment, only two tasks were used, namely the Ordering Task (upper left side) and the Monitoring Task (lower right side).Participants were asked to ignore the Refilling Task (upper right side; for a detailed description of this task see Domeinski, Wagner, Schoebel & Manzey, 2008).

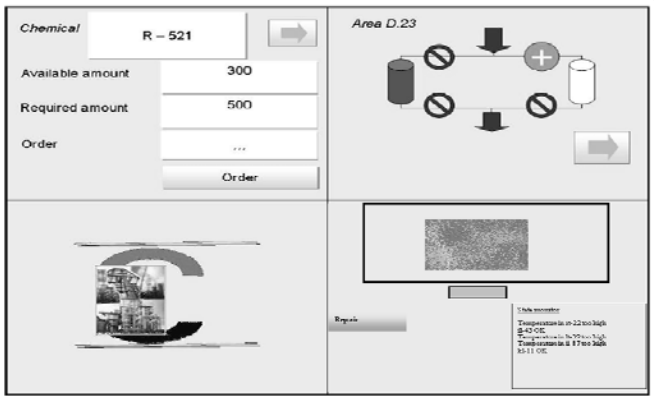

Figure 1. User interface of M-TOPS 2.

Ordering Task. The purpose of this task is to order chemicals which are needed to keep the chemical process running. Participants have to compare the required amount of a given chemical with the available amount, calculate the difference, and sent an appropriate order by clicking a button. A new ordering task shows up after 3 seconds. For every chemical, participants have 15 seconds to fulfill the demand. When participants fail to finish the task within the given time frame, the next task comes up automatically. Performance in this task is assessed by the number of correct orders sent.

Monitoring Task. The purpose of this task is to control the containers filled with the end product of the chemical process. Specifically, the participants need to check the appropriateness of the molecular weight of the content before it will be delivered to the customer. This task is supported by an automatic alarm system. Whenever a certain container arrives at the control station it gets automatically checked and the feedback of this process is signaled to the operator by a green or red light. A red light indicates an alarm pointing to an inappropriate molecular weight. A green light suggests that the content is in a proper state. In case of an alarm an additional message appears on the alarm state monitor which displays the automatically generated diagnosis for the container (i.e. "molecular weight too high"). When the alarm is not activated (green light) the message "container ok" is presented. For each event, the participants have to decide whether or not to initiate a corrective action. This can be accomplished by clicking the "repair" button. However, due to the overall complexity of the process the operators do not have access to any other information sources which would allow them to cross-check the validity of a given alarm. Therefore, responses need to be solely based on the experience of the participants with the alarm system and their knowledge about the system's reliability.

The system requires for controlling containers with a frequency of 7.5 containers per minute. For each single container the response of the operator is logged.

Procedure

Participants completed the experiment in groups of two to five people. When they arrived at the laboratory, they filled out a demographic questionnaire and then began reading the instructions. Both, the Ordering Task and the Monitoring Task, were explained and participants were familiarized with both tasks by practicing each of them separately for two minutes. This initial task training was followed by another practice block of 10-15 minutes which was used to familiarize the participants more deeply with the reliability of the alarm system (i.e. probability of misses and false alarms). During this task the participants had to work on the Monitoring Task only. A total of 100 containers was presented, and an auditory feedback was provided whenever the participants' response to a given alarm or its absence was wrong (i.e. repairing the container in case of a false alarm or ignoring it in case of a miss). This practice block allowed the participants to build a mental model of the system's alarm reliability. After the practice block, participants were asked to assess their trust

towards the system and the perceived reliability. Subsequently, participants completed two experimental blocks of 13 minutes each during which both tasks had to be performed concurrently. Participants were instructed to perform both tasks as best as possible in order to maximize their performance. They received 1.5 points for every correct order of chemicals. For the Monitoring Task, they collected 2 points for a correct response (i.e. repairing containers with a too high molecular weight and ignoring the ones which were ok). For a wrong response (i.e. ignoring damaged containers or repairing them without need), they lost 2 points. This payoff is based on an analysis of the time structure of both tasks and was chosen to make sure that the participants perceive the two tasks as being concurrent while having the same importance.

No on-line performance feedback was provided during these blocks. However, after completion of each 13-min block, participants were provided a visual feedback about their overall performance and they were asked to assess their trust in the automated alarm system. As the first experimental block was needed to gain more experience with the task and to develop stable working strategies, only the data from the second block and the trust measurement prior to this block were considered for statistical analyses.

Experimental Design

Five alarm systems without trial wise validity information were used. The systems differed with respect to the alarm reliability (i.e. PPV). The five different levels were: 0.10; 0.30; 0.50; 0.70; 0.90.

The experimental conditions simulated the use of an alarm system with given technical properties (i.e. d'=1.09, hit-rate=0.80 and false alarm-rate=0.4) in different settings where the base rates differed from 0.05 to 0.8. However, changes in the base rate do not only result in different PPVs but do also influence the NPV. The NPV should impact the responses of operators to no-alarm events ("green light"). However, these effects are not considered in this paper.

Measures

Behavior. For the Monitoring Task the response-rate to alarms was analyzed as the percentage of responses (clicking the repair button) across all alarms. To assess the performance in the Ordering Task the number of correct orders was used.

Perceived system reliability. After the practice block, the participants had to estimate the reliability of the system. In a 2x2 matrix they filled in the number of correct and wrong diagnoses (alarm or absence of an alarm) that they had experienced. The subjective estimation was used to calculate the perceived alarm reliability (PPV), comparing the number of correct alarms to all given alarms.

Trust. Three different aspects of trust were assessed: the trust towards the entire system, the trust regarding the given alarms and the trust regarding the non alarm trials. Participants had to indicate their extent of subjective trust on a 20 cm scale containing no dimension units but verbal orientation instructions ranging from "my trust is very strong" to "I barely trust the system". For the purpose of this study only the trust regarding the alarm trials was used. The trust ratings were assessed in cm and transformed onto trust dimensions ranging from 0 to 1.

RESULTS

The effects of the different alarm reliabilities on the mean response-rate, the rating of perceived reliability, and the subjective trust rating are shown in Figure 2. A one-way ANOVA with "Alarm Reliability" defined as within-subjects factor revealed a main effect for all three variables, i.e. response-rate, $F(4,51)=20{,}44$, $p<.001$;trust rating, $F(4,51)=3.42$, $p=.015$, and perceived reliability, $F(4,51)=18{,}52$, $p<.001$. However, the course of effects is different for the three variables.

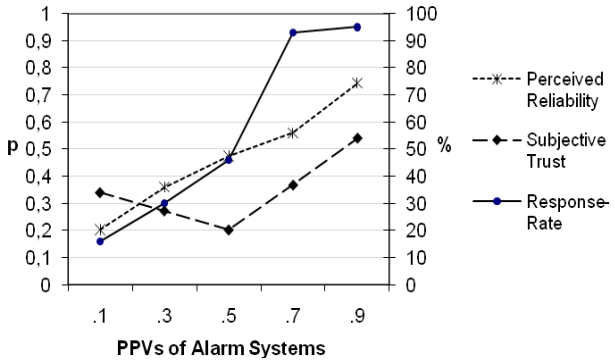

Figure 2. Perceived Reliability, Subjective Trust and Response-Rates for the five different PPVs.

As it becomes evident from Figure 2 the response-rate to alarms increased with the increasing alarm reliability. The lowest response-rate was found for the 0.10 reliability condition and the highest for the 0.70 and 0.90 conditions. A sharp increase of the response-rate occurred between the 0.50 and the 0.70 condition.

In contrast the rating of trust in the alarms showed a sort of U-shaped function with the lowest rating for the medium level of alarm reliability (0.50) which represents the most ambiguous condition. In contrast the trust ratings were higher for both, conditions with lower as well as higher levels of reliability. However, the observed effects on trust were considerably higher for the higher reliability levels, i.e. 0.70 and 0.90 than for the lower ones.

As expected, the rating of the perceived reliability shows the closest association with the real alarm reliabilities. This is reflected in an almost linear increase of perceived reliability across the different experimental conditions. However, the subjective estimations of the alarm reliability levels were not perfect. Specifically, a systematic bias of under- and overestimation, respectively, was found for the extreme levels of alarm reliability. This effect is shown in

Figure 3. Separate t-tests contrasting real and estimated reliabilities for each group revealed that the real alarm reliability was underestimated in the 0.90, $t(9)=-4.12$, $p=.004$, and the 0.70 conditions, $t(10)=-2.40$, $p=.038$. No significant differences were found for the 0.50, $t(11)= -0.49$, $p=.632$, and the 0.30 conditions, $t(10) = 1.46$, $p=.163$. A significant overestimation emerged for the 0.10 condition, $t(11)=2.35$, p=.038.

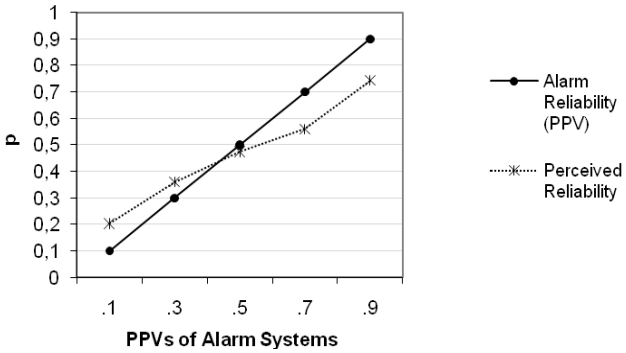

Figure 3. Perceived Reliability compared to the Alarm Reliability (PPV).

In order to explore the relationships between the perceived reliability, trust and response-rate in more detail, the bivariate correlations among all three variables were calculated across subjects. These analyses revealed significant relationships between all three variables with $r=.65$, p<.001, for perceived reliability and response-rate, $r=.30$, $p=.026$, for perceived reliability and trust, $r =0.35$, $p < .01$, and for trust and response-rate. Furthermore a mediator analysis (Baron & Kelly, 1986) was conducted in order to investigate to what extent the regression of the response-rate on the perceived reliability was mediated by trust. The results are shown in Figure 4. As it becomes evident, the response-rate to alarms seems to be directly dependent on the perceived alarm reliability level with no mediating effect of the expressed trust in alarms.

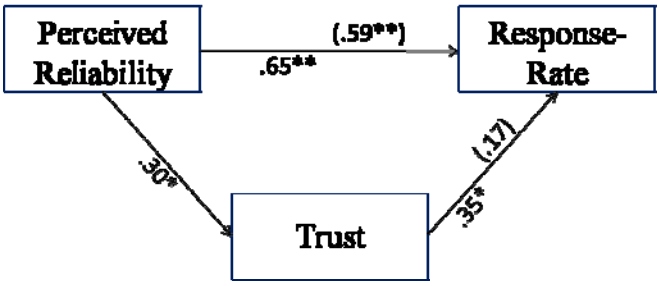

Figure 4. Results of mediator analysis with Perceived Reliability as independent variable, Trust as potential mediator variable and Response-Rate as dependent variable. Regression weights in brackets correspond to results of multiple regression; *p<0.05; **p<0.01.

DISCUSSION

The results of the current research suggest that operators' response frequencies to alarms are mainly guided by the perceived alarm reliability in terms of PPV. As was revealed by the close correspondence between the real and the subjectively perceived reliability of the alarms, the participants were able to build up a more or less correct mental model of the alarm reliability, even though they tended to underestimate high reliabilities and overestimate the lower ones. This perceived reliability was found to predict participants' response-rates on alarms. This result confirms earlier findings which also point to a close association between operators' behavior and perceived system properties (Bliss et al., 1995; Getty et al., 1995; Meyer, 2001; Bustamante, Anderson & Bliss, 2004; Gérard & Manzey, 2009).

The trust ratings did not seem to mediate the effect of system properties on response behavior even though the perceived reliabilities determined participants trust in the alarm system. These findings are in line with previous research by Bustamante (2009). This somewhat surprising effect might be related to the specific situational context which was used in the present study; i.e. participants had no possibility to actively verify a given alarm towards other available data. As Lee & Moray (1994) point out in their multidimensional theory of trust in automation, the operators' understanding of how the system works usually provides an important contribution to their level of trust towards the system. However, the development of such understanding requires that the operator can compare the decisions of an alarm system with the raw data the system's diagnose is build on. If the situation does not offer any possibility for such a sense making process the subjective trust cannot include this "process" dimension of trust. Hence, the observed trust ratings may represent mainly another dimension of trust. The low trust ratings for the 0.50 reliability system suggests that the trust ratings in the current study mainly reflect the predictability of alarm in the different experimental conditions. Predictability of an automated system has been supposed to represent an important aspect of trust in automation (Muir, 1994).

Lacking the possibility to actively verify a given alarm, operators' behavior might be more guided by rational considerations about the most efficient behavior than by their expressed trust in the system. The results show a high compliance (i.e. a response-rate over 90%) for the system with the 0.90 alarm reliability as well as for that one with the 0.70 reliability. For the lower reliabilities the response-rate corresponds nearly to the perceived reliabilities representing a kind of probability matching behavior. This suggests the existence of an internal criterion towards the effectiveness of the alarm system. If the alarm reliability is high enough to provide a sufficient benefit in terms of operator hit-rate, the operator decides to respond to nearly every alarm. If the alarm reliability decreases under a certain criterion the perceived costs in terms of time requirements become too high. In consequence the operator decides to focus on the parallel task.

Supposedly, operators try to maximize there expected value as proposed by Meyer (2004).

A limitation of the present study relates to the one-dimensional assessment of trust which does not allow further analysis of the proposed relationships among system properties, context of use and different dimensions of trust. In order to explore in more detail how the different dimensions of trust are affected by the properties of alarm systems and the contextual factors, a multi-dimensional assessment of trust needs to be executed. Furthermore it must be taken in consideration that subjective ratings of trust can always be biased to some extent.

Overall, the results of the present study suggest that the relationship between properties of alarm systems, operator trust and operator response behavior is complex and might be affected by the context of use. More research will be needed to elucidate these relationships further.

REFERENCES

Bliss, J. P. (2003). An Investigation of Extreme Alarm Response Patterns in Laboratory Experiments. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, 5, 1683-1687.

Bliss, J. P., Gilson, R. & Deaton, J. (1995). Human Probability Matching Behavior in Response to Alarms of Varying Reliability. *Ergonomics*, 38 (11), 2300-2313.

Breznitz, S. (1983). *Cry wolf: The psychology of false alarms.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bustamante, E. A. (2009). A Reexamination of the Mediating Effect of Trust among Alarm Systems' Characteristics and Human Compliance and Reliance. *Proceedings of the Human Factors and Ergonomics Society 53th Annual Meeting*, 249-253.

Bustamante, E. A., Anderson, B. L. & Bliss, J. P. (2004). Effects of Varying the Threshold of Alarm systems and Task Complexity on Human Performance and Perceived Workload. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 1948-1952.

Bustamante, E. A., Anderson, B. L. & Bliss, J. P. (2004). Effects of Varying the Threshold of Alarm systems and Task Complexity on Human Performance and Perceived Workload. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 1948-1952.

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.

Gérard, N. & Manzey, D. (2009). Are false alarms not as bad as supposed after all? A study investigating operators' checking behaviour in response to imperfect alarms. *Proceedings of the Europe Chapter of the Human Factors and Ergonomics Society Annual Meeting 2009*, Linköping, Sweden.

Getty, D.J., Swets, A., Pickett, R.M. & Gonthier, D. (1995). System Operator Response to Warning of Danger: A Laboratory Investigation of the Effects of the Predictive Value of a Warning on Human Response Time. *Journal of Experimental Psychology:Applied*, 1 (1), 19-33.

Kantowitz, B. H., Hanowski, R. J. & Kantowitz, S. C. (1997). Driver Acceptance of Unreliable Traffic Information in Familiar and Unfamiliar Settings. *Human Factors*, 39(2), 164-176.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35 (10), 1243-1270.

Lee, J.D. & See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46 (1), 50-80.

Lees , M. N. & Lee, J.D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, 50 (8), 1264-1286.

Madhavan, P., Wiegmann, D.A. & Lacson, F.C. (2006). Automation Failures on Task Easily Performed by Operatores Undermine Trust in Automated Aids. *Human Factors*, 48 (2), 241-256.

Meyer, J. (2004). Conceptual Issues in the Study of Dynamic Hazard Warnings. *Human Factors*, 46 (2), 196-204.

Meyer, J. & Bitan, Y. (2002). Why Better Operators Receive Worse Warnings. *Human Factors*, 44 (3), 343-353.

Muir, B. M. (1994). Trust in automation: Part 1. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37 (11), 1905-1922.

Parasuraman, R., Hancock, P.A. & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics*, 40 (3), S. 390-399.

Rice, S.R. (2009). Examining Single- and Multiple-Process Theories of Trust in Automation. *The Journal of General Psychology*, 136 (3), 303-319.

Stanton, N. S., Ragsdale, S.A. &Bustamante, E.A. (2009). The Effects of System Technology and Probability Type on Trust, Compliance, and Reliance. *Proceedings of the Human Factors and Ergonomics Society 53th Annual Meeting,1368-1372.*

Wiegmann, D. A. (2002). Agreeing with Automated Diagnostic Aids: A Study of Users' Concurrence Strategies. *Human Factors*, 44 (1), 44-50.