Dietrich Manzey, Nina Gérard, Rebecca Wiczorek

Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload

Journal article | **Accepted manuscript (Postprint)**
This version is available at https://doi.org/10.14279/depositonce-8628

This is an Accepted Manuscript of an article published by Taylor & Francis in ERGONOMICS on 16 Sep 2014, available online: http://www.tandfonline.com/10.1080/00140139.2014.957732.

Manzey, D., Gérard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. Ergonomics, 57(12), 1833–1855. https://doi.org/10.1080/00140139.2014.957732

Decision-making and response strategies in interaction with alarms: The impact of alarm reliability, availability of alarm validity information, and workload

Dietrich Manzey, Nina Gérard, Rebecca Wiczorek

Institute of Psychology and Ergonomics, Technische Universitaet Berlin, Berlin, Germany

Corresponding Author: Dietrich Manzey, TU Berlin, Marchstr. 12, F7, D- 10587 Berlin, phone +49 30 314 – 21340, fax +49 30 314 – 25434, email: dietrich.manzey@tu-berlin.de

Nina Gérard, TU Berlin Marchstr. 12, F7, D-10587 Berlin, Germany, phone +49 30 314– 25275, fax +49 30 314 – 25434, email: nirard@web.de.

Rebecca Wiczorek,TU Berlin Marchstr. 12, F7, D-10587 Berlin, Germany, phone +49 30 314– 25275, fax +49 30 314 – 25434, email: wiczorek@ tu-berlin.de

Abstract

Responding to alarm systems which usually commit a number of false alarms and/or misses involves decision-making under uncertainty. Four laboratory experiments including a total of n=256 participants were conducted to gain comprehensive insight into humans' dealing with this uncertainty. Specifically, it was investigated how responses to alarms/nonalarms are affected by the predictive validities of these events, and to what extent response strategies depend on whether or not the validity of alarms/nonalarms can be cross-checked against other data. Among others, the results suggest that, without cross-check possibility (experiment 1), low levels of predictive validity of alarms (≤0.5) led most participants to use one of two different strategies which both involved non-responding to a significant number of alarms (cry-wolf effect). Yet, providing access to alarm validity information reduced this effect dramatically (experiment 2). This latter result emerged independent of the effort needed for cross-checkings of alarms (experiment 3), but was affected by the workload imposed by concurrent tasks (experiment 4). Theoretical and practical consequences of these results for decision-making and response selection in interaction with alarm systems, as well as the design of effective alarm systems are discussed.

Keywords: decision-making, alarm system, compliance, reliance, cry-wolf effect

Practitioner Summary

Four laboratory experiments were performed to investigate the effects of false alarms and misses on the behavioural effectiveness of alarm-systems. The results provide insight in determinants of compliance with alarms, dependent on the alarm-system's reliability, the possibility to cross-check the validity of alarms, and the workload imposed by concurrent tasks.

1. Introduction

Alarm systems represent an important element of almost every complex human-machine system. Usually they are implemented to support human operators in detecting critical system states or events which require some intervention. The most basic forms of these systems represent binary decision aids which remain *silent* (or show a *green light*) as long as some assessed parameters are within a pre-defined target range of nominal operation, and provide an alerting signal (e.g. a loud tone or *red light*) in case of deviations. Due to inherent technical limitation and noisy data from the environment, binary alarm systems are never perfectly reliable. That is, they can miss critical events or produce false alarms. This represents a challenge for human operators who usually have the final responsibility for the overall safety of a process and needs to decide whether and how to respond to a given output of an alarm system. This raises a number of issues related to how operators' responses to alarms are affected by the perceived properties of the system, the experiences made with it, or the situational context (e.g. workload). For example, it has been shown, that humans who have made the repeated experience of false alarms start to mistrust the alarm system and tend to respond slower to alarms (Getty et al. 1995; Wickens and Colcombe 2007) or even to ignore future alarms completely (Bliss, Gilson, and Deaton 1995; Lees and Lee 2007). The practical significance of proper dealing with imperfect alarm systems is reflected by analyses suggesting that a significant number of accidents and incidents in process control or aviation can be traced back to inappropriate responses to alarms (Bliss and Fallon 2006; Bliss 2003a; Bransby and Jenkinson 1998). The current research includes a series of four laboratory experiments which were performed to gain a better understanding of human decision-making in interaction with alarms and its relationship to characteristics of the alarm system as well as to the situational context. In accordance with others, e.g., Sorkin and Woods (1985) and Meyer (2004), we propose that human interaction with alarm systems be considered an issue of decision-making under uncertainty. The following questions are addressed: To what extent depend human responses to alarms on the predictive value of alarms and what strategies and heuristics are applied to deal with uncertainty in this case? How are response strategies modified if the uncertainty of single alarms can be reduced by cross-checking their validity towards other information? What significance have context factors like effort of alarm verification or overall workload on decision-making and responses to alarms?

1.1. Background: Properties of Alarm Systems

Any binary alarm system represents an automated decision aid which classifies states of a process or the environment into two categories: normal vs. critical. The performance of such system can be described by its hit rate (pH), defined as proportion of correctly indicated critical events (hits) out of all events being in fact critical, and its false alarm rate (pFA), i.e. proportion of falsely emitted alarms out of all non-critical events. Formally, the basic properties of a system leading to this performance can be modelled by means of signal detection theory (SDT, Green and Swets 1966), i.e. in terms of its sensitivity (d') and response criterion (c). This has been described in detail elsewhere (cf. Meyer and Bitan 2002; Parasuraman, Hancock, and Olofinboba 1997; Sorkin and Woods 1985) and shall not be repeated here.

More important for an understanding of the behavioral efficiency of alarm systems are two other properties which have been referred to as the *positive* and *negative predictive value* (PPV and NPV; Getty et al. 1995; Meyer and Bitan, 2002). The PPV represents the conditional probability that there really is a critical event (E) in case of an alarm (A), which directly depends on the number of hits and false alarms produced by a system. Formally, it is defined as follows:

$$PPV = p(E \mid A) = \frac{hits}{hits + false\ alarms}$$

The NPV represents the conditional probability that there is no critical event (nonE) in case that no alarm (nonA) has emitted and is defined as follows.

$$NPV = p(nonE \mid nonA) = \frac{correct\ rejections}{correct\ rejections + misses}$$

Considering the PPV and NPV is especially important because it can be assumed that operators, based on their experience, may gain a more or less proper picture of the PPV and NPV of their alarms systems but not the underlying technical properties like pH and pFA. Notable in the current context is that the PPV and the NPV do not only depend on the technological features of a given alarm system alone, but also on the *a priori* probability or base rate of critical events (Getty et al. 1995; Meyer and Bitan 2002; Parasuraman et al. 1997). Specifically, the PPV of an alarm decreases significantly with decreasing base rates. For the NPV the opposite is true: with decreasing base rates NPV tends to increase. This has two important consequences: First, even alarm systems with high hit and low false alarm rates can have a very low PPV in case that the *base rate* of a failure is low (Meyer and Bitan 2002; Parasuraman et al. 1997). Second, PPV and NPV represent two distinct aspects of predictive validity of alarm systems which are negatively correlated.

1.2. Background: Responding to Alarms as Decision-Making under Uncertainty

Basically, working with alarm systems can be considered as a decision-making task, i.e. operators need to decide whether or not to become active and respond in a proper way in case an alarm has been triggered, or whether or not to remain passive in case of nonalarms. The specific challenges of making these decisions depend directly on whether or not alarm validity information (AVI) is available that can be accessed by human operators in order to cross-check the output of the alarm system (Sorkin and Woods 1985).

Let us first consider the situation where AVI is not available. Examples for this situation include remote smoke detectors or an in-flight icing alert in the cockpit. The decision to be made in this case typically represents a decision under uncertainty with the degree of uncertainty directly depending on the PPV and NPV of the alarm system (Meyer 2004). In accordance with these two properties, two different behavioral aspects have been distinguished and referred to as *compliance* and *reliance* (Meyer 2001, 2004). Compliance refers to the extent operators respond to a given alarm in accordance with the alarm, i.e. by initiating all actions necessary in case of a critical event or malfunction. Reliance refers to the extent to which operators refrain from any action unless the alarm goes off. Theoretically, rational compliance and reliance decisions can be modelled based on expected values considerations assuming that operators have detailed knowledge about the PPV and NPV as well as about the consequences (benefits and costs) of correct or false decisions (Meyer 2004). Operators following such rational considerations would be expected to apply what has been coined an extreme responding strategy in reponse to alarms (Bliss et al. 1995). That is, they would either respond to all alarms in the expected way or ignore all of them, dependent on whether the perceived PPV of the alarm system were higher or lower compared to their response criterion. However, as is known from many other domains, judgments and decisions under uncertainty hardly ever follow pure rational considerations but more or less proper decision heuristics (Gigerenzer and Todd 1999; Kahnemann, Slovic and Tversky 1982). In accordance with this, a response strategy other than extreme responding has also been observed in alarm research and referred to as probability matching (Bliss et al. 1995; Bliss 2003b). Applying this second strategy, operators follow the output of an alarm system in a frequency that corresponds more or less to the described or perceived reliability in terms of PPV and NPV. That is, if they are told (or experience) that about 80% of all alarms will be correct, they tend to follow about 80% of all alarms.

The decision-making process becomes even more complex if operators have the possibility to reduce the level of uncertainty related to a given output of an alarm system by cross-checking its validity with other available information. In this case, operators have three different behavioral alternatives to respond to a given alarm, i.e. they can directly comply with it, ignore it, or cross-check it before deciding how to respond. However, cross-checking usually is time consuming and, thus, involves costs that need to be traded-off against concurrent goals, e.g., to respond as quickly as possible, or to continue performing concurrent tasks which need to be done. As a consequence, it can be assumed that decision-making of operators in this case involves two different sub-stages (Allendoerfer, Pai and Friedman-Berg 2008). Decision-making on the first sub-stage is equivalent to the one described above for the situation without AVI. If the predictive validity of an alarm system's output is perceived as sufficiently high or too low, operators will probably either follow the system directly without investing the effort of own information sampling, or just ignore it. However, in all other cases a second sub-stage of more effortful decision-making will be entered by first cross-checking other information in order to reduce the uncertainty related to the advice of the alarm system before a final response decision is made.

1.3. Prior relevant research

Thus far, research has mainly focused on the compliance component in response to alarms and especially the so called "cry-wolf" effect in response to a high frequency of false alarms (Bresnitz 1984). Several studies support the notion that experiencing a high number of false alarms (indicative of a low PPV) leads operators to respond only slowly to alarms or to ignore alarms completely (e.g. Bliss et al. 1995; Bliss and Dunn 2000; Bliss, Jeans, and Prioux 1996; Bresnitz 1984, Dixon, Wickens, and McCarley 2007; Getty et al. 1995; Maltz and Shinar 2003). First insights in the relationship between PPV and compliance have been provided by studies of Getty et al. (1995) and Bliss et al. (1995). Getty et al. (1995) found immediate and quick responses predominating for PPVs of .39 and beyond, but a clear shift to slow response times reflecting a cry-wolf effect for the lowest PPV they considered (.25). The data of Bliss et al. (1995) provide some evidence for the significance of the *probability matching* heuristic in coping with the uncertainty of alarms. By analysing responses to alarms with PPVs of .25, .50 and .75 they found that the majority of participants chose a sort of probability matching strategy, leading to an increasing proportion of ignored alarms with decreasing PPV. Only a minority of participants (about 10%) applied the more rational extreme responding strategy by either responding to all or no alarms in the .75 and .25 condition, respectively.

Comparatively less is known about the effect of imperfect alarm systems on the reliance component and the relationship of compliance and reliance. According to Meyer (2004) compliance and reliance represent independent aspects of dependence on alarm systems with compliance mainly affected by the PPV (or false alarms) and reliance by the NPV (or misses). Whereas some research support this notion (Dixon and Wickens 2006; Meyer 2001; Meyer and Bitan 2002; Wickens and Colcombe 2007), other results challenge it by providing evidence for non-selective effects of false alarms on complicance as well as reliance (e.g. Dixon et al. 2007; Rice 2009; Meyer, Wiczorek and Guenzler 2013). Another important aspect which has been rarely addressed in previous research is how compliance and reliance of operators in interaction with alarms are affected by the availability of AVI (e.g. Bliss 2003b; Bustamante 2005). Bustamante (2005) analysed to what extent the provision of AVI would improve human decision-making in interaction with an alarm system with a PPV=.18. Not surprinsingly he found that providing AVI contributed to a better discrimination between true and false alarms under these conditions. More interesting results were presented by Bliss (2003b). His review of a set of his own studies suggests that the tendency for extreme responding to alarms with medium to high PPVs depends on the availability of AVI. Whereas only low rates of extreme responding (0-12%) were observed in studies where AVI was available, these rates increased up to 52% in situations without AVI. However, only conditions with alarm PPVs of 0.50 to 0.75 were included in this research which raises the question to what extent these results also apply to alarms with PPVs<0.50 which are much more common in real field settings.

1.4. Current Research

The main objective of the current research is to investigate how decision-making in response to alarms is affected by the PPV, and, even more important, to what extent this decision-making and response strategies change if AVI is available. The latter aspect is of considerable practical relevance. If it was shown, for example, that known biases and issues related to the uncertainty of alarms, e.g. the “cry wolf” effect, could be reduced by providing AVI, this would have direct implications for the design of effective alarm systems. Similar to other research (e.g. Bliss et al. 1995; Dixon et al. 2007), the alarm task used for the present research was part of a multi-task environment. This made it possible not only to evaluate the performance consequences of interacting with alarm systems of varying reliability on alarm decision-making itself, but also on concurrent task performance. However, in contrast to prior research a broader range of PPVs was considered, including extreme low and high PPVs. Furthermore, an alarm system was used that did not only produce false alarms but also misses, albeit to a considerably lesser degree. This made it possible to not only evaluate effects of different PPVs on responses to given alarms but also to explore possible effects of different NPVs on the reliance component at the same time with the same system.

A series of four experiments was conducted using essentially the same multi-task paradigm. Participants had to perform two or three concurrent tasks, including one quality control task which was supported by an alarm system. The PPV of alarms was varied systematically within each experiment by varying the base rate of critical events. More specifically, five different conditions were compared in each experiment with PPVs of alarms of 0.1, 0.3, 0.5, 0.7 and 0.9. Note that the variation of base rates producing these PPVs also involved varying NPVs for nonalarm events which differed between 0.98 and 0.41.

The first experiment addressed the impact of the different PPVs on responses of operators in a situation where no AVI was available. It was expected that, in this case, participants would base their decision about how to respond to an alarm directly on the PPV. More specifically, it was hypothesized that the majority of participants would use a sort of probability matching strategy for PPVs ranging from 0.3 to 0.7, and only a minority would exhibit a positive or negative extreme responding. This would be in accordance with the earlier findings of Bliss et al. (1995) and Bliss (2003a). However, with respect to the more extreme conditions (0.1 and 0.9) extreme responding becomes a specific case of probability matching. Therefore it was expected that in these conditions extreme responding would be the dominant response pattern, i.e. participants were expected to comply with every alarm in conditions 0.9, and to ignore every alarm, i.e. exhibit an extreme “cry wolf” effect, in conditions where PPV=0.1.

The second experiment replicated the first one in most aspects but offered the availability of AVI. Participants were able to cross-check the validity of a given alarm (or nonalarm) event against raw data. Yet, checking these data needed some time that had to be subtracted from another (concurrent) task. Thus, participants had to trade-off the possibility to reduce uncertainty to the possibility to work on the concurrent task. It was expected that in this case using the cross-check option in response to given alarms would follow an inverted-U shape function across different levels of PPV. Alarms would be cross-checked most often in case of maximum uncertainty (PPV=0.5) and least often in case of least uncertainty, independent of whether the PPV is extremely high (0.9) or low (0.1).[1] For events with least uncertainty, an extreme responding strategy was expected to remain the dominant behavior.

The third and fourth experiments capitalized on the second one and investigated how the pattern of effects found in the second experiment would be altered in case of raised effort needed for cross-checks (Exp. 3) or overall workload imposed by the addition of another concurrent tasks (Exp. 4). It was expected that these interventions would reduce the number of cross-checks and raise the probability of direct responding to or ignoring of alarms in conditions with PPV>0.5 and PPV<0.5, respectively.

Although the impact of the PPV on responses to alarms was the main focus of research, we also looked at possible effects of NPV on the reliance component, and to what extent participants distinguished between both aspects in their responses to alarms and nonalarms. Based on the original reasoning of Meyer (2004) it was assumed that participants would clearly discriminate between the PPV and NPV of the different outputs of the alarm system and that their responses to nonalarms would be mainly guided by the NPV of these events. In addition, we looked for any indications of cross-effects of false alarms and misses on the reliance and compliance component, respectively. However, since the design of the experiments was not specifically tailored to a detection of such effects, this part of research remained explorative.

2. Experiment 1

2.1. Method

2.1.1. Participants

[1] Note that alarm systems with PPVs of 0.1. and 0.9, albeit representing very different systems in terms of reliability are equivalent with respect to the level of uncertainty associated with alarms. Humans interacting with these systems can be sure that 90% of alarms are either wrong or correct, respectively, i.e. they face the same low uncertainty in making their decision how to respond.

56 students (27 females, 29 males, mean age: 26.98 years) participated in the experiment. Participants were screened not to suffer from any distortion of color vision which might interfere with the experiment (i.e. red-green color blindness). They were randomly assigned to the experimental conditions and received a basic payment of € 7 for their participation plus a bonus payment of maximal € 8 depending on their performance during the experiment.

2.1.2. Task

A PC-based laboratory task, M-TOPS-A (*M*ulti-*T*ask *O*perator *P*erformance *S*imulation for *A*larm Research), was used for the experiment. M-TOPS-A is a multi-task environment including up to three different tasks which have to be performed simultaneously. The tasks are chosen to require basic cognitive demands similar to those required from operators in the control room of a chemical plant, namely a *Resource Ordering Task*, a *Coolant Exchange Task,* and an *Alarm Task*. The latter task represents the most important task in the present context. The interface of M-TOPS-A is shown in Figure 1. The different tasks are described in the following:

Insert Figure 1 about here

Resource Ordering Task (ROT). This task represents a mental arithmetic task which is shown in the upper left quadrant of the interface. Participants are instructed that they always have to assure the availability of required chemicals in order to keep the chemical process running. For this purpose, the actual and the set value of an ascertained chemical is presented. Participants then have to calculate the difference, type the result in the designated ordering field, and initiate the order by mouse-clicking on the order button. Participants have 15 seconds to respond to a given request. After an order has been sent, a new task is presented after a fixed interval of three seconds. Alternatively, the participant can actively initiate a new task by clicking on the arrow button. As performance measure, the number of correctly sent orders per given time period is sampled.

Coolant Exchange Task (CET). This task demands the execution of different actions in a given sequence. According to the instruction these actions are needed to exchange the coolant in different sub-systems of the plant. For this purpose, different sets of two coolant tanks each are shown in the upper right part of the interface. In order to exchange the used coolant with new one, participants have to open and close the different valves of the sub-system in a predefined sequence. Due to inherent time-constants, a complete exchange-cycle takes a minimum of 40 seconds plus the time-delays produced by participants at different steps of the action sequence. After each completed exchange of coolant indicated by a change of the color of the tank from green (used coolant) to grey (empty) to blue (fresh coolant), participants have to activate the sub-system by a mouse-click on a button and a new set of tanks representing another sub-systems is displayed. As a performance measure the number of completed replacements of coolant within a given time period is counted.

Alarm Task (AT). This task demands decision-making in interaction with an alarm system. The interface for this task is shown in the lower right quadrant of Figure 1. Participants are instructed that the plant has a control system which automatically monitors and assesses the quality of the chemical end-product filled in single containers. In case the quality of a given container is assessed to fulfill a predefined quality criterion (“molecular weight ok”), a green light is emitted indicating that the state of the product is approved. However, in case the system detects an impaired product (“molecular weight too high”) a visual alarm (red light) goes off, combined with a message displayed on an alarm state monitor. Upon getting the output of the alarm system, the operator has to decide how to respond to it, i.e. by letting the container pass (staying passive) or by clicking the *rework* button. Note that the former is the suggested response to nonalarms and the latter the suggested response to alarms. Upon clicking the rework button, the relevant parameter (“temperature”) can be chosen by another mouse-click from an one-item menu. Because the alarm system is not perfectly reliable and the validity of given alarms cannot be double-checked with other data, this involves a typical decision under uncertainty, comparable to what is often the case in real-world settings.

For the present series of experiments the hit rate of the alarm system was defined as 0.8 and the false alarm rate was 0.4. In terms of signal-detection theory, these characteristics correspond to a sensitivity of d’=1.09 and a comparatively liberal response bias c = -0.29. That is, given low base rates of critical events (p<0.5), the system is clearly *false alarm prone* (like most real systems) but also produces at least some misses, depending on the specific base rate of critical events. Only with high base rates (>0.8) the system eventually turns into a miss prone system which produces more misses than false alarms. This characteristic makes it possible to study both, effects of the system on the compliance as well as on the reliance of the operator.

Containers passed the control station with a frequency of about six containers per minute. After entering the control station the containers remained there for five seconds. During this time window participants could decide whether or not to intervene in response to the output of the alarm system. For each container and output of the alarm system (“green” vs. “red”) the kind of response of the participant (*no response* vs. *rework*) was logged. In addition, it was recorded whether or not the response of the participant was correct. Correct decisions in response to the output of the alarm system include “no response” in case of correct rejections and false alarms emitted by the system, and “clicking on the rework icon” in case of hits or misses of the system, respectively.

2.1.3. Design

For experiment #1, only two of the three tasks had to be performed concurrently, namely the ROT and AT. The experimental design included two factors. The first factor (alarm reliability) was defined as a between-subject factor and included five levels corresponding to five different PPVs of the alarms emitted by the alarm system in the AT. The levels included PPVs equaling 0.1, 0.3, 0.5, 0.7 and 0.9 and were operationally established by varying the base rates of critical events to be detected by the alarm system. Note that this variation of base rates also affected the NPV of alarms, albeit to a much lesser extent. The full information about the different experimental conditions defining the first experimental factor in terms of number of participants in each group, base rates of critical events and basic characteristics of the alarm system (i.e. d', c, PPV and NPV) are provided in Table 1.

Insert Table 1 about here

The second factor included a two-level within-subjects factor capturing possible time-on-task effects on the development of response strategies. Participants had to perform two blocks of trials. During each block a total of 100 decisions had to be made in response to the outputs of the alarm system in the AT.

2.1.4. Dependent Variables

Three sets of data were sampled to analyse response strategies and their performance consequences in the alarm task: (1) *Response rates*, defined as the proportion of *direct responses* (activating the menu for selecting the impaired container and clicking on rework button) to given alarms ("red light") and nonalarms ("green lights"). Note that direct response to a nonalarm is indicative of a participant not relying on the system's output. (2) *Individual response strategies*. In order to describe the type of individual response strategies chosen in response to alarms and nonalarms, we looked at the number of participants who used an extreme strategy in response to given alarms and nonalarms. A response strategy was considered “extreme” if 90% or more of given outputs were either completely ignored (no response; “negative extreme responding”) or directly responded to (“positive extreme responding”). Any other response frequencies which represented a mix of alarms ignored and responded to are referred to as “mixed strategy” in the following. (3) *Overall performance,* defined as the overall number of correct decisions (human and alarm system together) in response to alarms (and nonalarms). A decision in response to a given alarm was considered to be correct if the operator responded to a hit of the alarm system with the expected response but did not respond to a false alarm. In case of nonalarms, decisions were considered to be correct if the operator did not respond to correct rejections of the system but overwrote the system and responded by mouse-clicking on the “rework” button if the system produced a miss.

In addition to these the data sets for AT performance, we also assessed the consequences of interaction with the AT for *concurrent task performance* in the ROT. Performance in this latter task was assessed by numbers of correct orderings sent.

2.1.5. Procedure

The experiment was conducted with groups of up to four participants. After completing a demographic questionnaire and reading the instructions, participants were familiarized with M-TOPS-A. They were told that the experiment was a simulation of a control room of a chemical plant and that their task was to keep the chemical process running while controlling the quality of the end-product simultaneously. After practicing the ROT and AT separately for two minutes each, participants completed a 100-trial training block with the AT only, in order to get familiarized with the characteristics of the alarm system. They were instructed that the alarm system would not work perfectly reliable and that they should use the training block to gain experience with its reliability. Only for that training block an auditory feedback was given that informed participants whether or not their decision in response to the output of the alarm system was correct. The feedback was provided via headphones and included an acoustical signal (“buzzer”) whenever the decision made in response to the output of the alarm system was wrong (i.e. in case that participants responded to a false alarm or did not respond to a green light although it was a miss). After completion of the practice block they provided subjective estimates about the characteristics of the alarm system in terms of numbers of hits, correct rejections, false alarms, and misses per 100 trials, which usually were quite close to the original numbers with slight deviations only in the most extreme PPV conditions. Then participants were explicitly informed about the actual characteristics by providing them a matrix showing the “real” distribution of the different outputs. This combination of providing information about the alarm system by experience as well as by description was done in order to avoid the impact of any decision bias related to only experience-based vs. description-based information (Hertwig and Erev 2009).

The following two experimental blocks again included 100 trials of the AT task and lasted about 13 minutes, each. Hits, CRs, FAs and misses of the alarm system were presented in a random and unpredictable order. Although it cannot be fully excluded, this presentation mode rendered it highly unlikely that the participants could base their guesses on the identification of specific patterns instead of the probability information. With respect to task priority setting during these blocks, participants were instructed that they had to perform both tasks concurrently with the same priority. This was also supported by the pay-off structure defining the number of points which could be “earned” in the different tasks. Participants receive 1.5 points for every correct order of chemicals (ROT). For the alarm task they were granted two points for a correct response but lost two points with each wrong response. This payoff-structure took into account that the two tasks had different time characteristics and should ensure that all tasks were treated as equally important. The financial compensation for participation in the experiments was partially dependent on the overall amount of points gained, i.e. participants could more than double their basic compensation with high scores.

2.1.6. Statistical Analysis

Performance data were analysed by a multifactorial analysis of variance (ANOVA). Given the sample size and the basic characteristics of the experimental design, the assumptions of this approach can be regarded as fulfilled. The F-test would even remain robust against slight deviations from the assumption of normal distribution of data or homogeneity of variances (e.g. Kirk, 1982; Stevens 2007). Post-hoc contrasts between single means were conducted by Scheffé tests which represent the most conservative approach for this purpose.

2.2. Results

2.2.1. Alarm Task: Mean response rates and individual response strategies

Mean response rates to alarms, i.e. average percentage of instances where participants complied with the alarm, are shown in Figure 2 (upper panel left). As becomes evident, mean response rates directly varied dependent on the PPV of alarms, and this general pattern became even more extreme as a result of task practice (block 2 vs. block 1). In block 2 only about 16% of all alarms were responded to in the conditions with the lowest PPV (0.1). In contrast, almost all alarms were complied with in the two conditions with highest alarm PPV (0.7, 0.9). Moreover, a noticeable step increase of response rates occurred between the 0.5 and 0.7 condition. A 5(condition) x 2 (block) ANOVA with repeated measures on the second factor revealed significant main effects of condition, $F(4, 51)= 14.78$, $p< .01$, $\eta_p{}^2= .54$, and a significant interaction effect, $F(4, 51)= 4.61$, $p<.01$, $\eta_p{}^2= .27$.

Insert Figure 2 about here

On first sight this pattern of mean response rates suggests that the participants had used an almost ideal probability matching strategy for responding to alarms, at least for conditions with PPV≤ 0.5. However, a closer look to individual data revealed, that this pattern was actually due to a different mix of response strategies used by the participants in the different conditions. This is illustrated in Figure 2 (medium left panel). The figure shows the number of participants who applied the different response strategies distinguished above in the different experimental conditions. Only data from block 2 were taken into consideration for this analysis. Aggregated across all conditions, the majority of participants (n=35) actually used a sort of extreme responding strategy. Positive extreme responding represented the strategy used by the vast majority of participants in the 0.7 and 0.9 condition. Actually, n=16 (out of 21) participants in these two conditions responded to almost all alarms, and just n=5 participants in the 0.7 condition used a mixed strategy. In contrast, negative extreme responders, who ignored more than 90% of all alarms, only were observed in conditions with an alarm PPV of 0.5 or less, and their number increased with decreasing PPV. Yet, in all of these conditions, also a significant number of participants were found who applied a mixed strategy, and in the 0.5 condition the number of participants using a mixed strategy was even higher than the number of extreme responders. In order to analyze to what extent the mixed strategies corresponded to a probability matching strategy, the individual response rates to alarms in the different conditions were contrasted to the PPV value by simple t-tests. Only the data of the participants using a mixed strategy were included in these analyses. It turned out that only for the 0.3 condition, the individual mixed-strategy response rates (mean: 0.47) did not differ significantly from the PPV, t(4)=1.56, p>.19. For all other conditions, significant differences were found, indicating that the participants tended to over-respond by different degrees to the alarms, with mean response rates of .20 (condition 0.1), t(3)=3.024, p<.06; 0.66 (condition 0.5), t(6)=3.55, p<.02; and 0.84 (condition 0.7), t(3)=11.43, p<.01. Note that similar tendencies of “over responding” were also reported by Bliss et al. (1995).

A different pattern emerged for responses to nonalarms. As becomes evident from Figure 2 (upper right panel) a significant number of direct responses to green lights, indicating that participants did not completely rely on the nonalarm signal, were only observed in the condition with the lowest NPV. In all other conditions the mean response rates were below 10%. The 5(condition) x 2(block) ANOVA revealed a significant effect of condition, $F(4,51)=7.21$, $p<.01$, $\eta_p^2= .36$. Post-hoc Scheffé tests revealed that the 0.41 condition differed significantly from all other conditions (all p<.03). No other pairwise comparisons between other conditions became significant. No main effect block emerged, $F(1,51)=2.18$, $p>.10$, $\eta_p^2= .04$, but a condition x block interaction, $F(4,51)=3.04$, $p<.03$, $\eta_p^2= .19$. The latter indicated that the tendency not to rely on nonalarms in the 0.41 condition even increased in the second block. This overall pattern was largely confirmed by an inspection of the individual data of the second block (Figure 2 medium right panel). All participants in the three conditions with the highest NPVs showed a negative extreme responding, i.e. relied completely on the output of the alarm system and did not intervene if the system did not indicate a critical state. This only changed in the conditions where the NPV decreased below .80. In these conditions a number of participants started to use a mixed strategy and responded to a certain number of green lights as if they would have been an alarm. In the condition with the lowest NPV (0.41) two participants even used a positive extreme responding strategy in the second block indicating a complete breakdown of their reliance on the alarm system's output.

2.2.2. Alarm Task: Overall performance

The number of overall correct decisions in response to alarms as well as nonalarms are displayed in Figure 2 (lower panel). As becomes evident, the number of correct responses to alarms (Figure 2, lower left) differed dependent on the degree of uncertainty associated with an alarm, i.e. showed a U-shaped relationship with a maximum number of correct responses for extreme low and high PPVs and a minimum of correct responses in the condition with a medium PPV of the alarms (0.5). This overall shape was the same for both blocks. Yet, the number of correct responses improved slightly from block #1 to 2. The data were analyzed by a 5(condition) x 2(block) ANOVA with repeated measures for the second factor. This analysis revealed a significant main effect of alarm reliability, $F(4, 51) = 11.58$, $p < .01$, $\eta_p^2 = .48$, and of block, $F(1, 51) = 10.31$, $p < .01$, $\eta_p^2 = .17$. The condition x block interaction did not become significant ($F < 2$). The number of correct decisions in nonalarm trials also increased slightly from block 1 to 2. In addition, it showed an almost perfect linear (inverse) relationship to the NPV of the signal. The 5(Condition) x 2(block) ANOVA revealed a main effect of alarm reliability, $F(4, 51) = 146.83$, $p<.01$, $\eta_p^2 = .92$, as well as a main effect of block that just approached the usual limit of significance, $F(1, 51) = 4.015$, $p=.05$, $\eta_p^2 = .07$.

Insert Figure 3 about here

2.2.3. Concurrent Task Performance

Performance in the ROT is illustrated in Figure 3. The number of ordered resources increased from block 1 to 2, reflecting some effect of practice, $F(1, 51) = 28.04$, $p< .01$, $\eta_p^2 = .36$. In addition, it also depended on the characteristics of the alarm system, $F(4, 51) = 3.84$, $p <. 01$, $\eta_p^2 = .23$. As becomes evident from Figure 3, the number of orders sent was higher in the three conditions with comparatively low PPVs of emitted alarms (≤ .5) compared to the conditions where alarm PPVs were high (0.7; 0.9).

2.3. Discussion

As expected, response frequencies to alarms were mainly guided by the PPV of the alarm system. This is in accordance with earlier findings (Bliss et al. 1995; Getty et al. 1995; Meyer 2001). The results also confirm that, dependent on the PPV, different strategies are used to cope with the uncertainty of the alarms. However, our specific expectation that probability matching would represent the most dominant response strategies for conditions with medium levels of PPVs, i.e. 0.3 to 0.7, and extreme responding only in conditions with extreme high or low PPVs, was only partially supported by the data. PPVs of 0.7 and 0.9 led most participants to apply the (in this case) most rational positive extreme responding strategy by complying with almost every alarm and even the majority of the few participants who applied a mixed strategy in the 0.7 condition still responded to more than 80% of the alarms in the expected way. This confirms assumptions of Bliss (2003b) who reported similar results based on a meta-analyses of response strategies to alarms with PPVs >0.50.

More interesting are the effects for the conditions with alarm PPVs≤0.5 which correspond more closely to real alarm systems. The results of the detailed analyses of the individual response patterns showed that the mean response rates to alarms in these conditions actually resulted from three different underlying strategies. A total of 16 of the 35 participants in these conditions applied a negative extreme responding strategy, i.e. exhibited an extreme cry-wolf effect by ignoring more than 90% of all alarms. This was expected for the 0.1 condition but a significant number of participants showed this behaviour also in the 0.5 condition where 50% of all alarms still were correct. One obvious benefit of ignoring a substantial number of alarms in case that alarm validity information is not available, is the maximizing of overall correct decisions (human and alarm system). Without access to alarm validity information, human operators have no basis other than their general experience with the system to discriminate false and true alarms. As a consequence, they principally cannot increase the number of correct decisions above what would be achieved by positive extreme responding when PPV>0.5 and by negative extreme responding in case of PPV<0.5. Thus, extreme responding can be considered a rational choice under these circumstances. A second beneficial consequence emerged with respect to concurrent task performance. Ignoring of alarms releases resources that can be used to perform concurrent tasks. In the present experiment this effect was directly reflected in a somewhat better concurrent task performance for conditions with PPV≤0.5. Most participants in conditions with PPV≤0.5 obviously decided to ignore at least a significant number of alarms and to focus their attention on the ROT instead. Given this, it is particularly remarkable that another 16 participants in the conditions with PPV≤0.5 did not apply this rational strategy but a sort of probability matching. Actually, the mean response rates of these participants in the different conditions were even a bit higher than the PPVs would suggest. And one participant in each of these conditions even applied a positive extreme responding strategy by complying with every alarm despite the fact that 50%, 70%, and 90% of these alarms were false. This result is in accordance with previous research (Bliss et al. 1995; Bliss 2003b) and suggests that other reasons than pure rational considerations play a role in interacting with alarms. Most likely these participants were hesitant to just ignore alarms and chose probability matching as a kind of compromise to optimize their overall performance in the multi-task setting with, at the same time, acknowledging the significance of alarms.

It has been supposed that the number of false alarms emitted by an alarm system may not only lead to a cry wolf effect and thus reduce the *compliance* with alarms but also affect the level of *reliance* of operators, i.e. the tendency to trust in the system alerting them when a critical event occurs (Dixon et al. 2007; Wiczorek et al. 2012). Our results do not provide evidence for such effect. Rather they confirm the original assumption of Meyer (2004) and results of other research (e.g. Dixon and Wickens 2007) suggesting that compliance and reliance are selectively affected by false alarms (PPV) and misses (NPV), respectively. This becomes most evident from considering the responses to the green light trials in the three conditions where the NPV was comparatively high (i.e. .98, .93 and .86). All participants showed high levels of reliance irrespective of the fact that the number of false alarms (and PPVs) differed considerably. Only in the most extreme condition where the alarm system turned from a false alarm prone to a miss prone system (NPV=0.42) two out of 10 participants showed a complete loss of reliance by responding to more than 90% of green light as if it were alarms. Note that in this condition the PPV was almost perfect and the alarm system emitted only very few false alarms.

Assessing the overall pattern of results it appears obvious that the finding of negative extreme response strategies and cry-wolf effects in response to alarms with PPV≤0.5 might be specific for a situation where the validity of alarms cannot be cross-checked towards other system data. In this specific case, extreme responding strategies present the most rational choice in order to optimize the overall rate of correct decisions. At least this holds true if the consequences of misses do not dramatically outweigh the consequences of false alarms. Note that this also means that disuse of alarm systems and *cry-wolf* effects can reflect a rational choice under such circumstances, albeit not with respect to safety. This raises the question to what extent decision-making and behavioral strategies in response to alarms would be altered if operators got the opportunity to cross-check the validity of a given alarm towards other data. On the one hand, this would provide the possibility to reduce the uncertainty associated with the output of the alarm system. Yet, on the other hand, it would involve some effort and cost effects in terms of time needed for cross-checking. It is hypothesized that cross-checking options would be used most often in case of moderate PPVs, and that extreme responding strategies would predominate if the PPVs of alarms were very high or low, respectively. That is, verification responses should exhibit a sort of inverted U-shaped relationship to PPV while the probability to ignore alarms or to directly comply with them should rise with PPVs becoming more extreme. This hypothesis was addressed in the second experiment.

3. Experiment 2

3.1. Method

3.1.1. Participants

80 students (40 females, 40 males; mean age = 27.24) participated as paid volunteers in the experiment. They received a basic payment of € 7 for their participation and a bonus payment of maximal € 15 depending on their performance during the experiment. Screening procedures corresponded directly to the ones in the first experiment. 16 participants each were randomly assigned to five experimental groups.

3.1.2. Task

The task was the same as in the first experiment with the important difference that the participants got access to alarm verification information (AVI). This provided them with the opportunity to cross-check the validity of a given output of the alarm system before responding, hence eliminating the uncertainty under which a decision had to be make. In order to check the raw data, they had to open a specific drop-down menu by clicking on a “check”-button and select the container they wanted to cross-check from a list of containers. Due to this selection, a coloured image of the heat distribution in the selected container got augmented after a delay of two seconds and provided always perfect valid data for a visual assessment of the appropriateness of the molecular weight. Once initiated, participants had as much time as needed for the cross-check and had to quit it actively by clicking on a “rework” or “continue” icon presented in the check menu. To complete this whole procedure took about eight seconds and prevented participants to work on the concurrent task. If the participant did not respond within five seconds to a given alarm by either clicking the “rework” or “check” icon on the main interface, the container passed the inspection unchanged, and a new container appeared.

3.1.3. Design

The experimental design corresponded in all details to the one of the first experiment.

3.1.4. Dependent Variables

The same set of variables as in experiment #1 was used to assess the overall performance in the AT and the concurrence task (ROT). In addition, the following variables were used to describe the individual responses in interaction with the alarm system in the AT: (1) the frequency of instances where the participants directly responded to an alarm or nonalarm by clicking on the “rework” button; (2) the frequency of ignoring the outputs of the alarm system, and (3) the frequency of trials where the participants cross-checked a given alarm or nonalarm against the other available data before they selected their response. Note that in this latter case the uncertainty linked to outputs of the alarm system is reduced to zero and hence always lead to a correct response.

3.1.5. Procedure and Statistical Analysis

The procedural details and the approach of statistical analyses corresponded in all aspects to the first experiment.

3.2. Results

3.2.1. Alarm Task: Mean Response Rates and Individual Response Strategies

Relative proportions of the different possible responses to alarms averaged across the two experimental blocks are shown in Figure 4 (upper left quadrant). As expected, the tendency to directly comply with an alarm was extremely low for alarms with low PPVs (0.1; 0.3) but increased considerably across the 0.5, 0.7 and 0.9 condition. For the 0.9 condition direct compliance with the alarms eventually was the most used response, although even in this condition only four/five out of the 16 participants applied a positive extreme responding strategy in the first/second block. All other participants just responded directly to a substantial percentage of alarms (23-82%), but cross-checked the others before eventually responding in the proper way. In all conditions with PPV ≤ 0.7, cross-checking of a given alarm represented the dominant sort of response. In contrast to the hypothesis this dominance also was valid for the two conditions with the lowest PPV. Actually, the frequency of instances where alarms were ignored was very low in these conditions. Only two participants in these conditions ignored more than 50% of all alarms in the second block, and not a single participant was found to have applied a negative extreme responding strategy. Separate 5(condition) x 2(block) ANOVAS of the mean frequencies of the three behavioral strategies revealed a significant influence of condition on direct responses, $F(4,75) = 29.37$ p < .001, η_p^2=.610, and cross-checks, $F(4,75) = 14.38$, p < .001, η_p^2=.434, but not on the frequency by which alarms were ignored, $F(4,75)$=2.22, p>.07, η^2=.106. The tendency to ignore alarms even decreased from block #1 to block #2, reflected in a significant effect of block, $F(1,75) = 6.91$, p<.01; . η_p^2=.084. Post-hoc Scheffé tests revealed that the frequency of cross-checking alarms did not differ among the 0.1 to 0.7 condition, but dropped significantly in the 0.9 condition (all contrasts p<0.001). In turn, the frequency of direct compliance with alarms was higher in the 0.7 and 0.9 conditions compared to all others (all contrasts p<0.05). The condition x block interaction did not become significant for either of these variables (all F<1.0).

Insert Figure 4 about here

The relative proportions of the three possible kinds of response to nonalarms in the different conditions, averaged across the two experimental blocks, are shown in the upper right quadrant of Figure 4. As becomes evident, almost no participant ever responded to a green light as if it was an alarm. This was also valid for the condition with the lowest NPV (0.41) and marks a contrast to the first experiment where this condition had provoked already a significant number of such responses. As a consequence, the effects for the other two kinds of responses were mirror-inverted. The participants' reliance on the alarm system, reflected by the proportion of events where they refrained from any action in case of a green light (no response), decreased significantly with decreasing NPVs, and the proportion of cross-check responses naturally showed an inverse pattern of effect. Not surprising, separate 5(condition) x 2(block) ANOVAS of the proportion revealed a significant influence of condition on both, "no reponse" $F(4,75) = 18.78$ p < .001, η_p^2=.501, as well as on cross-check responses, $F(4,75) = 16.56$, p < .001, η_p^2=.469. A more detailed analysis of the pattern of effect shown in Figure 4 suggests that the relationship between reliance and NPV was not linear but rather exhibited a sort of step profile. This is confirmed by post-hoc contrasts (Scheffé) of the proportion of instances where participants stayed passive in case of a nonalarm, i.e. showed a high level of reliance, in the different conditions. They showed (1) that the level of reliance in the condition with the highest NPV (.98) was significantly higher than in all other conditions (all contrasts p<.04), (2) that the level of reliance in the conditions with NPV=.93 and NPV=.86 was significantly higher than in the condition with NPV=.41 and .72 (all contrasts p<.02). Only the contrast between the .72 and .93 condition failed to reach the usual level of significance (p<.08). No block x condition interaction was found for either variable (all F<1.0).

3.2.2. Alarm Task: Overall Performance

The number of overall correct decisions in response to alarms as well as to nonalarm events are displayed in the lower half of Figure 4. Correct decisions in response to alarms differed significantly dependent on PPV, $F(4,75) = 9.09$, $p<.001$, $\eta_p^2 =.326$. The percentage of correct decisions was very high for the two conditions with lowest PPV (0.1: 98%; 0.3: 95%), decreased somewhat in the 0.5 condition and was lowest for the two conditions with highest PPVs (0.7: 82%, 0.9: 87%). Post-hoc Scheffé tests revealed that performance in the 0.1 and 0.3 conditions differed significantly from performance in the 0.7 and 0.9 condition, respectively (all contrast p< .05). Furthermore, an effect of block was found, $F(1,75)=6.97$, $p<.05$, $\eta_p^2=.085$, indicating a slight performance improvement across the two blocks. No interaction condition x block emerged, $F(4,75)=1.87$, $p>.10$, $\eta_p^2=.091$. For nonalarm trials, the percentage of correct responses decreased continuously with decreasing NPV, $F(4,75) = 4.75$, $p<.01$, $\eta_p^2=.216$. There was neither a significant main effect of block, $F(2,75) = 1.34$, p>.20, $\eta_p^2=.059$, nor a significant interaction, $F(2,75) = 1.17$, $p>.30$, $\eta_p^2=.018$.

3.2.3. Concurrent Task Performance

No significant differences were found in the frequency of correct orders sent by participants in the different conditions, *F*(4,75) = 1.96, *p*>.10, η_p^2=.095. However, a significant main effect of block emerged, *F*(1,75) = 77.88, p < .001, η_p^2=.509 indicating that participants' concurrent task performance was better in the second block (mean = 64.14 orders) than in the first one (58.78 orders). The condition x block interaction was not significant, *F*(4,75) = 1.25, *p*>.20, η_p^2=.062.

3.3. Discussion

Providing access to alarm validity information (AVI) significantly altered the decision-making and response strategies in interaction with the alarm system, compared to the first experiment where such possibility was not provided.

One of the main findings of the first experiment was that two different strategies were applied to cope with the uncertainty of given alarms in case that no AVI was available. The first one involved a mixed strategy, i.e. the participants responded to a certain percentage of alarms in the expected way but ignored others. The second one involved extreme responding strategies where participants ignored almost all alarms in case of low PPVs (≤ 0.5) and responded to all alarms in case of high PPVs (≥ 0.7). It was hypothesized that providing access to AVI would alter this pattern in a way that cross-checking of given alarms would exhibit an inverted U-shape function dependent on the PPV of the alarms, and that negative and positive extreme responding, in turn would only remain the major response in case of extreme low and high PPVs, respectively. This would have been in accordance with similar considerations of Bliss (2003b) and was mainly concluded from the fact that extreme PPVs in either direction reduce the level of uncertainty linked with alarms and hence reduce the benefit gained by alarm verification (see footnote 1). However, our results only support this hypothesis for conditions with PPVs ≥ 0.5. For these conditions, the percentage of direct compliance with alarms increased at the expense of cross-checks with increasing PPV. In the condition with the highest PPV direct compliance with alarms eventually represented the most used category of responses. Contradicting our expectations and, thus, most remarkable, is the finding that participants continued to cross-check almost all alarms in the conditions with low PPVs although most of these checks confirmed that the alarm was false. Even in the 0.1 condition only a minority of participants started to ignore a significant number of alarms and no one actually showed indications of negative extreme responding. Instead, cross-checking alarms represented the dominant strategy. This suggests an obvious asymmetry in tolerating uncertainty associated with alarms, dependent on whether a wrong response would result in a miss or a false alarm. Whereas participants in the 0.9 condition obviously tended to accept the risk of 10% to commit a false alarm and directly complied with most of the alarms, participants in the 0.1 condition avoided the same level of risk to commit a miss by almost always cross-checking the alarm instead of just ignoring it. Note, that this asymmetrical pattern of effects emerged although the pay-off structure of the experiment treated the commitment of both sorts of erroneous decisions the same way.

Related to this behavior are two consequences. The first one involves the elimination of the *cry-wolf* effect. Whereas ignoring alarms represented a common response to alarms with low to medium PPVs if no AVI was available (first experiment), this kind of response was hardly ever observed in the current experiment. Obviously, providing AVI represented an effective mean to prevent this effect and made participants very cautious and conscientious in responding to alarms. This, secondly, produced the somewhat paradoxical effect that the overall percentage of correct decisions was better in conditions with medium to low as compared to high PPVs. However, the high percentage of cross-checks even of alarms with very low PPVs can also be considered as a sort of "over-checking", which has the drawback that it needs effort and, thus, may lead to a waste of resources which otherwise could be invested in concurrent tasks. In the current experiment this was reflected in the fact that average concurrent task performance, in contrast to experiment 1, i.e. did not profit from a reallocation of resources in the conditions where the PPV of alarms was low. Furthermore, the mean number of resources ordered in the three conditions with PPV≤0.5 was lower in the current experiment (60.51) than in experiment 1 (74.67). This raises the question whether the same asymmetrical effect of cross-checking behavior would also be found in situations where the effort needed to verify an alarm or the overall workload would be higher than in the current experiment. For example, Bliss and Dunn (2000) showed that an increased workload stemming from concurrent tasks can significantly boost the *cry-wolf* effect, and it might be assumed that a similar effect emerges if the verification of alarms becomes more complex. This will be explored in experiments #3 and #4.

However, before turning to these experiments, we shortly want to discuss the effects of AVI on the reliance on the alarm system. This aspect has rarely been investigated before. As in the first experiment, responses to green light events seem to be mainly determined by the NPV which again supports the reliance-compliance distinction (Meyer 2004). Yet, it is remarkable that the participants seem to respond especially sensitively to changes in the NPV of the nonalarms if AVI is available. This is indicated by the fact that already the small difference in NPV between the first two conditions (NPV=.98 vs. .93), which did not provoke any visible change of responding behavior in the first experiment, now resulted in a strong behavioral effect reflected in a significant increase of cross-checks. This suggests that the reliance in an alarm system can already be affected by a very small number of misses and that this impact might be stronger than the one of false alarms on compliance.

4. Experiment 3

The main objective of this experiment was to investigate how an increase of effort needed to cross-check given alarms would alter the effects found in the second experiment. It was expected that increasing this effort would decrease the proportion of cross-checks particularly in the conditions with extreme low PPVs and extreme low NPVs in favor of direct compliance to alarms and no responses to green lights, respectively.

4.1. Method

4.1.1. Participants

60 students (40 females, 20 males; mean age: 25.98 years) participated in the experiment. The screening procedure and compensation directly corresponded to the one used in the second experiment. 12 participants each were randomly assigned to the five experimental conditions.

4.1.2. Task Environment

The participants performed the same basic tasks as in the second experiment. However, the procedure to cross-check a given output of the alarm system was made more complex. Specifically, a complete cross-checking of a given alarm now included the inspection of two different parameters, i.e. temperature and pressure. The sequence of necessary actions for a cross-check of the temperature status was the same as in experiment 2. The additional steps to also check the pressure status involved three more actions which corresponded to the procedure of the temperature check, i.e. selecting the correct container from a menu, activating a pressure display, waiting for two seconds until the indicated actual value could be compared with a nominal value, and clicking on the “rework” button in case the pressure needed to be corrected or a “continue” icon in case the pressure was okay. Both parameters needed to be checked in order to unambiguously cross-check the validity of a given output of the alarm system. Overall, this almost doubled the time needed to cross-check a given output of the alarm system (about 14 secs) compared to the procedure in the second experiment.

4.1.3. Design

The design corresponded to the one in the second experiment with the exception that only one experimental block of 100 trials had to be performed after the familiarization block. This was justified by the fact that no significant block x condition effects were found in the second experiment for either of the different response categories.

4.1.4. Procedure and Statistical Analysis

The procedural details and the approach of statistical analysis corresponded to the second experiment.

4.2. Results

4.2.1. Alarm Task: Mean Response Rates and Response Strategies

Relative proportions of the different possible responses to alarms are shown in Figure 5 (upper left quadrant). Again, the pattern of effects was very similar to the effects found in the second experiment although a non-significant trend in the expected direction could be observed. However, cross-checking of alarms again represented the most often response used in conditions 0.1-0.7 and only declined considerably in the 0.9 condition. A univariate between-subjects ANOVA of cross-checking rates revealed a main effect of condition, $F(4,55)=7.57$, $p<.01$, η_p^2 =.395. Post-hoc Scheffé tests showed that the proportion of cross-checking was higher in the 0.1-0.5 condition than in the 0.9 condition (all p<.02). This effect was mirrored by the results for direct compliance with alarms which was rarely observed in the conditions with medium and low PPVs but considerably increased in the 0.7 and 0.9 condition, $F(4,55)=17.81$, $p<.01$, η_p^2 =.71. Post-hoc contrasts showed that the percentage of alarms directly responded to were higher in the 0.9 condition than in all other conditions (all p<.03), higher in the 0.7 than in the 0.1 condition (p<.02), and tentatively higher in the 0.7 than in the 0.3 condition (p<.08). In contrast, only a comparatively weak impact of PPV with respect to the strength of effect was found on the proportion of alarms ignored by the participants, F(4.55)=2.59, p<.05, η_p^2 =.158. Yet, post-hoc Scheffé tests did not become significant for any pairwise comparison. This pattern of mean response rates described very well the findings on individual level. Inspection of individual data revealed that only three out of the 60 participants (one each in the 0.1, 0.3 and 0.5 condition) showed a significant *cry-wolf* effect by ignoring more than 80% of all alarms. In contrast, six participants in the 0.9 condition and two participants in the 0.7 condition exhibited positive extreme responding. All others cross-checked a significant number of alarms before choosing their response.

Insert Figure 5 about here

Results for the relative proportion of the three possible ways to respond to nonalarms are shown in the upper right quadrant of Figure 5. Again the overall pattern mirrored more or less the pattern found in the second experiment with the important exception that in the condition with the lowest NPV an increase of direct responses to green lights was observed. This resulted in a main effect of condition for this response category in a univariate between-subjects ANOVA, F(4,55)=2.57, p<.05, η_p^2 =1.57. However, inspection of the raw data revealed that this was due to only two out of the 12 participants in this experimental group who responded to more than 90% of green lights in this manner. The effects for the other two response categories resembled the effects found in the second experiment, albeit less extreme. Reliance on green lights was highest in the condition with the highest NPV and declined with decreasing NPVs, F(4,55)=9.121, p<.01, η_p^2 =.399. Post-hoc Scheffé tests revealed significant differences between the 0.98 and 0.72 condition (p<.05), the 0.98 and 0.41 condition (p<.01), and the 0.93/0.86 conditions and the 0.41 condition (p<.01). The effect for cross-checking events showed an inverted pattern, *F*(4,55)=4.13, *p*<.01, $\eta_p{}^2$ =.231, with a significant post-hoc test only for comparison of conditions 0.41 and 0.98.

In order to explore in more detail to what extent the use of cross-checks was affected by raising the check-effort, we compared the effects for cross-checking of alarms and nonalarms with those found in the second experiment (averaged across blocks). A 2(experiment) x 5(condition) between-subjects ANOVA revealed significant main effects of experiment for both cross-checks of alarms, *F*(1,130)=4.57, *p*<.04, $\eta_p{}^2$ =.034, as well as nonalarms, *F*(1,130)=12.69, *p*<.01, $\eta_p{}^2$ =.085, reflecting that the average number of cross-checks was somewhat lower in the third experiment. However, most important, the experiment x condition interaction did not become significant for either variable, both *F*<1.0, supporting the similarity of effects across conditions described above.

4.2.2. Alarm Task: Overall Performance

The number of overall correct decisions in response to alarms as well as to nonalarms are displayed in the upper half of Figure 5. Corresponding to the results of experiment #2, the percentage of correct responses to alarms was highest in the 0.1 condition and declined with increasing PPV, *F*(4,55)= 2.76, *p*<.04, $\eta_p{}^2$ =.167. However, post-hoc Scheffé tests only proved the difference between the 0.1 and 0.7 condition to be marginally significant (<.09). For nonalarm trials, the percentage of correct responses decreased continuously with decreasing NPV, *F*(4,55) = 13.27, *p*<.01, $\eta_p{}^2$=.491. Post-hoc contracts revealed significant differences between the 0.98/0.93 and 0.86/0.41 conditions, respectively (p<.01), as well as between the 0.5 and 0.9 condition (p<.04).

4.2.3. Concurrent Task Performance

No significant performance differences across conditions were found for performance in the ROT: *F*<1.0. On average the participants placed 58.98 correct orders.

4.3. Discussion

The results replicated the findings of the second experiment. Only few indications were found that doubling the effort for verification of given outputs of the alarm system would reduce the number of cross-checks. Notably, these became more obvious in responses to green lights, i.e. the reliance aspect, than in responses to alarms. Obviously the effort needed to cross-check a given output of the alarm system increased the reliance on the system's output but did not affect the compliance to the same extent. Although some trends also emerged with respect to compliance with alarms which were in line with the expectation (e.g. more ignored alarms in the 0.1 condition and more compliance in the 0.7 condition), they remained non-significant. Moreover, also the asymmetry effect found in the previous experiment, i.e. the tendency to rather accept the (low) risk of an erroneous response by complying with an alarm with a high PPV, as by ignoring an alarm with a very low PPV, did not become affected by raising the cross-check effort. Overall, this can be taken as evidence for the stability of effects found for compliance with alarms in the second experiment. Yet, it remains an open question whether this also would be valid for the situation where the overall workload imposed by concurrent tasks would be higher. As is suggested by research of, e.g., Bliss and Dunn (2000) and McBride, Rogers and Fisk (2011), increments of workload can affect the compliance with decision aids like alarms. Accordingly it might be assumed that raising the overall workload would lead operators to exhibit more extreme responding to alarms with extreme high and low PPVs even though AVI is available. This hypothesis was explored in the fourth experiment.

5. Experiment 4

5.1. Method

5.1.1. Participants

60 students (35 females, 25 males, mean age: 25.97 years) participated in the experiment and were randomly assigned to one of the five conditions, 12 participants each. Screening procedure and compensation directly corresponded to experiments #2 and #3.

5.1.2. Task Environment

As in the former two experiments the M-TOPS-A environment was used as simulation environment. However, in order to increase the overall workload, all three tasks of M-TOPS-A had to be performed concurrently.

5.1.3. Design

The design corresponded to the one in the third experiment.

5.1.4. Procedure and Statistical Analysis

The procedure and the approach of statistical analysis corresponded to the other experiments. However, the pay-off rules were enlarged by including the third task. As in the three previous experiments, participants received 1.5 points for every correct order of chemicals in the ROT they gained and lost two points for correct and false decisions in the *AT,* and were granted 7.5 points for every completed set of containers in the CET. This payoff pattern should support equal treatment of the three tasks, taking into account that completing a full refilling cycle in the CET took much longer than single trials of the ROT and AT.

5.2. Results

5.2.1. Alarm Task: Mean Response Rates and Response Strategies

Frequencies of the different possible responses to alarms are shown in Figure 6 (upper left quadrant). The pattern of effects revealed some obvious differences compared to the effects found in the second experiment. Direct compliance with alarms was lowest in the 0.1

condition but started to increase already in the 0.3 condition and represented the major category of responses not only in the 0.9 but also in the 0.7 condition. A univariate between-subjects ANOVA revealed a significant effect of condition, *F*(4,55)=13.06, *p*<.01, η_p^2 =.487. Post-hoc contrasts (Scheffé) showed that the percentage of direct responses were higher in the 0.7 and 0.9 conditions than in the 0.1 and 0.3 conditions (all p<.01). Another difference was found for “no response” events. These were rarely observed in conditions with PPV of 0.3-0.9 but increased considerably in the 0.1 condition, *F*(4,55)=6.95, *p*<.01, η_p^2 =.336. Post-hoc contrasts showed that the percentage of alarms not responded to was higher in the 0.1 than in all other conditions (all p<.05). Actually, three of the 12 participants in this condition exhibited negative extreme responding by ignoring more than 90% of the alarms. Cross-checking of alarms represented the major category only in the 0.3 and 0.5 condition but was less observed in the other condition, *F*(4,55)=5.68, *p*<.01, η_p^2 =.292. Yet, post-hoc Scheffé tests only revealed the differences between the 0.3 conditions and the 0.7/ 0.9 condition to be significant (all p<.02). Comparing the effects for alarm verification across conditions with those found in the second experiment by means of a 2(experiment) x 5(condition) between-subjects ANOVA resulted in a significant main effect of condition, *F*(1,130)=19.92, *p*<.01, η_p^2 =.133, as well as a significant experiment x condition interaction, *F*(4,55)=3.24, *p*<.02, η_p^2 =.091.

Insert Figure 6 about here

Results for the frequency of the three possible kinds to respond to *non*alarms are shown in the upper right quadrant of Figure 6. Almost no participant did ever respond to a green light as if it was an alarm. As a consequence the patterns for the other two response categories were directly inverted to each other. The level of reliance as reflected in the proportion of green lights which were neither responded to nor cross-checked decreased continuously with decreasing NPV, *F*(4,55)=17.17, *p*<.01, η_p^2 =.555. Post-hoc Scheffé tests confirmed that reliance in the condition with the highest NPV was significantly higher than in the two conditions with lowest NPV (p<.01). In addition, also the comparisons between the condition with NPV=.41 and the conditions with NPV=.86 and .72 became significant (p<.05). The general pattern of this effect resembled the pattern found in the second and third experiment with the important exception that the level of reliance remained high for both conditions with NPV>.90. Compared to the second experiment, where already a decrement of NPV from .98 to .93 led to a significant increase of the percentage of cross-checked green lights from zero up to about 40%, the same change of NPV only led to a negligible increase to about 9% in the present experiment. This difference resulted in a significant main effect of experiment, *F*(1,130)=20.31, *p*<.01, η_p^2 =.135, as well as a significant experiment x condition interaction, *F*(1,130)=2.87, *p*<.03, η_p^2 =.081, when we compared the effects of the second and current experiment by means of a 2(experiment) x 5(condition) between-subjects ANOVA.

5.2.2. Alarm Task: Overall Performance

The number of overall correct decisions in response to alarms as well as to no-alarm events are displayed in the upper half of Figure 6. Correct decisions in response to alarms differed significantly dependent on PPV, *F*(4,55) = 6.93, *p*<.001, η_p^2 =.335. The percentage of correct decisions was highest in the 0.1 condition and decreased almost linearly with increasing PPV except a small improvement in the 0.9 condition. Post-hoc Scheffé tests revealed that performance in the 0.1 conditions differed significantly from performance in the 0.5 and 0.7 condition, respectively (both contrasts p< .02). In addition, the contrast between the 0.3 and 0.7 condition just failed to reach the conventional level of significance (p=.051). For nonalarm trials, the percentage of correct responses decreased continuously with decreasing NPV, *F*(4,55) = 11.96, *p*<.01, η_p^2=.465. Post-hoc contracts revealed significant differences between the 0.1 and 0.7/0.9 condition (p<.01) and the 0.3 and 0.9 condition (p<.05).

5.2.3. Concurrent Task Performance

No significant performance differences across conditions were found in either of the two concurrent tasks, ROT: $F(4,55)=1.65$, $p>.15$, $\eta_p^2=.107$; CET: $F<1.0$. Averaged across conditions, a mean of 49.68 orders were sent correctly in the ROT, and 11.77 cycles of coolant exchange were completed in the CET.

5.3. Discussion

It was expected that raising the workload imposed by concurrent tasks would lead operators to choose extreme responding strategies to alarms with extreme high and low PPVs even though AVI is available. The results of the fourth experiment at least partially support this hypothesis. In contrast to the second experiment where cross-checking of alarms was found to be by far the most often used response across a large range of PPVs (0.1-0.7), participants now started to comply directly with most of the alarms already in the condition with PPV= 0.7 and ignored the majority of alarms in the 0.1 condition. These behavioral aspects resembled the situation where no AVI was available (first experiment) and suggest that participants facing high workload started to trust in the output of the alarm system more than in situations where the overall workload was lower (second experiment). This effect is in accordance with previous findings of Bliss and Dunn (2000) who reported an increasing tendency of *cry-wolf* effect in response to a high false alarm rate if overall workload was high. It is further in line with results of Biros, Daly, and Gunsch (2004) and McBride et al. (2011) who found that increased level of workload led to increments of compliance with moderately reliable (.60-.70) decision aids in different contexts.

A similar effect also was found for the reliance aspect. Compared to the second experiment, where already a decrement of NPV from .98 to .93 led to a significant increase of the percentage of cross-checked green lights, the same change of NPV just entailed a very weak behavioural effect in the present experiment. Even in the conditions with NPV=.72 only about 30% of green lights were checked. This result suggests that the level of reliance on the output of the alarm system became much higher in the current experiment with high overall workload as compared to the second experiment where workload imposed by concurrent tasks was significantly lower.

To some extent the results suggests that the raised overall workload provoked a more rational decision-making and response selection in case of extreme PPVs and NPVs. Nevertheless, the previously found asymmetry of acceptance of uncertainty in interaction with alarms also persisted to some extent in the current experiment. This becomes evident by comparing the proportions of alarm verifications in the 0.3 and 0.7 conditions. Although the risk to commit an erroneous decision was the same by ignoring a given alarm in the 0.3 condition as compared to comply with it in the 0.7 condition, the number of alarm verification was found to be significantly higher in the former than the latter conditions. This corresponds to the finding of “over-checking” behavior in the second and third experiment.

6. General Discussion and Conclusions

The results of the four experiments provide clear empirical evidence for the assumption outlined in the introduction that PPV and NPV are crucial triggers for participants’ decision-making and response selection in interaction with alarms. Overall, this finding confirms similar results reported by Getty et al. (1995). Given the fact that the technical characteristics of the alarm system (i.e. d’, response bias) remained essentially the same and the differences of PPV (NPV) were just induced by changing the base rates of critical events, the results further empirically supports the claim made by Parasuraman et al.

(1997) that the same alarm systems can entail very different behavioural effects dependent on the *a priori* probabilities of critical events.

However, the specific response characteristics are highly dependent on whether or not AVI is available. In the case that a cross-checking of the output of the alarm system is not possible, the decision how to respond to it represents a decision under uncertainty with the level of uncertainty directly reflected by PPV and NPV. The results of the first experiment suggest that humans take these differences into account in their decision-making, i.e. adapt their behavioural strategies to the level of uncertainty of alarms and nonalarms. This was most true for high PPVs and NPVs (≥0.7) where the vast majority of participants in the first experiment applied an extreme responding strategy by complying with all alarms and not responding to nonalarms, thereby accepting the small risks to commit a false alarm or to miss a critical event. However, a more diverse set of strategies was found in case that the PPVs of the alarms were ≤0.5. In this case two subgroups were found. The first one applied a negative extreme responding by not responding to 90% or more of all alarms. The second group showed a sort of mixed strategy, i.e. responded to a certain percentage of alarms and only ignored the others. Obviously, participants choosing this strategy are reluctant to stay passive in case of alarms and think that they can compensate to some extent the unreliability of the alarm system and to meet real critical events by chance. However, such strategy does not reflect a rational choice. This is given by the fact that without access to AVI human operators do not have any information which would enable them to discriminate false and true alarms. As a consequence, they principally cannot increase the number of correct decisions by some strategy above that which would be achieved by positive extreme responding when PPV>0.5 and negative extreme responding in case of PPV<0.5. In the first experiment this became evident in the percentage of overall (human and alarm system) correct responses. Due to the fact that a substantial number of participants chose a sort of probability matching in conditions with PPV<0.5, the mean number of correct decisions in these conditions did not reach those which were to be expected if all alarms were ignored. However, a better record was achieved for conditions with PPV>0.5 where the vast majority of participants applied a positive extreme responding strategy.

Two practical implications might be derived from these results for situations where AVI cannot be made available. First, for environments where alarm systems with PPVs greater than 0.5 are technically feasible, there is no real need to keep human operators in the loop, and the best choice would be to fully automate the response to alarms. This is suggested by the fact that human operators would tend to respond to all alarms anyway. Second, for conditions with alarm PPVs<0.5, strategies that involve at least to some extent a cry-wolf effect seem to represent the most probable and, to a high degree, also rational choice. This has particular implications for situations where misses are much more consequential than false alarms like in high risk environments (e.g. power plants or airplanes). In these situations, human operators should get clear and binding orders what to do in case of an alarm or nonalarm, instead of leaving this decision with them. Specifically direct responses to each alarm independent of its PPV should be requested, as well as careful repeated monitoring of raw data in case that a high NPV of an alarm system cannot be assured. Note that the former, e.g., is the usual prescription of responses to outputs of the traffic alert and avoidance system (TCAS) in cockpits of commercial airplanes and also the most applied procedure in control rooms of chemical plants. Another possibility is the provision of what has been referred to as likelihood alarm systems (Sorkin, Kantowitz and Kantowitz 1988). These sorts of systems provide a more distinct information about the relative likelihood of critical events and haveshown to improve decision-making significantly if other informations sources are not available (Wiczorek and Manzey 2014).

The situation changes if AVI is available. Investigating the behavioural effects of AVI availability was the major focus of the present research. Providing AVI offers the possibility to reduce the uncertainty linked to outputs of the alarm system before selecting a response.

Therefore it was assumed that the option to cross-check outputs of the alarms system before responding would be mostly chosen in cases where this uncertainty is comparatively high, i.e with medium PPVs and NPVs, but not if the uncertainty is low. The results of the second experiment supported this assumption only for the upper parts of PPV and NPV range, i.e. PPV and NPV > 0.7. For the lower part of the investigated PPV range, which is more realistic for characteristics of alarm systems in the field, cross-checking of the outputs of the alarm system remained the major response even in the condition where 90% of all alarms were false. This general effect also persisted when the effort to check the AVI was increased. Although some indications were found that doubling the check effort indeed reduced the checking of AVI in response to alarms as well as green lights (third experiment), this reduction was only small and occurred in all conditions, i.e. did not change the overall pattern of effects. This only changed if the overall workload imposed by concurrent task was increased (fourth experiment). This intervention eventually led to results that were more close to expectations. In accordance with previous findings of Bliss and Dunn (2000) participants who had to perform two tasks (instead of one) concurrently to the AT started to comply with and ignore more alarms in case the PPV was extremely high or low, respectively. For the case of extreme low PPV this involved the "re-appearance" of the cry-wolf effect even with AVI available. However, independent of the raised workload, cross-checking remained to be the major response for alarms with a PPV as low as 0.3. This is in accordance with the assumption stated earlier by Bliss (2003b) that AVI generally would reduce extreme responding behaviour and has several implications.

It suggests that providing AVI that can be easily checked, combined with keeping the overall workload of operators on moderate levels can significantly reduce the *cry-wolf* effect which constitutes one of the major issues of the "psychology of false alarms" (Breznitz 1984). This insight contributes to the current discussion on the significance of the *cry-wolf* effect in different fields (e.g. Lees and Lee 2007; Wickens et al. 2009). Furthermore, the finding that our participants tended to “over-check” alarms in case of extreme low but not in case of extreme high PPVs suggests that the tolerance of uncertainty in interaction with alarms differs dependent on the sort of error which might result from a wrong decision. That is, humans are obviously more motivated to reduce the uncertainty associated with the output of an alarm system if an erroneous action would lead to a miss than to a false alarm. Indications of this asymmetry were found in all experiments where AVI was available and, thus, emerged independent on check effort and overall workload. On first sight, this might be regarded as an obvious (and safety oriented) behavioural effect in interaction with alarm systems. Independent of the pay-off structure which provided the same penalties for misses and false alarms in the experiments reported above, the participants might have perceived the risk of committing a miss as higher as the one of a false alarm. This might have also been supported by the cover story that they would be operators in a chemical plant. However, also a more general effect might be responsible for this finding. Committing an erroneous decision by complying with an alarm might be perceived as less detrimental than committing an error by ignoring the alarm. Such “psychological bias” might not be specific for alarm systems but could be valid for all kinds of automated decision aids (e.g. navigation systems in cars). That is, humans interacting with a decision aid might be biased to actively respond to the aid in the expected way and perceive even committing an error by complying with a system as more tolerable than making a false decision by just ignoring the automated cue. A similar effect has been described in other decision-making contexts and referred to as “action bias” (e.g. Bar-Eli et al. 2007). First results from another series of experiments in our lab provide support for this assumption (Guenzler and Manzey 2013) but more research will be needed before such far-reaching claim can be made.

A final conclusion regards the distinction of two behavioural indicators of dependence on alarm systems, i.e. compliance and reliance (Meyer 2004). The results of all four experiments show that participants discriminated between these two aspects in a straightforward manner. No obvious evidence of cross-effects between the effects of false alarms and misses on compliance and reliance were found. This provides support for the assumption and previous findings that these two behavioral aspects represent largely independent behavioural consequences of alarms and nonalarms (Dixon and Wickens 2006; Meyer 2004; Meyer and Bitan 2011; Wickens and Colcombe 2007). However, it contrasts some recent results at least suggesting that reliance might be affected by both false alarms and misses (Dixon et al. 2007; Wiczorek et al. 2012). Admittedly, the current experiments were not specifically designed to investigate the dependence of both aspects and, thus, more research will be needed to explore possible moderating factors which might affect the dependence-independence of both aspects.

7. Limitations

The present research has been based on laboratory experiments using a sort of micro-world task. Whereas this ensured a high level of internal validity of the effects reported, it suffers from the usual disadvantages of laboratory research in terms of reduced complexity compared to real-world settings and, thus, limited external validity. First, the current research considered decision-making to single binary alarms, mainly based on probability information in terms of PPV and NPV. Thus, the complexity of decision-making usually found in the field was considerably reduced. Specifically, other relevant context factors than just probability information, or aspects like dealing with multiple more or less connected alarms, were not taken into account. Moreover, the model chosen for our research included the most basic form of alarm system, i.e. a simple binary threshold detector whereas in the field also more complex systems are available (e.g., Pritchett 2001). Thus, our findings might only be externally valid for situations where operators have to interact with single binary alarm systems in contexts where their experiences concerning the probability of false alarms and misses represent the main basis for their decision-making and response selection. Second, direct responding to a valid or false alarm in our paradigm only included the click on a re-work button . This does not compare to real-world settings where responding to alarms usually involves much work. It cannot be excluded that this characteristic of the current studies has led to an overestimation of direct compliance with alarms in the upper PPV range. Furthermore, the risks usually associated with misses and false alarms in the field are rarely known in detail and often unbalanced in a way that misses are much more consequential than false alarms. In our series of experiments, the pay-off structure underlying erroneous decisions was completely transparent, and false alarms were treated as consequential as misses. This might have affected in particular the observation of high percentages of ignored alarms in the first experiment already in conditions with moderate PPVs. Finally, the AVI provided in experiments 2-4 always reduced the uncertainty involved in the decision-making process to zero. Admittedly this also was a specific characteristic of our laboratory research that does not apply to all real-world settings. Thus, it remains an open question to what levels consulting AVI must reduce uncertainty in order to evoke the behavioural effects observed in the present research.

Acknowledgements

Thanks are due to Marcus Bleil for programming the M-TOPS-A software, and Marlene Vogel, Ulrike Schmuntzsch and Julia Zorn for their support with data collection and analysis. We further thank Magali Balaud and three anonymous reviewers for their helpful comments to earlier versions of this paper.

Funding

The research presented in this article was supported by a scholarship provided to the second author by the German Science Foundation (DFG).

References

Allendoerfer, K. R., S. Pai, and F. J. Friedman-Berg. 2008. “The complexity of signal detection in air traffic control alert situations.” *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*: 54-58.

Bar-Eli, M., O.H. Azar, I. Ritov, Y. Keidar-Levin, and G. Schein. 2007. Action bias among elite soccer goalkeepers: The case of penalty kicks. *Journal of Economic Psychology* 28: 606-621.

Biros, D. P., M. Daly, and G. Gunsch. 2004. “The influence of task load and automation trust on deception detection.” *Group Decision and Negotiation* 13: 173-189.

Bliss, J. P. 2003a. “Investigation of alarm-related accidents and incidents in aviation.” *International Journal of Aviation Psychology* 13: 249-268.

Bliss, J. P. 2003b. “An investigation of extreme alarm responses of extreme alarm response patterns in laboratory experiments.” *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, 1683-1687. Santa Monica, CA: Human Factors and Ergonomics Society.

Bliss, J. P., and M. Dunn. 2000. “Behavioural implications of alarm mistrust as a function of task workload.” *Ergonomics* 43: 1283-1300.

Bliss, J. P., and C. K. Fallon. 2006. “Active warning: False alarms.” In *Handbook of warnings*, edited by M. S. Wolgater, 231-242. New York: Lawrence Erlbaum.

Bliss, J. P., R. D. Gilson, and J. E. Deaton. 1995. “Human probability matching behaviour in response to alarms of varying reliability.” *Ergonomics* 38: 2000-2012.

Bliss, J. P., S. M. Jeans, and H. J. Prioux. 1996. “Dual-task performance as a function of individ-ual alarm validity and alarm system reliability information.” *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, 1237-1241. Santa Monica, CA: Human Factors and Ergonomics Society.

Bransby, M. and J. Jenkinson. 1998. Alarming performance. *Computing and Control Engineering Journal 9: 61-67.*

Breznitz, S. 1984. *Cry-wolf: the psychology of false alarms.* Hillsdale, NJ: Erlbaum.

Dixon, S. R., and C. D. Wickens. 2006. “Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high work-load.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48: 474 – 486.

Dixon, S. R., C. D. Wickens, and J. S. McCarley. 2007. “On the independence of compliance and reliance: Are automation false alarms worse than misses?” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49: 564-572.

Getty, D. J., J. A. Swets, R. M. Pickett, and D. Gonthier. 1995. “System operator response to warnings 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: 19-33.

Gigerenzer, G., and P. M. Todd. 1999. *Simple heuristics that make us smart*. New York: Oxford University Press.

Green, D. M., and J. A. Swets. 1966. *Signal detection theory and psychophysics*. New York: Wiley.

Guenzler, T., and D. Manzey. 2013. “Asymmetries in human tolerance of uncertainty in interaction with alarm systems: effects of risk perception or evidence for a general commission bias? ” *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society* (1362-1366). Santa Monica, CA: Human Factors and Ergonomics Society.

Hertwig, R., and I. Erev. 2009. “The description – experience gap in risky choice.” *Trends in Cognitive Science* 13: 517-523.

Kahnemann, D., P. Slovic and A. Tversky (eds.) (1982). *Judgement under uncertainty: heuristics and biases*. New York: Cambridge University Press.

Kirk, R.E. (1982). Experimental design. Procedures for the behavioural sciences. 2nd ed. Belmont: Brooks/Cole.

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

Maltz, M., and D. Shinar. 2003. “New alternative methods of analyzing human behavior in cued target acquisition.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 45: 281-295.

McBride, S.E., W. A. Rogers and A. D. Fisk. 2011. “Understanding the effect of workload on automation use for younger and older adults.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53: 672-686.

Meyer, J. 2001. “Effects of warning validity and proximity on responses to warnings. ” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43: 563-572.

Meyer, J. 2004. “Conceptual issues in the study of dynamic hazard warnings.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46: 196-204.

Meyer, J., and Y. Bitan. 2002. “Why better operators receive worse warnings.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 44: 343-354.

Parasuraman, R., P. Hancock, and O. Olofinboba. 1997. “Alarm effectiveness in driver-centred collision-warning systems.” *Ergonomics* 40 (3): 390–399.

Sorkin, R.D., Kantowitz, B.H., and Kantowitz, S.C. (1988). Likelihood alarm displays. Human Factors, 30, 445–459.

Sorkin, R. D., and D.D. Woods. 1985. “Systems with human monitors: a signal detection analysis.” *Human-computer interaction* 1: 49-75.

Stevens, J.P. (2007). Intermediate statistics. A modern approach. 3rd ed. New York: Lawrence Erlbaum.

Wickens, C. D., and A. Colcombe. 2007. “Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49: 839-850.

Wickens, C. D., S. Rice, D. Keller, S. Hutchins, J. Hughes, and K. Clayton. 2009. “False alerts in air traffic control conflict alerting system: Is there a “cry wolf” effect?” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 51: 446-462.

Wiczorek, R. and Manzey, D. (2014). Supporting attention allocation in multitask environments: effects of likelihood alarm systems on trust, behavior, and performance. Human Factors, DOI: 10.1177/0018720814528534

Wiczorek, R., J. Meyer, and T. Günzler. 2012. “On the Relation Between Reliance and Compliance in an Aided Visual Scanning Task.” *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society*, 253-257. Santa Monica, CA: Human Factors and Ergonomics Society.

Table 1: Experimental conditions (between subjects) of experiment 1

Groups	Number of Participants	Base Rate of Critical Events	Characteristics of Alarm System			
			d’	c	PPV	NPV
1	12	.05	1.09	-0.29	.10	.98
2	11	.18	1.09	-0.29	.30	.93
3	12	.33	1.09	-0.29	.50	.86
4	11	.54	1.09	-0.29	.70	.72
5	10	.81	1.09	-0.29	.90	.41

Figure Captions

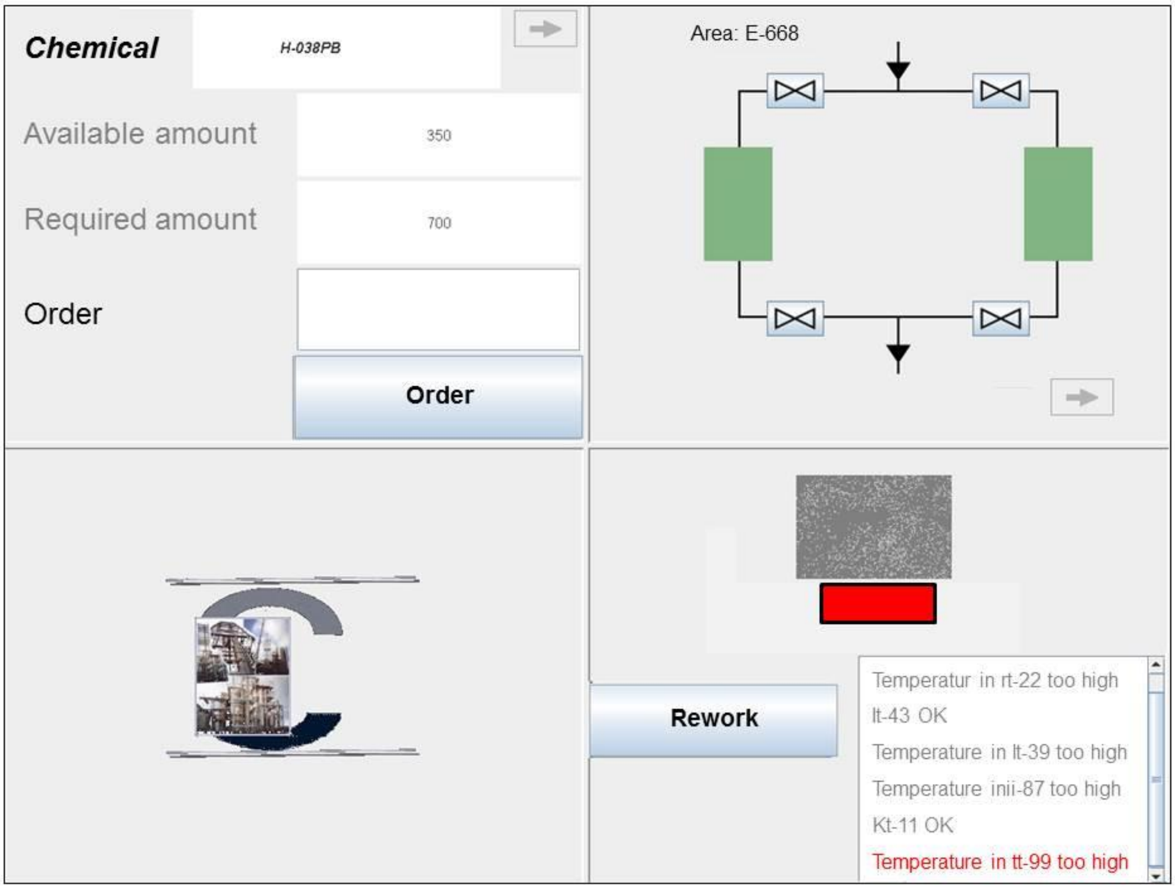
Figure 1: User interface of M-TOPS 2 with ROT (upper left quadrant), CET (upper right quadrant) and AT (lower right quadrant). The figure shows the version used in experiment 1, i.e. without possible access to alarm validity information in the AT. Figure 2: Results of experiment #1. Upper panel: Means and standard errors of proportions of direct “rework” responses to alarms (left) and nonalarms (right) across experimental conditions and the two experimental blocks. Medium panel: Percentage of participants showing negative extreme responding, positive extreme responding and mixed strategies as defined in the text (only the second experimental block considered). Lower panel: Means and standard errors for overall correct decisions in response to alarms left) and nonalarms (right) across conditions for the two different experimental blocks.

Figure 3: Concurrence tasks performance in experiment #1. Shown are means and standard errors oft correct resource orders sent in the ROT for the five different conditions and the two blocks.

Figure 4: Results of experiment #2. Upper panels: Means and standard errors of proportions of different kinds of responses to alarms (left) and nonalarms (right). Lower panels: Means and standard errors for overall correct decisions in response to alarms (left) and nonalarms (right).

Figure 5: Results of experiment #3. Upper panels: Means and standard errors of proportions of different kinds of responses to alarms (left) and nonalarms (right). Lower panels: Means and standard errors for overall correct decisions in response to alarms (left) and nonalarms (right).

Figure 6: Results of experiment #4. Upper panels: Means and standard errors of proportions of different kinds of responses to alarms (left) and nonalarms (right). Lower panels: Means and standard errors for overall correct decisions in response to alarms (left) and nonalarms (right).



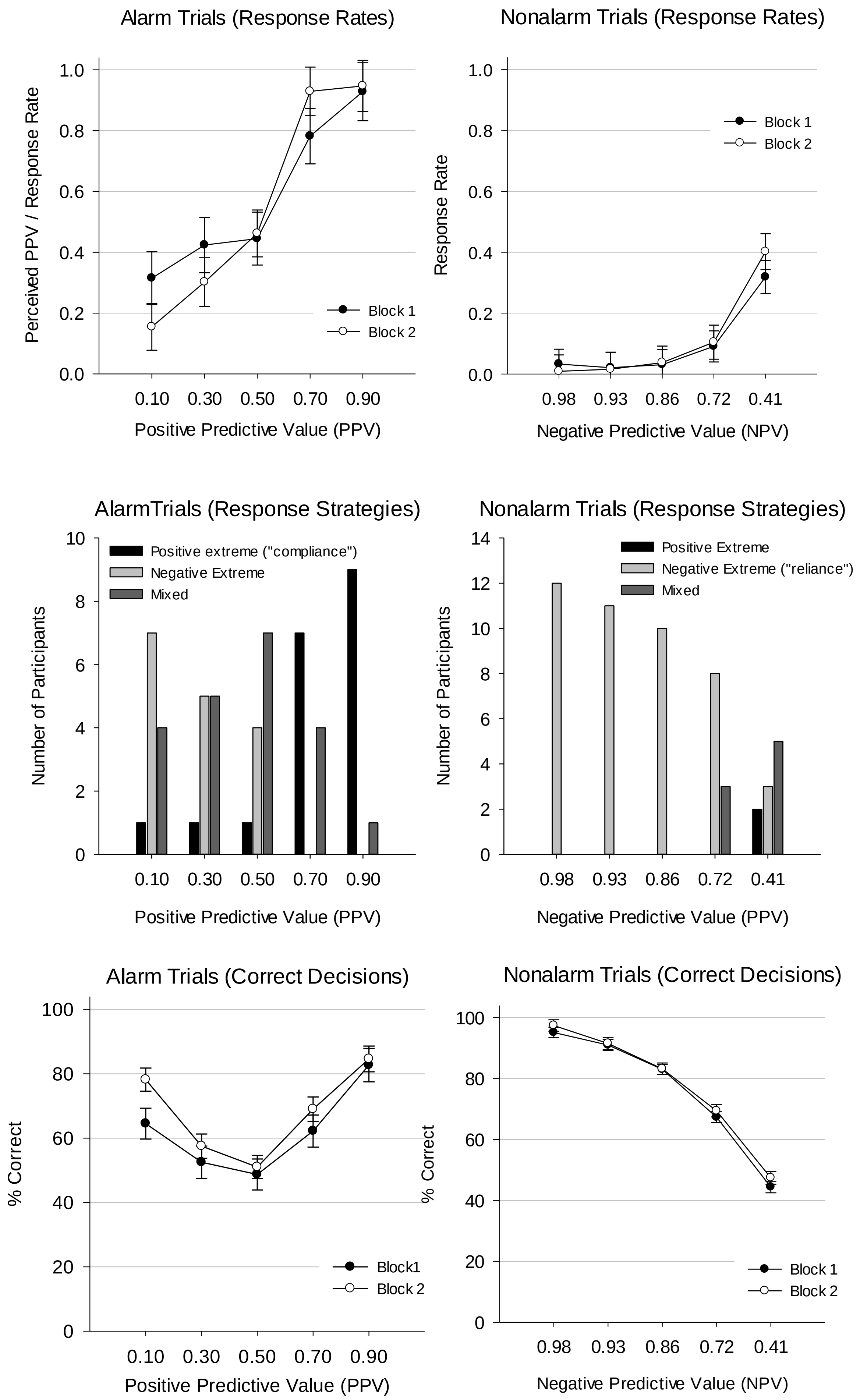
Chemical H-038PB

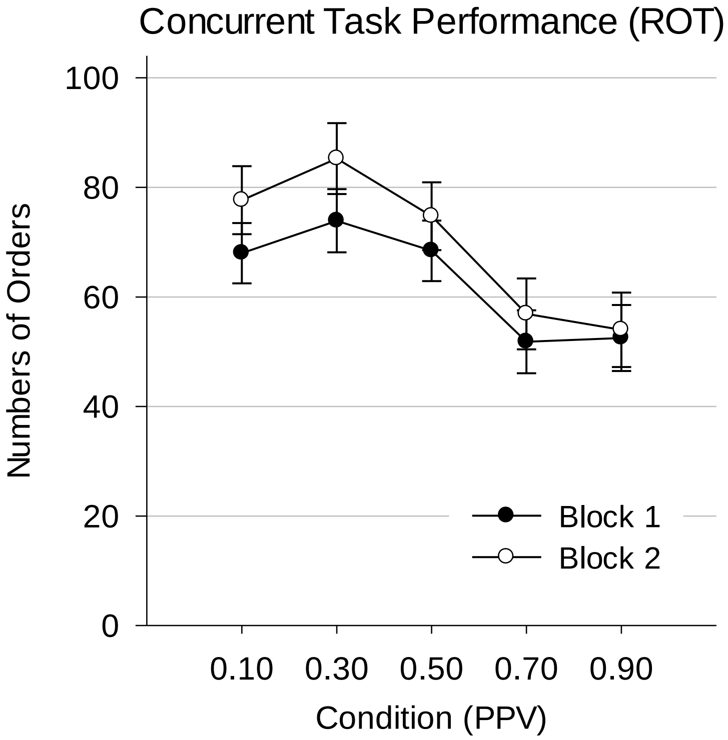
Available amount 350

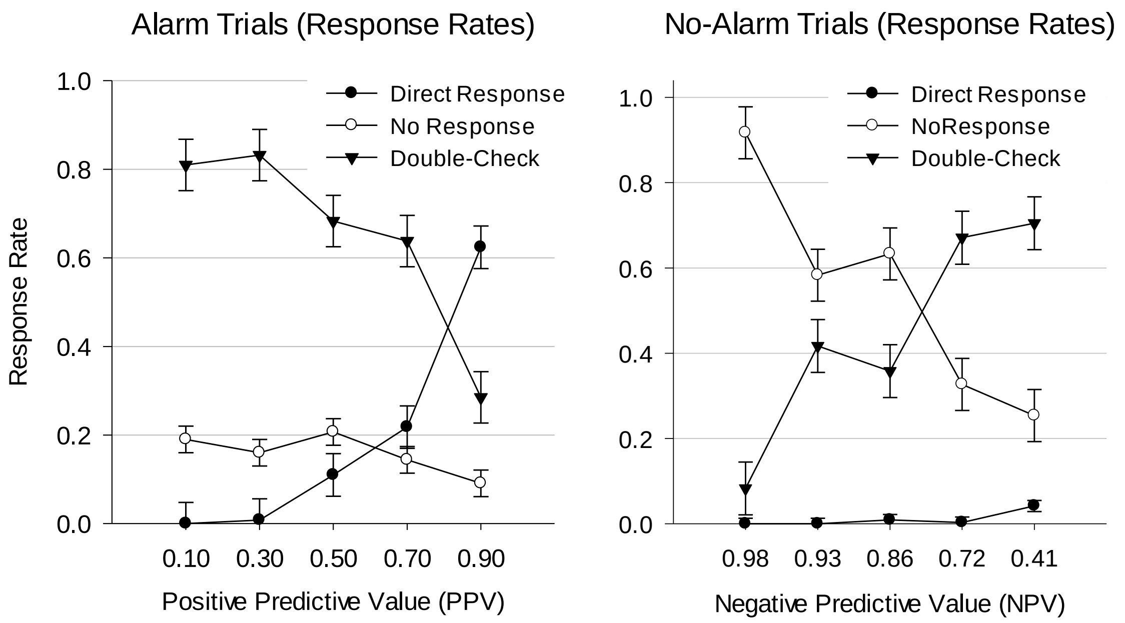
Required amount 700

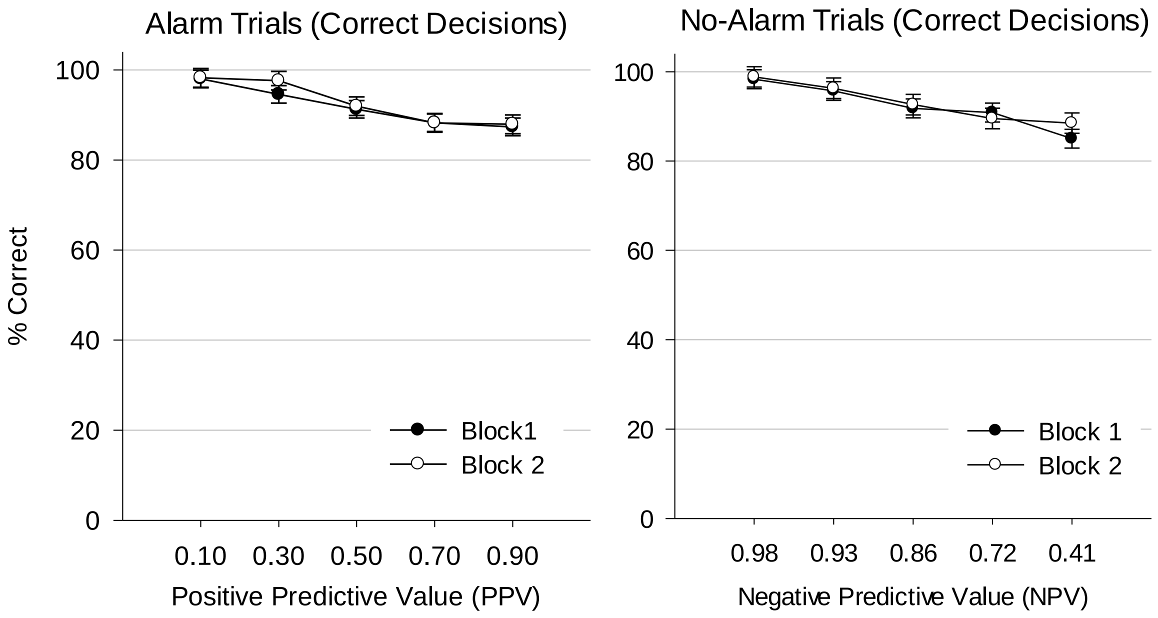
Order

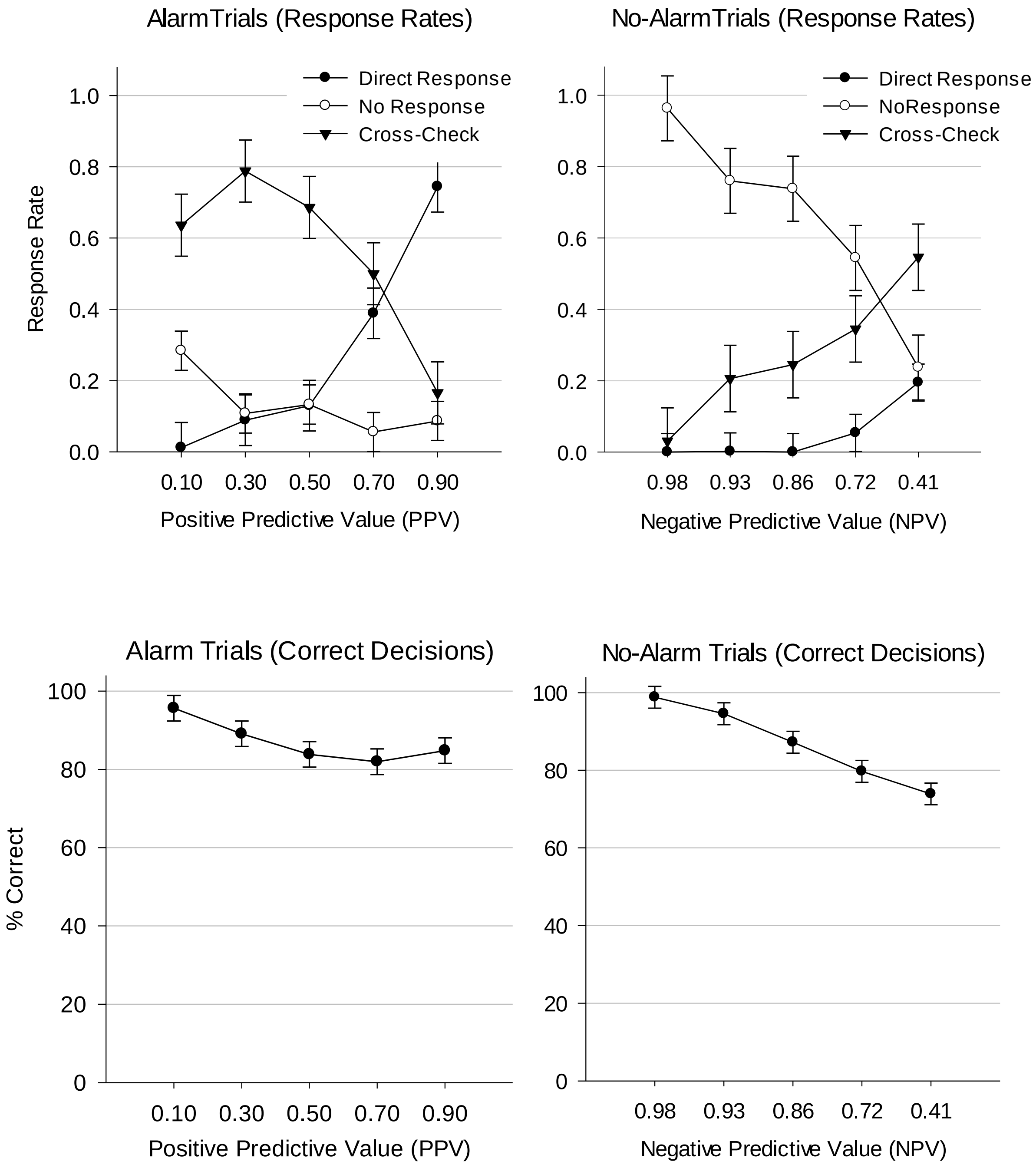
```
Temperatur  in rt-22 too high
lt-43 OK
Temperature  in lt-39 too high
Temperature  inii-87 too high
Kt-11 OK
Temperature  in tt-99 too high
```

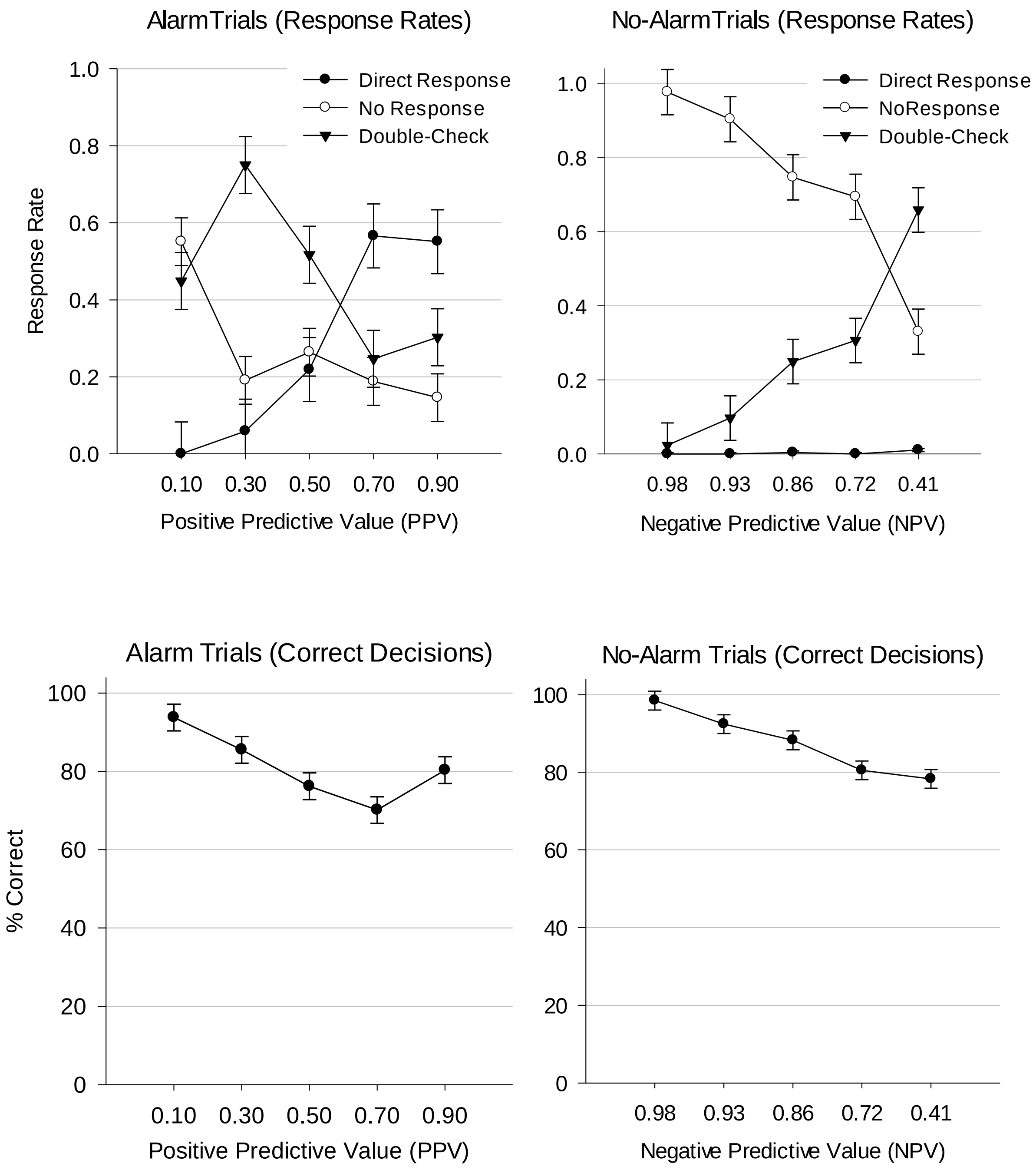

Alarm Trials (Response Rates)
Perceived PPV / Response Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.10 0.30 0.50 0.70 0.90
Positive Predictive Value (PPV)
Block 1
Block 2
Nonalarm Trials (Response Rates)
Response Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.98 0.93 0.86 0.72 0.41
Negative Predictive Value (NPV)
Block 1
Block 2
AlarmTrials (Response Strategies)
Number of Participants
10
8
6
4
2
0
Positive extreme ("compliance")
Negative Extreme
Mixed
0.10 0.30 0.50 0.70 0.90
Positive Predictive Value (PPV)
Nonalarm Trials (Response Strategies)
Number of Participants
14
12
10
8
6
4
2
0
Positive Extreme
Negative Extreme ("reliance")
Mixed
0.98 0.93 0.86 0.72 0.41
Negative Predictive Value (PPV)
Alarm Trials (Correct Decisions)
% Correct
100
80
60
40
20
0
Block1
Block 2
0.10 0.30 0.50 0.70 0.90
Positive Predictive Value (PPV)
Nonalarm Trials (Correct Decisions)
% Correct
100
80
60
40
20
0
Block 1
Block 2
0.98 0.93 0.86 0.72 0.41
Negative Predictive Value (NPV)


Concurrent Task Performance (ROT)
Numbers of Orders
100
80
60
40
20
0
0.10
0.30
0.50
0.70
0.90
Condition (PPV)
Block 1
Block 2


Alarm Trials (Response Rates)
No-Alarm Trials (Response Rates)
Response Rate
1.0
0.8
0.6
0.4
0.2
0.0
Direct Response
No Response
Double-Check
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
Direct Response
NoResponse
Double-Check
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)


Alarm Trials (Correct Decisions)
No-Alarm Trials (Correct Decisions)
% Correct
100
80
60
40
20
0
Block1
Block 2
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
Block 1
Block 2
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)


AlarmTrials (Response Rates)
No-AlarmTrials (Response Rates)
Direct Response
No Response
Cross-Check
Direct Response
NoResponse
Cross-Check
Response Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)
Alarm Trials (Correct Decisions)
No-Alarm Trials (Correct Decisions)
% Correct
100
80
60
40
20
0
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)


AlarmTrials (Response Rates)
Response Rate
1.0
0.8
0.6
0.4
0.2
0.0
Direct Response
No Response
Double-Check
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
No-AlarmTrials (Response Rates)
1.0
0.8
0.6
0.4
0.2
0.0
Direct Response
NoResponse
Double-Check
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)
Alarm Trials (Correct Decisions)
% Correct
100
80
60
40
20
0
0.10
0.30
0.50
0.70
0.90
Positive Predictive Value (PPV)
No-Alarm Trials (Correct Decisions)
100
80
60
40
20
0
0.98
0.93
0.86
0.72
0.41
Negative Predictive Value (NPV)