Supporting Attention Allocation in Multitask Environments: Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance

Rebecca Wiczorek and **Dietrich Manzey**, Technische Universität Berlin, Berlin, Germany

Objective: The aim of the current study was to investigate potential benefits of likelihood alarm systems (LASs) over binary alarm systems (BASs) in a multitask environment.

Background: Several problems are associated with the use of BASs, because most of them generate high numbers of false alarms. Operators lose trust in the systems and ignore alarms or cross-check all of them when other information is available. The first behavior harms safety, whereas the latter one reduces productivity. LASs represent an alternative, which is supposed to improve operators' attention allocation.

Method: We investigated LASs and BASs in a dual-task paradigm with and without the possibility to cross-check alerts with raw data information. Participants' trust in the system, their behavior, and their performance in the alert and the concurrent task were assessed.

Results: Reported trust, compliance with alarms, and performance in the alert and the concurrent task were higher for the LAS than for the BAS. The cross-check option led to an increase in alert task performance for both systems and a decrease in concurrent task performance for the BAS, which did not occur in the LAS condition.

Conclusion: LASs improve participants' attention allocation between two different tasks and therefore lead to an increase in alert task and concurrent task performance. The performance maximum is achieved when LAS is combined with a cross-check option for validating alerts with additional information.

Application: The use of LASs instead of BASs in safety-related multitask environments has the potential to increase safety and productivity likewise.

Keywords: automation, graded warnings, decision support systems, compliance, trust, safety, signal detection theory

Address correspondence to Rebecca Wiczorek, Institut für Psychologie und Arbeitswissenschaften, Technische Universität Berlin, Sekr. F 7, Marchstr. 12, Berlin, 10587, Germany; e-mail: rwi@zmms.tu-berlin.de.

HUMAN FACTORS
Vol. 56, No. 7, November 2014, pp. 1209–1221
DOI: 10.1177/0018720814528534

INTRODUCTION

Binary Alarm Systems

Today's workplaces of complex human–machine systems, such as control rooms or flight decks, usually require human operators to monitor several different systems while dealing with other concurrent tasks at the same time. Regarding the monitoring tasks, operators are often supported by alarm systems, which alert them in case of critical events.

Ideally, the alarm systems should guide operators' attention allocation between monitoring tasks and concurrent tasks. In alarm-free periods operators may *rely* on the alarm system (Meyer, 2001, 2004) and focus on other ongoing tasks. Yet whenever an alarm goes off they are supposed to *comply* with it (Meyer, 2001, 2004) by allocating their attention to the event they were alerted to and by initiating a proper action. In reality, however, operators are often found to use alarm systems in unintended ways or not at all (Parasuraman & Riley, 1997). The underlying reasons for operators' disuse of alarm systems can be understood by considering the nature of the most widely used system—the *binary* alarm system (BAS).

Modern (binary) alarm systems are very sensitive, but none of them are 100% reliable. Based on a modeling of alarm systems in terms of signal detection theory (SDT; cf. Green & Swets, 1966) the normal operating state (noise) and the presence of a critical event (noise + signal, in the following referred to as signal) can be considered as being represented by two density distributions with a certain overlap (see Figure 1). In this area of uncertainty it is not clear if a value derives from the noise or the signal distribution. Hence, the decision whether or not to generate an alarm depends on the threshold setting of the alarm system. To not miss any critical event, the

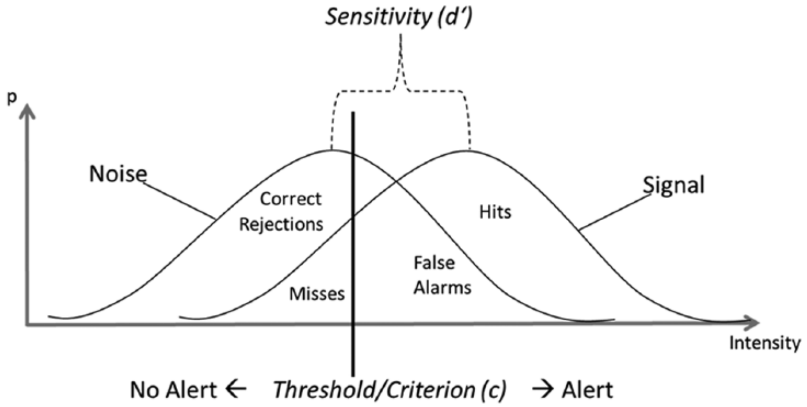


Figure 1. Representation of signal detection theory with the overlapping noise and signal distributions, sensitivity (d'), threshold (c), and the resulting number of correct decisions: hits and correct rejections as well as the resulting number of wrong decisions—misses and false alarms.

system thresholds are often set very liberal (low; Swets, 1992). Yet minimizing the number of misses will inevitably lead to an increase of false alarms (FAs).

Research has shown that not only misses but also FAs can be associated with potentially dangerous consequences. After repeated experiences of FAs operators often perceive the system as unreliable and lose their trust in it, which in turn can result in a reduction of compliance with alarms (e.g., Lees & Lee, 2007; Madhavan, Wiegmann, & Lacson, 2006). This effect is called the *cry wolf phenomenon* (Breznitz, 1984) and becomes manifest in reduced response frequencies or increased response times to alarms (e.g., Bliss, Dunn, & Fuller, 1995; Getty, Swets, Pickett, & Gonthier, 1995).

Recent research has shown that the cry wolf effect can be mitigated by providing participants access to additional information such as raw data that can be used to cross-check the validity of given alarms before responding (Gérard & Manzey, 2010). However, this behavior also entails the risk of unintended side effects. Since cross-checking is time consuming and needs attention, participants might focus too much on the alarm-supported task and thus neglect their other tasks.

Based on this evidence it seems that BASs, although still representing the most used types of alarm systems today, do not support operators' attention allocation in an optimal way. Therefore, so-called *likelihood* alarm systems (LASs) have been proposed as a possible alternative that might circumvent some of the disadvantages of BASs (Sorkin, Kantowitz, & Kantowitz, 1988).

Likelihood Alarm Systems

LASs represent a special case of graded alerts. Most systems, which use graded alerts, monitor the development of *analogue* signals over time. Typical examples are prealarms in the process industry or collision warning systems. They provide additional information regarding urgency or remaining time to collision (e.g., Lee, Hoffman, & Hayes, 2004; Marshall, Lee, & Austria, 2007). LASs, on the contrary, monitor *discrete* signals (e.g., blocked valves) and provide additional information about the *likelihood* of the indicated critical event. Whereas a collision warning system may generate first a warning and later an alarm in case no corrective action occurs and time to collision diminishes, the LAS generates only an alarm *or* a warning, depending on the intensity of the detected signal.

The schematic picture of a SDT model of a three-stage LAS is presented in Figure 2. As becomes evident, the LAS can be modeled as a signal-detection system with two thresholds. The first threshold corresponds to the one of a BAS, separating the nonalert from the alert

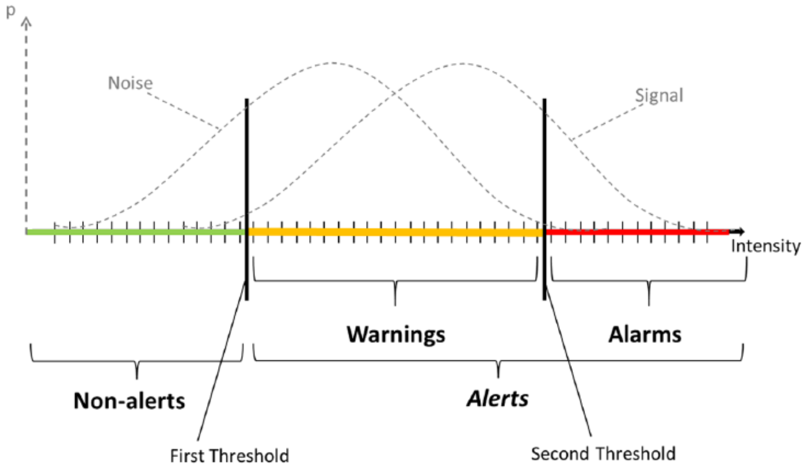


Figure 2. Schematic picture of a three-stage likelihood alarm system with two thresholds and the representation of the underlying noise and signal distributions.

stage. The second threshold then splits the alert stage into two stages, warnings and alarms. As a consequence, alarms have a comparatively high likelihood to truly indicate a critical event, whereas warnings are associated with a considerably higher level of uncertainty, informing the operator that there *might be* a critical event.

One potential advantage of LASs is the decreased number of false *alarms.* Instead, the systems generate a considerably high number of false *warnings*. This might have beneficial effects on operators' trust in the systems. Whereas an FA is usually perceived as a system error, false warnings present a different case. Unlike an alarm, the warning just indicates that there *might be* a critical event. Thus, the absence of the critical event does not necessarily prove false the system's diagnosis. Thus false warnings might not be considered as system errors, and therefore operators perceive the LASs as more reliable and more trustworthy than the BASs. In line with this assumption, Bustamante (2008) presumed that the two types of alerts would also lead to distinct behavioral consequences. Specifically, he expected that operators would respond more often to alarms and less often to warnings compared to their compliance with BAS alerts. This should result in an improved and therefore safer alert task performance. Furthermore, Sorkin et al. (1988) expected LASs to improve operators' attention allocation, especially when operators have access to additional information. If they would not cross-check all of the given alerts but only the warnings and comply directly with the alarms, they could save time and attention resources and invest them in concurrent tasks.

Even though the concept of LASs was proposed more than 25 years ago, comparatively little research is available, thus far, that has addressed the possible performance consequences of LASs. A first set of studies that compared the use of LASs to BASs provided evidence that LASs have the potential to increase alert task performance (Bustamante, 2008; Sorkin et al., 1988). Specifically, LAS users were found to respond more often to true alerts and less often to false alerts, compared to BAS users (Bustamante & Bliss, 2005). This pattern came along with a reduction of overall response rate with alerts for LAS, compared to BAS. Whereas Sorkin et al. (1988) could find the positive effect of LAS over BAS only under conditions of high workload, the results of Bustamante (2008) did not point to such a restriction, as benefits of LAS emerged also under low workload conditions. However, the comparison of these results is complicated by the fact that the LASs used in the two studies differed with regard to both the number of system stages (four vs. three) and the threshold setting, resulting in different numbers of FAs produced by the systems.

Whereas no study exists that investigated the impact of number of stages of LASs two studies did address the relevance of threshold settings in

LASs and BASs. The results of Wickens and Colcombe (2007) suggest that lowering the threshold and thus increasing the number of FAs might result in a stronger increase in reaction time to alerts from the LAS than to alarms from the BAS. Clark, Peyton, and Bustamante (2009) found that an LAS with a liberal first threshold did not improve performance compared to a miss-prone BAS.

Some data of these studies also provide insights into the impact of LASs on attention allocation in multitask environments. However, none of these studies have reported any benefits of LASs compared to BASs in this respect (Sorkin et al., 1988; Wickens & Colcombe, 2007). That seems to be somewhat surprising given the finding of Bustamante and Bliss (2005) that LASs can reduce the overall response rates to alerts and thus free resources that principally could be invested in other tasks. Finally, Bustamante (2008) also investigated the possible effects of the combination of cross-check option and LAS in his study. He found that BAS users' performance improved with access to additional information, whereas LAS users could not benefit from the cross-check option.

However, some of the reported results should be interpreted carefully. For example, the study of Sorkin et al. (1988) included only four participants and thus represents a rather explorative approach with somewhat limited conclusiveness. The alarm system used in the air traffic control study of Wickens and Colcombe (2007) alerted participants of possible conflicts by indicating different distances between the airplanes. It can be questioned whether such alarm system really can be considered an LAS. Indicating distances seems to be much more similar to the graded alerts described above known from automatic monitoring of dynamic systems than to LASs. Finally, some concerns can also be raised regarding the findings of Bustamante (2008). In this study, LAS users did not benefit from the cross-check option, which was realized through the display of gauge deviations. This additional information did not reduce uncertainty to 0% but only to a minimum of 12%. Therefore, it is possible that LAS users avoided the extra effort of cross-checking because the alarm system itself already provided sufficient additional information for a proper response selection.

The Current Study

The objective of the current study was to compare the human performance consequences of BASs and LASs dependent on whether or not a cross-check option for validation of given alerts was available. For this purpose a laboratory multitask was used that included two tasks, supported by either a BAS or LAS. The systems compared had the same first threshold, separating the nonalert and the alert stage; that is, both generated the same amount of *alerts* in total. In the LAS condition, a second threshold divided the alert stage. Whereas all alerts in the BAS condition represented *alarms*, most alerts in the LAS condition represented *warnings*, and only a few represented *alarms.* Performance consequences addressed in the study included participants' trust in the systems, their response behavior, and their performance in both tasks. Our hypotheses were as follows:

Hypothesis 1: With respect to subjective trust in the different systems, it was assumed that participants would trust LAS more than BAS independent of whether or not a cross-check option is available.

Hypothesis 2: With respect to the overall compliance with alerts (alarms for BASs, warnings + alarms for LASs), a complex pattern of effects was assumed, reflected in an interaction between type of alarm system and cross-check option. For the condition *without* cross-check, it was expected that LAS would lead participants to overall comply less often with alerts than participants in the BAS condition. This should result from the assumptions that (a) LAS users' compliance with *alarms* will be higher than BAS users' compliance with *alerts* and (b) LAS users' compliance with *warnings* will be lower than BAS users' compliance with *alerts*. For conditions *with* the cross-check option, a reverse pattern of effects was assumed. It was expected that BAS users would cross-check most alarms prior to responding. In contrast, LAS users should comply directly with most of the alarms and restrict cross-checking to warnings. Overall this will lead to a higher rate of *direct* compliance with *alerts* in the LAS compared to the BAS condition.

Hypothesis 3: With respect to alert task performance, it was assumed that an interaction between the type of alarm system and the cross-check option would occur. LAS will lead to fewer wrong decisions in response to alerts than BAS, *without* cross-checking, because the LAS provides additional information and thus offers a better basis for decision making. *With* a cross-check option, however, no differences should emerge since users of BAS and LAS will cross-check most alarms and most warnings, respectively, which should reduce the number of wrong decisions in both conditions.

Hypothesis 4: Finally, a general performance benefit of LAS compared to BAS with respect to concurrent task performance was expected. This will result from participants working with LASs responding or cross-checking fewer alerts in conditions without and with the cross-check option, respectively (see Hypothesis 3).

METHOD

Participants

A total of 60 participants (30 females, 30 males; average age = 26.74) with normal or corrected-to-normal vision and without reported color vision deficiency were randomly assigned to one of four experimental conditions. Participants were paid €10 ($13) and could receive an additional performance-related bonus of up to €10. On average, participants received €17 ($22).

Task Environment

The PC-based laboratory environment Multi-Task Operator Performance Simulation (M-TOPS) was used for the experiment. M-TOPS simulates cognitive task demands similar to those of control room operators in chemical plants. A picture of the experimental screen is shown in Figure 3.

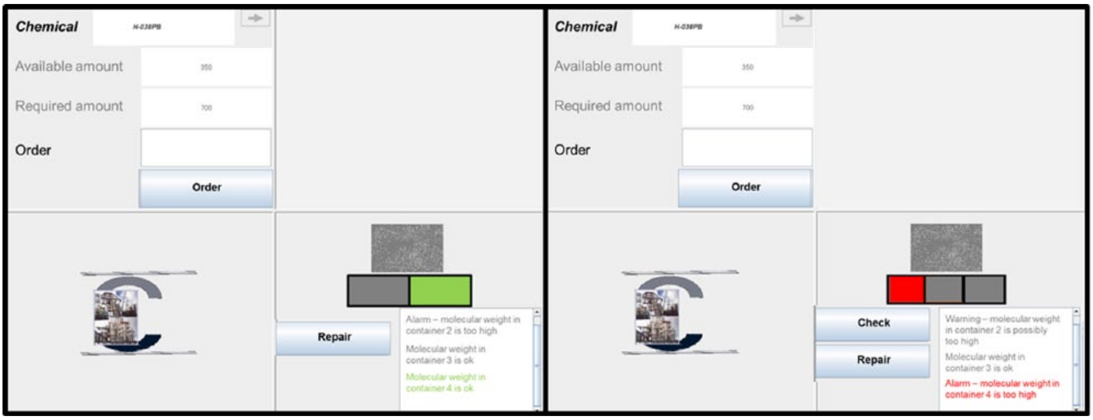
Ordering task. In the ordering task (upper-left part of the interface; see, Figure 3), participants have to ensure the availability of required chemicals to keep the chemical process running. Therefore, they have to calculate the difference between the actual and the set value, type the result in the ordering field, and send it by clicking a button within 15 s. Afterward the next order appears automatically.

Alert task. In this task, which simulates a quality control (lower-right part of the interface; see Figure 3), participants are supported by either a BAS or a LAS. Every 8 s, one container holding the chemical end product is automatically checked by the alarm system, which generates a visual diagnosis regarding the appropriateness of the molecular weight. No auditory alert information is provided, as alert modality does not seem to play a crucial role in the comparison of BASs and LASs (Sorkin et al., 1988; Wickens & Colcombe, 2007). The systems' diagnoses are indicated in a color display (with two or three panels for the BAS and the LAS, respectively) under the container and as redundant information in an alert state monitor on the right side under the display. A green light in the display and the announcement "molecular weight is ok" in the monitor indicate the absence of an alert, whereas a red light and the announcement "molecular weight is too high" are used to alarm participants. These two states are identical for BAS and LAS. The LAS has an additional warning stage with an amber light and the announcement "molecular weight is possibly too high."

In the condition without the cross-check option participants have to decide whether or not they want to respond to a diagnosis by clicking the "repair" button to fix the molecular weight. In the cross-check condition, participants also have the possibility of validating the system diagnoses by clicking the "check" button. Raw data of the container are displayed as a colored picture (see Figure 4). When it shows 15 green marks on red ground, the container is okay, whereas 16 marks imply a faulty container. Participants can click either the "repair" or the "continue" button. Then the picture is closed and the process continues.

Payoff Matrix

Participants received 1.5 points for every correct order in the ordering task and lost 2 points for every wrong decision in the alert task (e.g.,



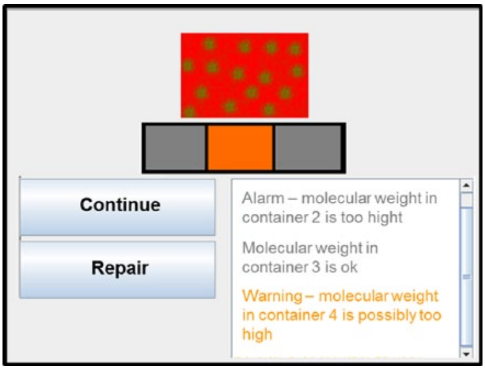
Chemical H-038PB
Available amount 350
Required amount 700
Order

Alarm – molecular weight in container 2 is too high
Molecular weight in container 3 is ok
Molecular weight in container 4 is ok

Chemical H-038PB
Available amount 350
Required amount 700
Order

Warning – molecular weight in container 2 is possibly too high
Molecular weight in container 3 is ok
Alarm – molecular weight in container 4 is too high

Figure 3. Two screenshots of the M-TOPS task environment with the ordering task on the upper-left side, the alert task on the lower-right side, and the M-TOPS logo ("C" for M-TOPS "chemical plant") on the lower-left side. The left part of the figure shows the condition with binary alarm system and without the cross-check option. The latest announcement in the alarm state monitor verbalizes the actual information whereas the prior announcements are displayed above. The right part of the figure shows the condition with likelihood alarm system and cross-check option.



Alarm – molecular weight in container 2 is too hight
Molecular weight in container 3 is ok
Warning – molecular weight in container 4 is possibly too high

Figure 4. M-TOPS screenshot with raw data of a container being not faulty (condition with the likelihood alarm system and cross-check option).

repairing a container that was okay or not repairing a faulty container). This payoff with gains in the concurrent task and losses in the alert task was chosen to produce a competition between both tasks, as is the case in real-world settings.

Alarm System Configuration

In safety-related work environments, base rates of critical events are usually very low (Parasuraman, Hancock, & Olofinboba, 1997). In research settings, we need higher base rates to not extend the experiment duration. To deal with this trade-off, the alert task used in our study was declared as quality control, which is also in real-world settings characterized by more moderate base rates. Base rates (pS), sensitivity (d'), and (first) threshold (c) values were identical for the BAS and LAS and calculated according to SDT formulas (cf. Green & Swets, 1966). The LAS had an additional threshold (c*), which was calculated using the formula for c. The first threshold was liberal, as is common for BASs (Swets, 1992). The second threshold was chosen to generate the same percentage of true (88%, i.e., 14) and false (12%, i.e., 2) alarms as in Bustamante's (2008) study. System configurations and resulting outcomes are shown in Tables 1 and 2.

Design

The study consisted of a 2 × 2 between-subject design. The two factors were "type of alarm system" with BAS versus LAS and "cross-check" with versus without such option. The dependent variables were defined as follows.

Trust ratings. Participants indicated their trust toward the entire alarm system on a visual

TABLE 1: Configurations of SDT parameters for the Binary Alarm System (BAS) and the Likelihood Alarm System (LAS)

	BAS	LAS
Sensitivity (d′)	1.8	1.8
Criterion (c)	–1.05	–1.05
Criterion (c*)	—	0.29
Base-rate (pS)	0.3	0.3

analog trust scale, in the form of one line containing no dimension units but five verbal anchors ranging from *my trust is very strong* to *I barely trust the system*. Answers were assessed in cm and transformed into trust dimensions ranging from 0 to 100. Trust was assessed before and after the experimental block to see if it remains stable over time.

Behavioral data. Compliance was defined as clicking the "repair" button in response to alarms or warnings. Checking was defined as clicking the "check" button in response to alarms or warnings. Compliance and checking rates were the frequencies of compliance or checking, respectively, qualified by the total number of diagnoses of the specified category: alarm, warning, or alert.

Performance data. Performance in the alert task was operationalized as the sum of wrong decisions participants made in interaction with the alarm system (e.g., repairing an intact container [FA] or not repairing a faulty container [miss]). Performance in the ordering task was the number of correct orders.

Procedure

After filling in a demographic questionnaire, participants read the instructions. They were told to be responsible for the ordering task and the alarm task, and that both tasks were equally important. Participants then practiced the two tasks separately for 2 min each. Afterward they completed a 60-trial block with the alert task only to be familiarized with alarm system characteristics. During this block they received auditory feedback whenever they made a wrong decision to provide information about the likelihood of alarms and warnings to be correct. Participants were told that the alarm system would not be perfectly reliable and that they should use the training block to gain experience with its reliability. After this they filled in the first trust questionnaire. The following experimental block consisted of 100 trials (approx. 15 min). Participants had to work on both tasks simultaneously without on-line feedback. A visual feedback of achieved points was given subsequently. After the experimental block, participants filled in the second trust questionnaire.

RESULTS

Statistical analyses were done with SPSS 12.0. The level of significance was .05. We used two-way ANOVAs with the alarm system and cross-check factors for the analyses of response rates and performance data, a three-way ANOVA with repeated measures for the analysis of trust ratings, and *t*-tests for the single comparisons (the significance level was not adjusted as they were defined a priori).

TABLE 2: Number of Resulting Outcomes of the Binary Alarm System (BAS) and the Likelihood Alarm System (LAS) for a Block of 100 Trials

		BAS			LAS		
		Signal	No Signal	Total	Signal	No Signal	Total
Nonalerts		**1**	31	32	**1**	31	32
Alerts	Alarms	29	**39**	68	14	**2**	16
	Warnings	—	—	—	15	**37**	52
				100			100

Note. Misses, false alarms, and false warnings are indicated by bolding.

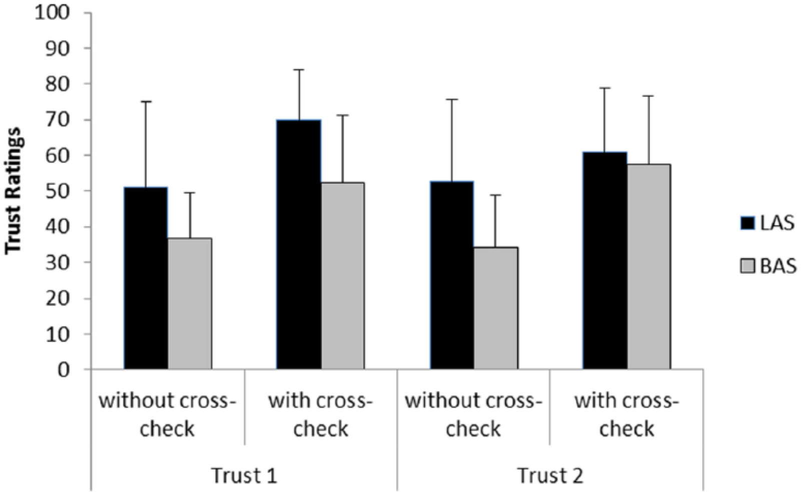


Figure 5. Means and standard deviations of trust ratings from both times of trust measurement for likelihood alarm systems (LASs) and binary alarm systems (BASs) in the conditions with and without the cross-check option.

Trust

Because of missing data, only 58 of the 60 participants were included in the statistical analysis of trust. Figure 5 shows the trust ratings for LASs and BASs under the conditions without and with cross-check, separated for the two times of assessment. As expected, ratings were higher for the LASs than for the BASs in both cross-check conditions, resulting in a main effect for alarm system, $F(1, 54) = 9.92, p = .003, \eta^2_p = .16$. Furthermore, it was found that validating a given alert had an effect on participants' trust ratings. When cross-checking was possible, BAS and LAS users trusted their systems more than without such option, leading to a cross-check main effect, $F(1, 54) = 14.67, p < .001, \eta^2_p = .21$. No significant difference could be found between the first and the second measurement, $F(1, 54) = 0.3$, *ns*; however, there was a significant triple interaction, $F(1, 54) = 4.22, p = .045, \eta^2_p = .07$, indicating a different development of the trust in both systems only when cross-checking was possible. Trust in both systems remained stable over time in the condition without cross-check. In the condition with cross-check, trust in BAS slightly increased, whereas a decrease of trust in LAS was found. None of the other interactions became significant.

Response Rates

Compliance rates with alerts. Compliance with alerts is presented in Figure 6. The figure shows that direct compliance occurred less often when a cross-check option was available, cross-check main effect, $F(1, 56) = 75.62, p < .001, \eta^2_p = .58$. In the condition without cross-check, compliance with alerts was lower in the LAS (47%) than in the BAS (67%) condition. As expected, the reverse picture was found when cross-checking was possible. LAS users complied with 16% of the alerts, whereas BAS users complied with only 1%. A significant interaction, $F(1, 56) = 9.59, p = .003, \eta^2_p = .15$, confirms these findings. The main effect for alarm system, $F(1, 56) = 0.27$, *ns*, was not significant.

Checking rates and compliance rates with alarms and warnings. The analysis of checking and compliance with alarms and warnings was done in a descriptive way. In addition, we made two single comparisons regarding compliance with BAS alerts versus LAS alarms and BAS alerts versus LAS warnings.

Figure 7 shows the proportion of the different behavioral options, compliance in the condition without cross-check and compliance and checking in the cross-check condition. Without the cross-check option, BAS users complied with

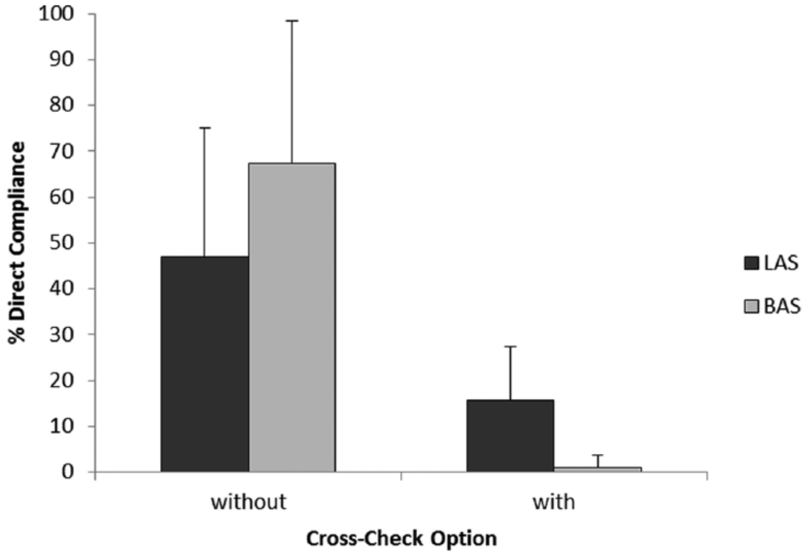


Figure 6. Means and standard deviations of compliance with binary alarm system (BAS) and likelihood alarm system (LAS) alerts in the conditions with and without the cross-check option. (Note that alerts were alarms for BASs, warnings + alarms for LASs).

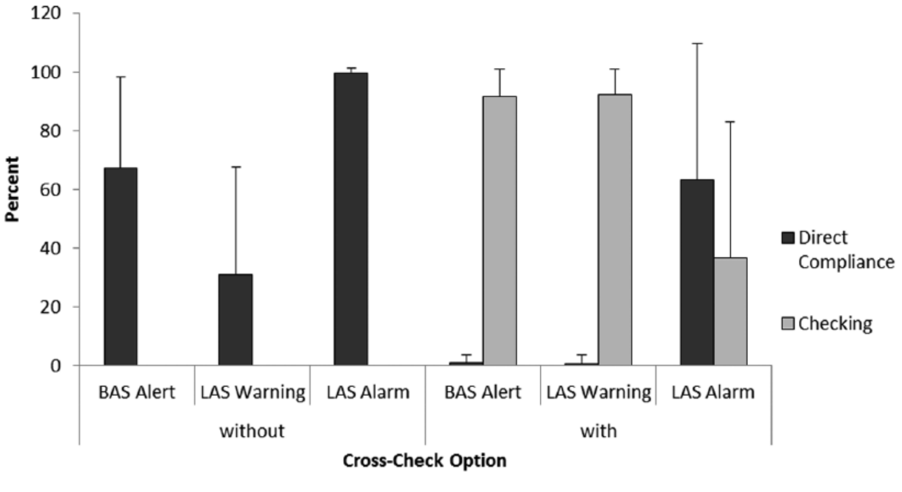


Figure 7. Means and standard deviations of checking rates and compliance rates for binary alarm system (BAS) alerts, likelihood alarm system (LAS) warnings, and LAS alarms separately in the conditions with and without the cross-check option. The percentages refer to the number of events of the specified category (e.g., responding to all of the 16 alarms in the LAS condition represents a compliance rate of 100%).

70% of the alarms, whereas LAS users complied with 99% of the alarms but only with 31% of the warnings. A priori defined single comparisons of LAS alarms with BAS alerts, $t(28) = -4.01$, $p < .001$, and LAS warnings with BAS alerts, $t(28) = 2.67$, $p < .05$, confirm that LAS users differentiated their behavior toward the two types of alerts in the expected way. They responded more often to alarms and less often to warnings compared to BAS users' compliance

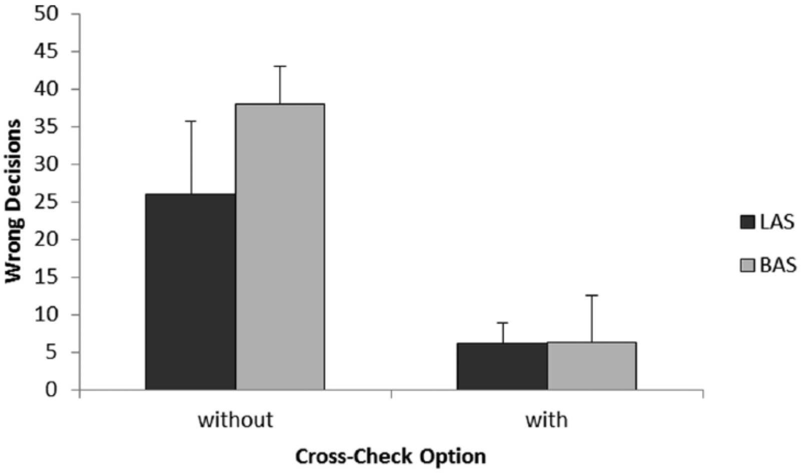


Figure 8. Means and standard deviations of alert task performance for binary alarm systems (BASs) and likelihood alarm systems (LASs) in the conditions with and without cross-check option.

with alerts. A differentiation was found also in the condition with the cross-check option. BAS users checked 92% of the alarms and complied with only 1%. Exactly the same response rates were found for LAS warnings. In contrast, only 37% of LAS alarms were checked and participants complied with 63% of the alarms.

Performance

Figure 8 presents alert task performance in terms of wrong decisions, that is, repaired intact containers (FAs) and unrepaired faulty containers (misses), for both types of alarm systems under the two cross-check conditions. As can be seen from the graph, the number of wrong decisions was much higher without a cross-check option, resulting in a significant main effect for cross-check, $F(1, 56) = 239.57$, $p < .001$, $\eta^2_p = .81$. Furthermore, the use of LAS as compared to BAS led to significant fewer wrong decisions, resulting in a main effect for alarm system, $F(1, 56) = 13.21$, $p = .001$, $\eta^2_p = .19$. However, this main effect is further qualified by the significant interaction effect, $F(1, 56) = 12.34$, $p = .001$ $\eta^2_p = .18$. The difference between BAS and LAS emerged only without the cross-check option, and no difference was found when cross-checking was possible.

Figure 9 presents ordering task performance in terms of correct orders for both systems under the two cross-check conditions. As can be seen, ordering task performance was significantly higher without than with the cross-check option available, $F(1, 56) = 4.16$, $p = .046$, $\eta^2_p = .07$. More interesting, the use of LASs generally led to significantly improved performance compared to the BASs, which emerged irrespective of cross-check availability, $F(1, 56) = 7.011$, $p = .011$, $\eta^2_p = .11$. No interaction effect was found between the two factors, $F(1, 56) = 0.35$, *ns*.

DISCUSSION

The aim of the current study was to examine the effects of different alarm systems and the availability of additional information on participants' trust, behavior, and performance. Therefore, BASs and LASs were compared in a dual-task paradigm with and without a cross-check option.

Our first hypothesis was partly confirmed as LAS users reported higher trust in the system than BAS users, which remained stable over time. This finding supports the assumption that the experience of FAs affects the perceived reliability of an alarm system considerably more than the experience of false warnings. However,

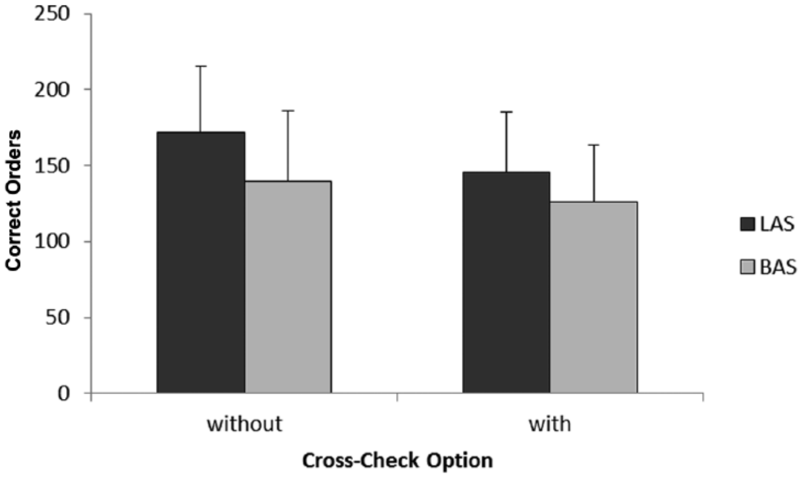


Figure 9. Means and standard deviations of ordering task performance for binary alarm systems (BASs) and likelihood alarm systems (LASs) in the conditions with and without the cross-check option.

results of the cross-check condition were rather unexpected. When cross-checking was possible, trust ratings not only were higher than without cross-checking, but also developed differently for the two types of alarm systems. Whereas BAS users' trust increased over time, LAS users' trust decreased while they were working with the system. This, on first sight, surprising effect might be due to the fact that BAS users cross-checked most of the alarms, whereas LAS users cross-checked most of the warnings but only fewer than half of the alarms. As the BAS alarms had a higher likelihood to be true than the LAS warnings, the perceived reliability might have been reduced more for the LAS than for the BAS. An alternative explanation refers to the possibility of complete uncertainty reduction. As uncertainty and vulnerability represent the preconditions for trust decisions (Lee & See, 2004), it is possible that their absence has led to the abolition of the trust concept in this specific situation. Further research should assess participants' perceived reliability and their actual level of uncertainty to answer this question.

The analysis of compliance with alerts confirmed our second hypothesis regarding the interaction of type of alarm system and cross-check availability. In the condition without cross-check, overall compliance with alerts (warnings and alarms) was lower for LAS (47%) as compared to BAS (67%). A similar effect has also been reported by Bustamante and Bliss (2005). However, when cross-checking was possible, the opposite result was found, as overall compliance with alerts was higher in the LAS condition (16%) than in the BAS group (1%). This finding can be explained by the behavioral differentiation of LAS users toward the two types of alerts.

Without cross-check, LAS users' compliance with alarms (99%) was higher and their compliance with warnings (31%) was lower than was BAS users' compliance with alerts (67%), confirming Hypotheses 2a and 2b. The insufficient compliance with BAS alerts can be interpreted as a classical cry wolf effect, which has been found in several studies before (e.g., Bliss et al., 1995). The LAS, for its part, led not to the elimination of but rather to a shift of the cry wolf effect from the alarm stage into the warning stage. This explains the reduction of overall compliance with *alerts*, given that the LAS generated 52 warnings and only 16 alarms.

When cross-checking was possible, BAS users validated nearly every alert (92%) and complied with only 1%. This finding is in line with earlier research of Gérard and Manzey (2010). The LAS users showed a similar behavior in response to warnings, but checked only 37% of the alarms and complied directly with the remaining 63%. The behavioral differentiation of LAS users toward alarms and warnings had positive effects on the performance in both the alert task and the ordering task.

For alert task performance we found an interaction effect, confirming our third hypothesis. Differences between the two types of alarm systems emerged only when no cross-check was available. In line with previous findings (e.g., Bustamante, 2008), performance in the alert task was significantly better for LAS than for BAS in the condition without cross-check. Participants had higher response rates to alarms, which had a high likelihood to be true, and lower response rates to low likelihood warnings, therefore reducing the number of misses and FAs likewise. This behavioral pattern corresponds to earlier findings of Bustamante and Bliss (2005). When cross-checking was possible, BAS and LAS users made equally fewer wrong decisions than without the cross-check option. In this study, unlike the findings of Bustamante (2008), users of both types of alarm systems could benefit from the availability of cross-check information.

An even more interesting result, however, is the increased concurrent task performance of LASs over BASs, which has not been shown before but confirms our fourth hypothesis. This result has been found for both conditions with and without cross-check. We assume that LAS users achieved a higher performance because they had more free time and attentional resources available for this task than did participants of the BAS group. In the condition without cross-check, the additional resources resulted from LAS users' reduced compliance with alerts. Fewer interactions required less time and attention. In the condition with cross-check, the validation of raw data was more time-consuming than was the direct compliance. As LAS users complied more often directly, they had more remaining resources for the concurrent task. These results suggest that attention allocation was supported by LASs in both conditions with and without cross-check option, as the increase in ordering task performance was not at the expense of alert task performance.

We think that some general conclusions can be drawn from the findings of this experiment. First, it seems that users' subjective perception of alarm systems does not only depend on the actual reliability of the system, as participants' trust increased when additional likelihood information was provided. Second, users are apparently able to make use of the additional information for their behavioral calibration, as participants responded more often to high likelihood alarms and less often to low likelihood warnings. The most important theoretical implication of the current results is that users' behavior can be guided in a meaningful way through the use of multistage approaches in alarm systems. As a practical implication, we consider the potential of LASs for mitigating the trade-off between safety and productivity through an improved attention allocation.

LIMITATIONS

Even though the results of this study provide new insight regarding the behavioral effects of LASs and the resulting performance, some limitations have to be considered. One critical issue regards the cross-check option used in this experiment. The frequent use of the cross-check possibility could be explained by the fact that it was neither very time-consuming nor difficult. Furthermore, the artificial task allowed the complete reduction of uncertainty. In real working environments, such as the process industry, the access to additional information is more challenging and the interpretation of data often rather complicated. Further research is needed to investigate the influence of availability and information content on the use of cross-check options. It also should be noted that the base rate used here does not correspond to real-world base rates of critical events in safety-related environments. Furthermore, it has to be mentioned that most of the alarm systems used in the process industry monitor analog signals rather than discrete ones. The other critical point regards the threshold setting of the LAS. The threshold in our experiment was the same as used by Bustamante (2008). Therefore it is not clear whether or not the results can be generalized also to LASs with different thresholds. Additional studies that systematically investigate effects of threshold setting in LASs have to be conducted.

ACKNOWLEDGMENTS

We thank Mandy Dotzauer, Ulrike Wiedensohler, and Frank Bolton for proofreading. Furthermore, special thanks are due to Marcus Bleil for programming the M-TOPS environment and implementing all adaptations needed for the present experiment.

KEY POINTS

- We compared a binary alarm system (BAS) with a likelihood alarm system (LAS) under conditions with and without the possibility to validate alarms via cross-check.
- We assessed participants' trust ratings, response rates, and performance in the alert task and the concurrent task.
- Compared to BAS users, we found LAS users to have more trust in the system, to comply more often with alarms, and to perform better in the concurrent task
- The cross-check option led to better performance in the alert task for BASs and LASs but reduced concurrent task performance in the BAS condition, which did not occur in the LAS condition.

REFERENCES

Bliss, J. P., Dunn, M. C., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills*, *80*, 1231–1242.

Breznitz, S. (1984). *Cry wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum.

Bustamante, E. A. (2008). Implementing likelihood alarm technology in integrated aviation displays for enhancing decision-making: A two-stage signal detection modeling approach. *International Journal of Applied Aviation Studies*, *8*, 241–261.

Bustamante, E. A., & Bliss, J. P. (2005). Effects of workload and likelihood information on human response to alarm signals. In *Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 81–85). Oklahoma City, OK: Wright State University.

Clark, R. M., Peyton, G. G., & Bustamante, E. A. (2009). Differential effects of likelihood alarm technology and false-alarm vs. miss prone automation on decision making. In *Proceedings of the Human Factors and Ergonomics Society annual meeting* (pp. 349–353). Santa Monica, CA: Human Factors and Ergonomics Society.

Gérard, N., & Manzey, D. (2010). Are false alarms not as bad as supposed after all? A study investigating operators' responses to imperfect alarms. In D. de Waard, A. Axelsson, M. Berglund, B. Peters, & C. Weikert (Eds.), *Human factors. A system view of human, technology and organisation* (pp. 55–69). Maastricht, Netherlands: Shaker.

Getty, D., Swets, J. A., Pickett, R. M., & Gonthier, D. (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.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley.

Lee, J. D., Hoffman, J. D., & Hayes, E. (2004). Collision warning design to mitigate driver distraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 65–72). New York, NY: ACM.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

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

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*, 241–256.

Marshall, D. C., Lee, J. D., & Austria, P. A. (2007). Alerts for in-vehicle information systems: Annoyance, urgency, and appropriateness. *Human Factors*, *49*, 145–157.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*, 563–572.

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

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

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

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*, 522–532.

Wickens, C. D., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors*, *49*, 839–850.

Rebecca Wiczorek works as a postdoc researcher at the Technische Universität Berlin in Germany and earned her PhD in human factors from the Technische Universität Berlin in 2012.

Dietrich Manzey is professor in the Department of Psychology and Ergonomics at the Technische Universität Berlin. He earned his PhD in psychology from the University Kiel in Germany in 1988.

Date received: May 3, 2013
Date accepted: February 20, 2014