Investigating Benefits of Likelihood Alarm Systems in Presence of Alarm Validity Information

Rebecca Wiczorek, Magali Balaud & Dietrich Manzey
Technische Universität Berlin, Germany

Providing operators additional information helping them to validate alarms has been found to be a countermeasure for problems related to the cry wolf effect (i.e., operators ignoring alarms). Adding information can be realized with likelihood alarm systems (LAS) or with access to alarm validity information (AVI). The two studies presented here examined behavior and performance consequences of the combination of LAS and AVI in multi-task settings. It was investigated to what extent concurrent task performance and alert task performance depend on characteristics of the LAS (i.e. proportion of different alert types) and cost of cross-checking AVI. Results suggest that those LAS characteristics varied here do not influence participants' performance. Secondly, no benefit of LAS over binary alarm systems (BAS) emerged when increasing the cost of accessing AVI. Results are further discussed with regard to participants' response patterns.

INTRODUCTION

Alarm systems are a basic form of automation designed to inform users about critical events. Unfortunately, they often produce false alarms, which incorrectly direct the users' attention to events that actually do not present critical states. One reason for high numbers of false alarms is that alarm systems base their decisions on data, which can be ambiguous (Swets, 1992) or evolves over time (Thomas et al., 2003). Furthermore, they use sensors that vary in their ability to detect and analyze critical events (Sorkin & Woods, 1985). A commonly used indicator describing the reliability of an alarm system is the Positive Predictive Value (PPV). PPV is the conditional probability that, given an alarm, a critical event actually exists. In other words, it reflects the ratio between true alarms and the total number of times an alarm went off [true alarms/(true alarms+false alarms)]. The PPV is of great interest for human-alarm interaction researchers because it corresponds to users' mental representation of the system reliability and therefore has an impact on their behavior (e.g. Getty, Swets, Pickett & Gonthier,, 1995; Manzey, Gerard & Wiczorek, 2014). The higher the PPV, the more operators usually comply with the given alarms. However, most systems emit a lot of false alarms and are thus, characterized by low PPVs. For example, in the medical domain 80% to 99% of the alarms produced by systems are false (e.g. Lawless, 1994). Two main reasons explain these high false alarm rates. First, engineers tend to follow an approach called "fail-safe engineering" (Swets, 1992) so that alarm systems emit an alarm even with little evidence for a critical event. Secondly, base rates of critical events are low in most settings giving even highly reliable systems only few possibilities for true alarms (Parasuraman, Hancock & Olofinboba, 1997).

High false alarm rates are actually a well-known problem when using alarm systems because they can lead to the so called cry wolf effect (Breznitz, 1984). That is, when experiencing a low PPV, operators might lose trust in the system (Madhavan, Wiegmann & Lacson, 2006) and, thus, react slower to the emitted alarms (Getty et al., 1995) or ignore alarms partially or completely (Bliss, Gilson & Deaton, 1995). Such behaviors present severe risks because they can result in losses of safety and productivity (e.g. Edworthy, 2013).

One suitable solution to reduce or eliminate the cry wolf effect is to provide operators additional cues increasing their ability to differentiate between true and false alarms. The use of likelihood alarm systems (LAS) is an approach that has been proposed as a promising alternative to binary alarm systems (BAS) in this respect (Sorkin, Kantowitz & Kantowitz, 1988). Instead of emitting only one type of alert as BAS usually do, LAS generate different types of messages depending on the *likelihood* that a critical event is actually present. In other words, LAS are composed by two or more alert stages each of them with a different PPV. Colors or labels can be used to represent different alert stages. For example, LAS could use red and the label 'alarm' to indicate high likelihood alerts with high PPV and amber and the label 'warning' for low likelihood alerts with low PPV. These design characteristics are based on findings from studies that investigated the effects of colors, sounds or wording on operators' perception of the hazardousness of an alarm (e.g. Braun & Silver, 1995; Chapanis, 1994). Other studies provide evidence for potential benefits of LAS over BAS in terms of decision making and performance (e.g. Bustamante & Bliss, 2005; Ragsdale, Lew, Dyre & Boring, 2012; Wiczorek, Manzey & Zirk, 2014). They show that operators can use the extra information provided by LAS to adapt their response behavior to each stage in a way that increases the probability to comply with true alarms while it reduces the probability to comply with false alarms.

It has also been found that the availability of alarm validity information (AVI) reduces the cry wolf effect in BAS. When participants have the possibility to cross-check whether an alarm is correct or not they validate almost every alarm and reduce wrong decision making to a minimum (Manzey, Gérard & Wiczorek, 2014). Whether LAS would also lead to performance benefits compared to BAS when providing AVI was investigated recently by Wiczorek & Manzey (2014). No performance benefits of LAS compared to BAS emerged in the alert task when AVI was available. Wrong decision making was reduced to a minimum with both systems.

However, using a multi-task environment, benefits of LAS with respect to concurrent task performance could be identified. Overall, participants supported with LAS used AVI less frequently for cross-checking the outputs emitted by the alarm system than BAS users. More precisely, LAS users did cross-check all emitted warnings but directly complied with the majority of alarms. Because cross-checks consumed time and attentional resources, this behavior led to higher performance in a concurrent task. Such effect was already suggested by Sorkin et al. (1988) but never before shown empirically.

These results raise two questions regarding the combination of LAS and AVI. First, it would be interesting to know whether advantages in concurrent task performance depend on the actual proportions of alarms and warnings of the LAS. If that is true, a relative increase of alarms and decrease of warnings should be beneficial for concurrent task performance while a relatively higher number of warnings and lower number of alarms should reduce this potential benefit of LAS. Second, it is also of interest to investigate what happens when availability of AVI becomes more costly. It has been shown that an increase in required effort reduces cross-checking in BAS (Manzey et al., 2014). In contrast to BAS, LAS allow a more precise reduction of check frequencies. That is, participants can decide to reduce their use of AVI only in case they get a warning, knowing that the warning-PPV is low and therefore reduce the probability to miss a critical event. BAS users, on the other hand, are not able to use any additional likelihood information to decide which alarms do not need to be checked. As a consequence, alert task performance with high-cost AVI should be worse for BAS than LAS because participants working with the latter one would miss less critical events.

Two experiments have been designed in order to answer these questions. The first experiment aimed to investigate how a variation of alarm and warning frequencies of three different LAS would affect participants' performance in one, respectively two concurrent tasks. Participants were expected to comply directly with the majority of alarms and to check almost every warning. Thus, it was assumed, that concurrent task performance would increase with a decreasing number of warnings. The second experiment was conducted in order to test the influence of AVI checking cost on alert task performance with BAS and LAS. The effort and cost required from participants to check AVI was greater than in the experiment of Manzey et al. (2014) in order to obtain larger effects on participants' behavior and performance. Benefits of LAS over BAS were expected for the condition with high-cost AVI, while no differences should emerge when cross-checking required only low effort. Participants were expected to reduce checking frequencies for both systems. While BAS users could only reduce overall checking, participants working with LAS could benefit from the differentiated information offered by LAS. They could therefore reduce their checking behavior more purposeful, i.e. check only when warnings are presented. This difference with regard to checking behavior should lead to a stronger decrease in alert task performance when using BAS compared to LAS with high-cost AVI.

METHOD EXPERIMENT 1

Participants

Sixty-two master students (32 females, mean age: 26.24) participated and were randomly assigned to one of three conditions.

Task Environment

The PC-based laboratory simulation environment M-TOPS (Multi-Task Operator Performance Simulation), which represents a simulation of cognitive requirements of control room operators in chemical plants, was used for the experiment (see Figure 1). M-TOPS consists of three tasks: the objective of the *Resource Ordering Task* (ROT, in the upper left side) is to order certain amounts of chemicals. Therefore, participants calculate the difference of a current and a required amount of chemicals and order it by clicking the 'send' button within 15 seconds. A new task appears automatically. The aim of the *Coolant Exchange Task* (CET, upper right side) is to keep the chemical process at the right temperature. Therefore, participants have to exchange coolant by regulating warm and cold water supply. They do so by opening the valves in a predefined manner. This task includes dead times. When finishing with the actual set, a new set of coolants appears automatically.

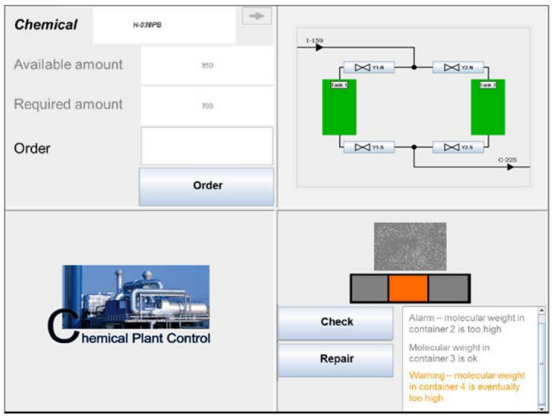

Figure 1. User interface of M-TOPS

In the *Alert Task* (AT, lower right side), participants have to control the quality of the chemical end-product by checking its molecular weight. In that task they are supported by an automatic alarm system. Containers enter one by one in the control station. The alarm system generates a green signal and states 'molecular weight ok' to inform participants that everything is normal; it generates a red signal and states 'molecular weight too high' to inform that there is a problem (low quality); and it generates an amber signal and states 'molecular weight possibly to high' to inform that there might be a problem. After receiving the visual diagnose from the alarm system, participants can decide whether they want to fix the problem directly by clicking the 'repair' button, to access further information by clicking the 'check' button, or to not engage in any action and let the container pass. When clicking the 'check' button a picture of the container content is presented: a certain number of green marks on a red background. Fifteen marks mean the molecular weight is ok. If the number is 16, the weight is too high and needs to be fixed. After checking, participants can either click the 'repair' button or click the 'proceed' button and let the container pass. Responses of the participants are logged.

Every correct order in the ROT was rewarded with 1.5 points, every set of coolants with 7.5 points and every wrong decision in the AT was penalized with -2 points. This allocation of points was based on an analysis of the time structure and was chosen to produce a competition between the tasks. Participants were rewarded dependent on the number of points they received in total.

Experimental Design

The experiment consisted of a 3(type of alarm system) x 2(workload) design with repeated measures on the second factor. All alarm systems had a sensitivity d' of 1.8 and emitted the same number of total alerts (68), but differed with regard to the resulting numbers of alarms and warnings (LAS-1: 40 alarms and 28 warnings; LAS-2: 16 alarms and 52 warnings; LAS-3: 9 alarms and 59 warnings). The base rate of critical events was 0.3. In the low workload condition, participants had to perform the AT and the ROT; while in the high workload condition, all three tasks (AT, ROT, CET) had to be performed simultaneously. The order of high and low workload was counterbalanced.

Procedure

In a 1-hour training session, participants were familiarized with the different tasks and the nature of the alarm system. Afterwards, they completed two experimental blocks with 100 containers each. Finally, participants were paid and debriefed.

Measures

Behavior: Direct response frequencies defined as the number of clicks of the 'repair' button without prior cross-check of AVI, and check frequencies defined as the number of clicks of the 'check' button were used as behavioral measures.

Performance: The number of correct decisions, i.e. repairing low quality containers and not repairing intact containers, as performance in the AT, the number of points obtained in ROT and the number of points obtained in CET as measure of concurrent task performance.

RESULTS EXPERIMENT 1

Behavior

Direct response frequencies and check frequencies were analyzed separately with two-way ANOVAs with repeated measures. With regard to direct response frequencies, no significant difference was found between the three LAS. Also, the variation of workload did not lead to significant differences. Comparison of check frequencies between the three LAS did not reveal significance but a difference was found for the two workload conditions $F(1, 59)=20.62$, $p<.0001$. Participants checked less often when the workload was high. None of the possible interaction effects revealed significance.

Performance

The number of correct decisions and the number of points in ROT were analyzed with two-way ANOVAs with repeated measures. The number of points in CET was obtained only for one block and therefore analyzed with a one-way ANOVA. Results of correct decisions did not reach significance with regard to the different LAS or the interaction effect, but only for variation of workload $F(1, 59)=13.85$, $p<.0001$. The same is true for the number of points in ROT, $F(1, 59)=150.87$, $p<.0001$. A decrease in performance occurred for both, the AT and ROT when workload increased due to the additional third task. No differences between the three LAS with regard to number of points in CET were found.

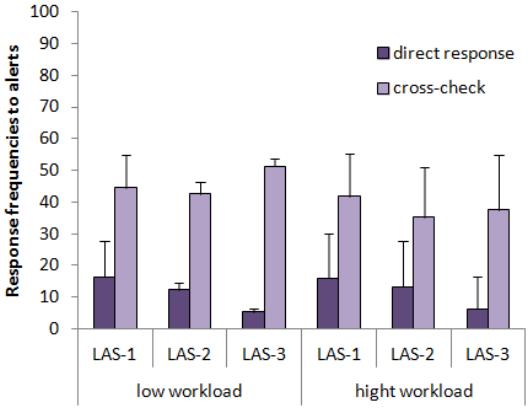

Figure 2. Means and standard deviations of participants' response rates depending on the type of alarm system and the level of workload

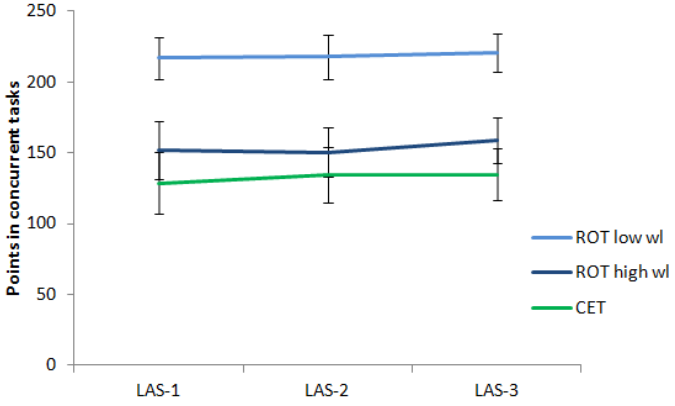

Figure 3. Means and standard deviations of numbers of point for the two concurrent task with high and low workload

METHOD EXPERIMENT 2

Participants

Sixty-one master students (28 females, mean age: 26.08) participated in the study and were randomly assigned to one of four conditions.

Task Environment

The same task environment as described above was used but pictures of container content differed from those used in Experiment 1 and were varied between conditions. That was done in order to manipulate participants' effort and cost to obtain AVI. In the high-cost condition, participants had to wait five seconds after clicking the 'check' button before AVI appeared. AVI consisted of a picture displaying 40 randomly chosen letters from A to Z. The presence of the letter K indicated that the molecular weight was too high. In the low-cost condition AVI was displayed immediately after clicking the 'check' button. In addition, in case of low quality containers the letter K had to be detected among 40 L's used as distractors. It made detection easy due to the pop-out effect.

Experimental Design

A 2(type of alarm system) x 2(checking cost) design was used. Type of alarm system (BAS vs. LAS) and checking cost (low vs. high) were manipulated between groups. As in the

first experiment, alarm systems had a d´ of 1.8 and the base rate of critical events was 0.3. The BAS emitted 66 alarms and the LAS 24 alarms and 42 warnings. PPV of BAS was .35 and LAS had an alarm-PPV of .78 and a warning-PPV of .21. Measures and procedure were the same as in the high workload condition of Experiment 1. Participants performed only one block of 100 trials.

RESULTS EXPERIMENT 2

Behavior

Direct response frequencies and check frequencies were analyzed separately with two-way ANOVAs. With regard to direct response frequencies, a main effect of type of alarm system was found, $F(1, 57)=5.06$, $p<.05$, as well as a main effect of checking cost, $F(1, 57)=10.39$, $p<.0001$. Participants using the LAS had higher direct response frequencies as those working with BAS. When checking was more costly, participants responded directly more often in both alarm system conditions. A similar pattern was found for check frequencies with a main effect of type of alarm system, $F(1, 57)=9.16$, $p<.0001$, and of checking cost, $F(1, 57)=31.82$, $p<.0001$. BAS users checked more often than participants working with LAS and both groups' checking frequencies were lower when checking was more costly. None of the interaction effects reached significance.

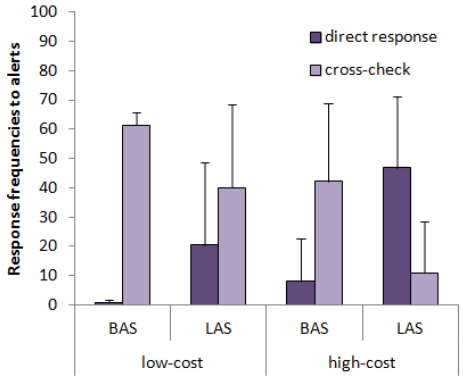

Figure 4. Means and standard deviations of participants' response rates depending on the type of alarm system and the cost of cross-checking

Performance

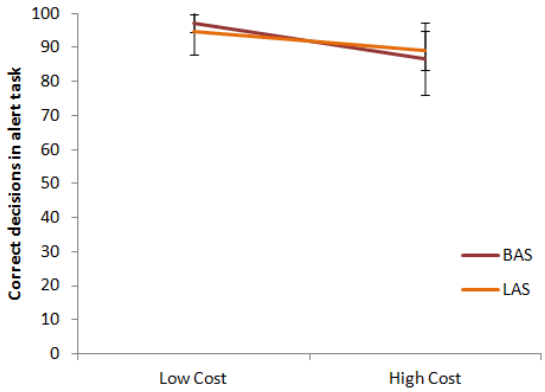

Figure 5. Means and standard deviations of number of correct decisions depending on the type of alarm system and the checking cost

The number of correct decisions and the number of points in ROT and CET were analyzed with two-way ANOVAs. One participant was removed from the analysis of CET because of its outlying performance values. Results of correct decisions in AT did not reveal significant differences with regard to the type of alarm system used. A significant difference was found for variation of checking cost, $F(1, 57)=55.41$, $p<.0001$. Participants made more correct decisions in the low-cost condition. Regarding the number of points obtained in ROT none of the effects reached significance. For the number of points in CET, a significant difference was found for variation of checking cost, $F(1, 56)=9.57$, $p<.0001$, but not for type of alarm system or the interaction. Participants performed better in the CET, when checking was more costly.

DISCUSSION

The aim of the two studies was to gain further insight in performance consequences of LAS in settings where AVI is available. One advantage of LAS over BAS, when providing participants with AVI that was postulated by Sorkin et al. (1988), refers to a performance increase in concurrent tasks when operating the LAS. Sorkin et al. (1988) argue that graduation of alerts based on their PPV offers users the possibility to differentiate their behavior towards alarms and warnings. In their opinion, it would be rational to cross-check AVI only in case of warnings because they are more likely to be false. Alarms instead, should not be checked but participants should comply directly with them, as they are more likely to be true. Overall this would lead to a higher rate of direct compliance with LAS alerts than with BAS alerts. As checking is more time-consuming than complying directly, LAS users can save time and invest it in other activities. This should increase their concurrent task performance. Evidence for this general effect has already been provided in an earlier experiment (Wiczorek & Manzey, 2014). Emphasizing on this result, the first experiment investigated whether varying the proportions of alarms and warnings of the LAS would affect participants' performance in one or two concurrent tasks. It was suggested that a decreased number of warnings would reduce cross-check frequencies. This should save time and thus, lead to higher concurrent task performance.

However, results of the current experiment do not support this assumption. Neither a significant decrease in cross-check frequencies nor a significant higher concurrent task performance for LAS with a lower proportion of warnings could be found. When workload was higher due to the added third task, direct compliance with alarms increased and performance in both, the AT and ROT, decreased across all the three LAS.

Experiment 2 focused on alert task performance rather than concurrent task performance. It has been shown that benefits of LAS over BAS in terms of correct decisions with alerts disappear when participants are provided with AVI (Wiczorek & Manzey, 2014). Users of both systems increase their performance equally when using AVI. This is because participants working with BAS cross-check almost every alert and users of LAS cross-check almost every warning. Therefore, both groups reduce wrong decision making to a minimum. Another study, however, revealed some evidence that specific costs associated with AVI may affect users' behavior (Manzey et al., 2014). Therefore, it was assumed that an increase in effort needed to obtain AVI would result in a

decrease in participants' overall check frequencies. Less checking on its part should result in a reduction of correct decisions, i.e. AT performance. This effect was suggested for both types of systems. However, reducing check frequencies was expected to harm AT performance stronger when working with the BAS than with the LAS. The reason is that LAS users should reduce check frequencies in the warning stage that has a low warning-PPV. Missing a true critical event is not very likely for them. BAS users have a greater risk to commit misses because the alarms they decide not to check have a higher possibility to be true alarms. In the current experiment, the effort associated with obtaining AVI was increased in two ways simultaneously, time-based, i.e. longer time to wait for AVI and cognitive, i.e. more resources needed to interpret information.

Results show the expected decrease in check frequencies as a consequence of high-cost AVI for both, the LAS and the BAS. Resulting AT performance, however, was not in line with hypotheses. Both groups showed equally reduced AT performance; no advantage of LAS over BAS emerged. An unexpected result was found for CET performance. While performance in ROT did not differ between conditions, CET performance for both systems was higher with the high-cost AVI. It seems that participants reduced check frequencies not only in order to maintain actual ROT and CET performance but rather shifted resources from AT to CET when obtaining AVI was related with high costs.

Results of both experiments do not completely correspond with assumptions made before. To understand the reasons one should have a closer look to participants' behavior with the LAS. The underlying basic hypothesis of Sorkin et al. (1988) assumes that LAS users are cross-checking most of the warnings and complying with most of the alarms. A behavior, that has been referred to as 'extreme responding' (Bliss, 2003). In the context of BAS, it has been shown that extreme responding only occurs when PPV is 0.7 or higher. A PPV lower than 0.5, participants rather engage in a behavior called 'probability matching' (Bliss et al., 1995). They try to imitate alarm validity with their response frequency (Manzey et al., 2014). While this behavior is less efficient than extreme responding, it is more widespread because most BAS have low PPVs (Parasuraman et al., 1997). Most LAS used in the experiments here had alarm-PPVs higher than 0.7 (the only exception was LAS-1 in the first experiment with an alarm-PPV of 0.63, which is still above 0.5). Therefore it seemed reasonable to expect extreme responding behavior. A closer look to the descriptive data suggests that at least in some conditions LAS users might have shown a probability matching behavior because the number of containers they directly responded to was below the number of alarms. That is the case with LAS-1 in the first experiment and the LAS in the second experiment, whose alarm-PPV was 0.78.

Therefore, a possible interpretation of results might be that also when using LAS, participants tend to show probability matching with alarms when alarm-PPVs go below a certain PPV. It is also possible that this 'critical alarm-PPV' is higher compared to the critical value of PPV in BAS as direct response frequencies for in the second experiment suggest, where alarm-PPV was above 0.7. However, this interpretation can only be made with caution because analyses of means can bias individual response patterns.

To confirm these assumptions further research is needed investigating extreme responses and probability matching on an individual level as suggested by Bliss (2003).

LITERATURE

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

Bliss, J. P. (2003). 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* (pp.1683-1687). Santa Monica, CA: Human Factors and Ergonomics Society.

Braun, C. C., & Silver, N. C. (1995). Interaction of signal word and colour on warning labels: differences in perceived hazard and behavioural compliance. *Ergonomics*, *38*, 2207–2220.

Breznitz, S. (1984). *Cry wolf: the psychology of false alarms*. Hillsdale N.J.: Lawrence Erlbaum Associates.

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.

Chapanis, A. (1994). Hazards associated with three signal, words and four colours on warning signs. *Ergonomics*, *37*, 265–275.

Edworthy, J. (2013). Alarms are still a problem! *Anaesthesia*, *68*(8), 791-794.

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*(1), 19–33.

Lawless, S. T. (1994). Crying wolf: false alarms in a pediatric intensive care unit. *Critical care medicine*, *22*(6), 981-985.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 241–256.

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*, 1-23.

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

Ragsdale, A., Lew, R., Dyre, B. P., & Boring, R. L. (2012). Fault Diagnosis with Multi-State Alarms in a Nuclear Power Control Simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 2167-2171). Boston, MA: Human Factors and Ergonomics Society.

Sorkin, R.D., Kantowitz, B.H. & Kantowitz, S.C. (1988). Likelihood alarm displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society,30,* 445-459.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1*(1), 49–75.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*(4), 522-532.

Thomas, L. C., Wickens, C. D., & Rantanen, E. M. (2003). Imperfect automation in aviation traffic alerts: A review of conflict detection algorithms and their implications for human factors research. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 344-348). Santa Monica, CA: Human Factors and Ergonomics Society.

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Wiczorek, R., Manzey, D. & Zirk, A. (2014). Benefits of Decision-Support by Likelihood versus Binary Alarm Systems: Does the number of stages make a difference? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*, 380-384. Santa Monica: Human Factors and Ergonomics Society.