Anna Zirk, Rebecca Wiczorek, Dietrich Manzey

Do We Really Need More Stages? Comparing the Effects of Likelihood Alarm Systems and Binary Alarm Systems

Zirk, A., Wiczorek, R., & Manzey, D. (2019). Do We Really Need More Stages? Comparing the Effects of Likelihood Alarm Systems and Binary Alarm Systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 1872081985202. https://doi.org/10.1177/0018720819852023

Human Factors, online first, © 2019 Human Factors and Ergonomics Society, DOI: 10.1177/0018720819852023

Title: Do we really need more stages? Comparing the effects of likelihood alarm systems and binary alarm systems

Anna Zirk
Research Assistant
Junior Research Group FANS
Department of Psychology and Ergonomics
Technische Universität Berlin

Rebecca Wiczorek
Group leader of the Junior Research Group FANS
Department of Psychology and Ergonomics
Technische Universität Berlin

Dietrich Manzey
Chair of Work, Engineering & Organizational Psychology
Department of Psychology and Ergonomics
Technische Universität Berlin

Running head: Likelihood and Binary Alarm Systems

Address correspondence to Anna Zirk, Institute of Psychology and Ergonomics, Technische Universität Berlin, Sec. MAR 3-2, Marchstrasse 23, 10587 Berlin; e-mail:

27 zirk@tu-berlin.de

Abstract

Objective: This research investigates the potential behavioral and performance benefits of a 4-stage likelihood alarm system (4-LAS) contrasting a 3-LAS, a binary alarm system with a liberal threshold (lib-BAS) and a BAS with a conservative threshold (con-BAS).

Background: Prior research has shown performance benefits of 3-LASs over conventional lib-BASs due to more distinct response strategies and better discriminating true from false alerts. This effect might be further enhanced using 4-LASs. However, the increase of stages could cause users to reduce cognitive complexity by responding in the same way to the two lower and the two higher stages, thus treating the 4-LAS like a con-BAS.

Method: All systems were compared using a dual task paradigm. Response strategies, number of joint human machine (JHM) false alarms (FAs), misses, and sensitivity were regarded.

Results: Compared to the lib-BAS, JHM sensitivity only improved with the 4-LAS and the con-BAS. However, the number of JHM misses was lowest for the con-BAS compared to all other systems.

Conclusion: JHM sensitivity improvements can be achieved by using a 4-LAS, as well as a con-BAS. However, only the latter one may also reduce the number of JHM misses, which is remarkable considering that BASs with conservative thresholds a priori commit more inbuilt misses than other systems.

Application: Results suggest implementing conservative BASs in multi-task working environments to improve JHM sensitivity and reduce the number of JHM misses. When refraining from designing systems which are miss prone, 4-LASs represent a suitable compromise.

Key Words: warning, threshold setting, decision-making, signal detection theory, automation

Précis: Using a multi-task paradigm, we compared the behavioral and performance

consequences of a four-stage likelihood alarm system (4-LAS) to a 3-LAS and two binary alarm systems (BASs), one with a liberal and one with a conservative threshold and found the conservative BAS to lead to the best performance and behavior.

In many safety critical domains, such as aviation or process industry, operators must carry out supervisory tasks, including monitoring the time dynamics of processes or monitoring of parameters (e.g., temperature or pressure), in order to evaluate the current state of the system as nominal or critical. Frequently, these tasks must be performed concurrently with other tasks (e.g., manual operations, communication with others). However, in case of a critical event, requiring intervention under time pressure by the operator, the supervisory task must immediately be prioritized over the other tasks. Automated monitoring systems with integrated alarm or warning functionalities often assist the operator in their priority setting by guiding their attention to critical events and supporting their decision-making. Alarms emitted by these systems usually provide the most salient and only cue for an operator to decide upon a proper action. This applies to all sorts of remote monitoring devices implemented, for example, in intensive care units of hospitals, in centralized control rooms, or in aircraft cockpits. In these settings, monitoring devices usually provide alarms indicating critical states without the operator being able to cross-check the alarm validity towards other directly available information.

Currently, most of such alarm systems are binary alarm systems (BASs) that remain silent (e.g., show a green light) as long as all data assessed suggest a nominal operation and emit an alarm (e.g., show a red light) as soon as deviations from nominal operation are detected. Due to imperfect reliability caused by inherent technical constraints and ambiguous (noisy) data, the alarm systems can err. These errors can either be false alarms (FAs), defined as alarms generated without an underlying critical event, or misses, i.e., no alarm is generated in the presence of a critical event. These errors are not independent of each other but inevitably linked through the choice of threshold setting for the emittance of alarms. Specifically, if low (liberal) threshold settings are used and alarms are already emitted in response to weak deviations of the nominal state, the number of misses is kept low but only at the expense of a considerable number of FAs. The opposite holds true for choosing higher

(more conservative) threshold settings.

In most safety critical domains, designers of BASs prefer to use liberal thresholds, i.e., they prefer false-alarm prone systems over miss prone systems. This reflects the commonly applied fail-safe engineering approach (Swets, 1992; Parasuraman & Riley, 1997). However, experiencing many FAs can reduce operators' trust in the alarm system (Lee & See, 2004; Madhavan, Wiegmann, & Lacson, 2006). As a consequence, their response time to alarms can increase (e.g., Getty, Swets, Pickett, & Gonthier, 1995; Wickens & Colombe, 2007), or they may even completely ignore given alarms (e.g., Bliss, Gilson, & Deaton, 1995; Lees & Lee, 2007; Meyer, Bitan, Shinar, & Zmora, 1999). This effect has been referred to as 'cry wolf' phenomenon (Breznitz, 1984) which is related to the problem of alarm fatigue (Graham & Cvach, 2010; Sendelbach; 2013) and can compromise safety by specifically enhancing the risk of missing a critical event. Goel, Datta and Mannan (2017) have provided a recent review of incidents caused by such inappropriate alarm responses.

The current research investigates to what extent an improvement of adequate responding to alarms can be achieved by providing operators with more complex Likelihood Alarm Systems (LAS; Sorkin, Kantowitz, & Kantowitz, 1988). LASs do not only inform users about the absence and presence of a critical event, but also provide a sort of staged information about the relative likelihood that the emitted alert is actually true.

BACKGROUND: RESPONDING TO ALARMS

Alarm systems should support operators in detecting critical events. This implies that operators are expected to adjust their behavior according to the alarm systems' outputs. Specifically, they are expected to continue with their tasks and refrain from any action if the alarm system remains silent but need to initiate immediate proper action when an alarm is emitted. According to Meyer (2001), the former behavior is referred to as reliance and the latter as compliance. However, operators do not always behave as intended. For

example, the cry wolf effect mentioned above reflects a clear lack of compliance, based on the repeated experience of FAs. Since in most of the cases operators do not know the exact threshold setting of their alarm systems, their decision whether or not to respond to alarms is usually based on the perceived alarm reliability, which has been referred to as the positive predictive value of an alarm system (PPV; Getty et al., 1995). Formally defined, the PPV is the conditional probability of a critical event, given an alarm is emitted. It is calculated by dividing the number of hits by the total number of alarms (i.e., hits plus FAs). The corresponding characteristic of the non-alert stage is the negative predictive value (NPV), defined as the number of correct rejections (CRs) divided by the number of non-alert events (i.e., CRs plus misses; Meyer & Bitan, 2002).

Consistent findings over the past twenty years have shown that response frequencies to alarms decrease with decreasing PPV (e.g., Bliss et al., 1995; Bustamante, Bliss, & Anderson, 2007; McCarley, Rubinstein, Steelman, & Swanson, 2011; Manzey, Gèrard, & Wiczorek, 2014). More specifically, response behavior in interaction with alarms often mirrors one of two different strategies: probability matching or extreme responding (Bliss, 2003). Probability matching represents a sort of response heuristic in which operators try to adjust their response rates to the PPV, with lower PPVs leading to successively lower and higher PPVs leading to successively higher response rates. In contrast, extreme responding mirrors an all-or-nothing strategy, leading to either ignoring most alarms (negative extreme responding) or to responding to most alarms (positive extreme responding). While probability matching has been found to be the dominant strategy for medium PPVs, negative and positive extreme responding often are applied in response to alarms with low and high PPVs, respectively (e.g., Bliss, 2003). For example, in the study of Manzey et al. (2014, Exp. 1) the portion of participants who preferred positive extreme responding over probability matching increased from 8% to 90% with the PPV increasing from .5 to .9. In contrast, incidents of negative extreme responding, indicating a cry wolf effect, increased

considerably for PPVs lower than .4. In most domains where alarms systems are implemented the base rate of critical events is usually low. Consequently, even highly sensitive BASs become false-alarm prone to a considerably high degree – with PPVs of BASs frequently less than .3 (Parasuraman & Riley, 1997; Parasuraman, Hancock, & Olofinboba, 1997). In this case, both strategies mentioned above would directly lead to a high rate of ignored alarms during the interaction with BASs. One possible countermeasure to prevent or at least mitigate such an effect is the use of LASs. By providing various alerts with different PPVs, LASs provide more options than BASs to guide users' behavior. Thus, they enable operators to better distinguish between true and false alerts than BASs and to adapt their behavior accordingly. However, the full potential of LASs has not yet been investigated in its entirety.

LIKELIHOOD ALARM SYSTEMS

The basic concept of LASs has already been suggested thirty years ago as an alternative to BASs by Sorkin et al. (1988). In contrast to BASs, LASs have more than one threshold for emitting various alerts, which then differ in their PPV and therefore inform the operator about the relative likelihood of an underlying critical event. Compared to control conditions with classical BASs, LASs were found to improve decision-making and performance in terms of accuracy (e.g., Clark, Peyton, & Bustamante, 2009; Ragsdale, Dyre, & Boring, 2012; Wiczorek & Manzey, 2014), particularly under high-workload conditions and for low base rates (Bustamante 2005, 2008; Clark & Bustamante, 2008). Moreover, it has been shown that LASs are especially useful to improve proper responding to alerts in case that the validity of an alert cannot be easily verified towards other available information (Wiczorek & Manzey, 2014). Only a few studies did not find benefits of LASs over BASs (e.g., Wickens & Colombe, 2007).

The common procedure of designing a LAS is keeping the initial low threshold of a typical

liberal BAS, which separates non-alerts from alerts, but grading the alert level further by adding (at least) one additional threshold (Bustamante 2005, 2008; Clark & Bustamante 2008; Clark et al., 2009; Clark, Ingebritsen, & Bustamante, 2010; Ragsdale et al., 2012; Vargas & Bustamante 2011; Wiczorek, 2017; Wiczorek & Manzey, 2014, Wiczorek, Manzey, & Zirk, 2014).

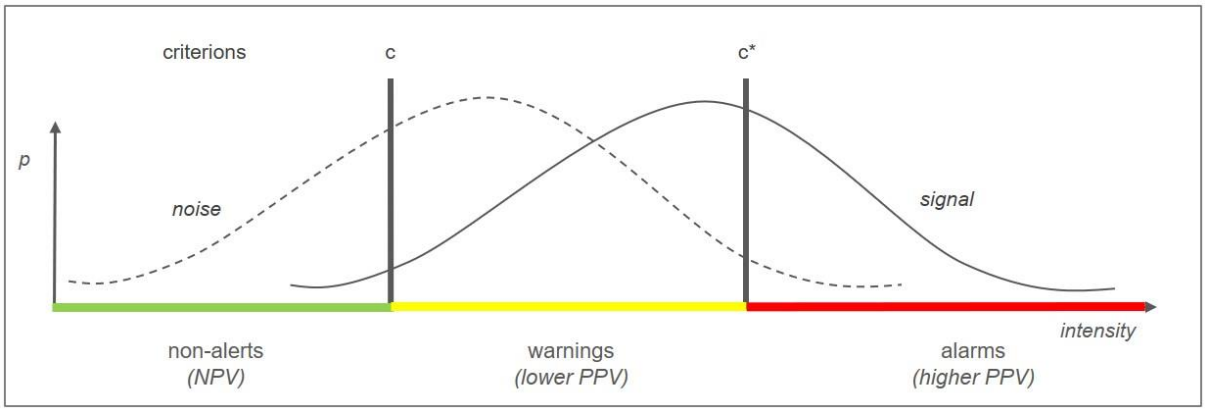

Figure 1. Schematic representation of a three-stage LAS.

Most common are LASs with two thresholds which consist of three stages (3-LAS) as depicted in Figure ***1***. Parameters above the first (original) and below the second threshold trigger a warning (i.e., relatively low PPV) and parameters exceeding the second threshold trigger an alarm (i.e., relatively high PPV). Such systems have been found to improve operators' decision-making by increasing responses to true alerts and, at the same time, reducing responses to FAs. Specifically, participants interacting with a 3-LAS were found to apply probability matching to warnings, but to choose positive extreme responding in response to alarms when they do not have the chance to validate the alarm system's diagnoses (Wiczorek, 2017; Wiczorek & Manzey, 2014). Thus, for 3-LASs, the cry wolf effect is still visible to some extent but almost exclusively in interaction with warnings, which have a lower likelihood to truly indicate a critical event anyway.

Based on these findings, the question arises whether this benefit of 3-LASs could be further enhanced by an even more graduated 4-LAS. Adding a fourth stage by separating the former

3-LAS warning stage into *higher*-PPV and *lower*-PPV warnings could shift the cry wolf effect further to the lower-PPV warnings of the 4-LAS. Due to their lower likelihood to truly indicate a critical event, ignoring them is less likely to result in missing a critical event (compared to the 3-LAS warnings). This in turn could cause performance improvements from the 3-LAS to the 4-LAS. However, an obvious trade-off that must be considered is one between the benefits of more distinct information and the disadvantages of a higher complexity for operators to adjust response behavior to the different sort of alerts. Thus, it remains to be seen whether a four-stage LAS (4-LAS) really enhances its value beyond that of a 3-LAS or leads operators to reduce the raised complexity, for example, by responding in the same way to the two lower and higher stages, respectively. In the latter case, the more distinct information provided by the 4-LAS would not be used and the whole system would be treated like a 3-LAS or even a BAS with a relatively conservative threshold.

Thus far, only few studies have investigated the performance consequences of LASs with more than three stages. For example, already in their classical work, Sorkin et al. (1988) contrasted a 4-LAS with a conventional BAS. However, they created the fourth stage by further separating the non-alert stage into *lower*-NPV and *higher*-NPV non-alerts. Consequently, the alarm stage and the warning stage corresponded to those known from most 3-LASs.

St. John and Manes (2002) went even further and investigated the performance consequences of a six-stage alerting system that supported participants in a visual search task. Participants had the option to validate the system's diagnoses by rolling over a location with the mouse and to hold for one second to get a clearer view on the target. They found that the six-stage alerting system led to a better performance than a BAS. However, given that the authors did not compare the six-stage alerting system with a simpler 3-LAS, it remains unclear whether the benefits were linked to the six stages or resulted from a more general effect of graduated alerts at all. In a study directly contrasting different types of

LASs, Shurtleff (1991) compared three different LASs consisting of four, six, and eight stages to a BAS. Participants were provided with different polygons which they had to identify as friends or foes while being supported by one of the alarm systems. Shurtleff (1991) found significant performance improvements for the 4-LAS and the 8-LAS compared to the BAS in a target detection task, but not between any of the LASs. Moreover, again the three complex LASs were not contrasted with a basic 3-LAS and therefore no clear conclusion can be drawn whether the performance benefits of the LASs were due to the number of stages > 3 or just the graduation of alert levels in general.

CURRENT RESEARCH

The current research compares the behavioral effects and performance consequences of a four-stage LAS (4-LAS) with a three-stage LAS (3-LAS) and two sorts of BAS, the latter differing in whether they had a conventional liberal threshold (lib-BAS) or a more conservative (con-BAS) threshold for emitting alarms. All alarm systems were modeled based on the signal detection theory (SDT; Green & Swets, 1966). They consisted of the same good but not perfect sensitivity $d' = 1.7$. The base rate of critical events was set to $p = .3$ in every condition. These parameters were chosen to allow the comparison with prior studies including LASs using similar sensitivities and base rates of $d' = 1.8$ and $p = .3$, respectively (Wiczorek & Manzey, 2014; Wiczorek, Balaud & Manzey, 2015, Wiczorek 2017). Choosing such a relatively high base rate reflects a compromise between simulating a realistic situation which often is characterized by much lower base rates of critical events and the necessity to elicit enough events for a reliable behavior assessment in a time limited

testing session. Threshold settings and resulting NPVs and PPVs are displayed in Figure *2*.

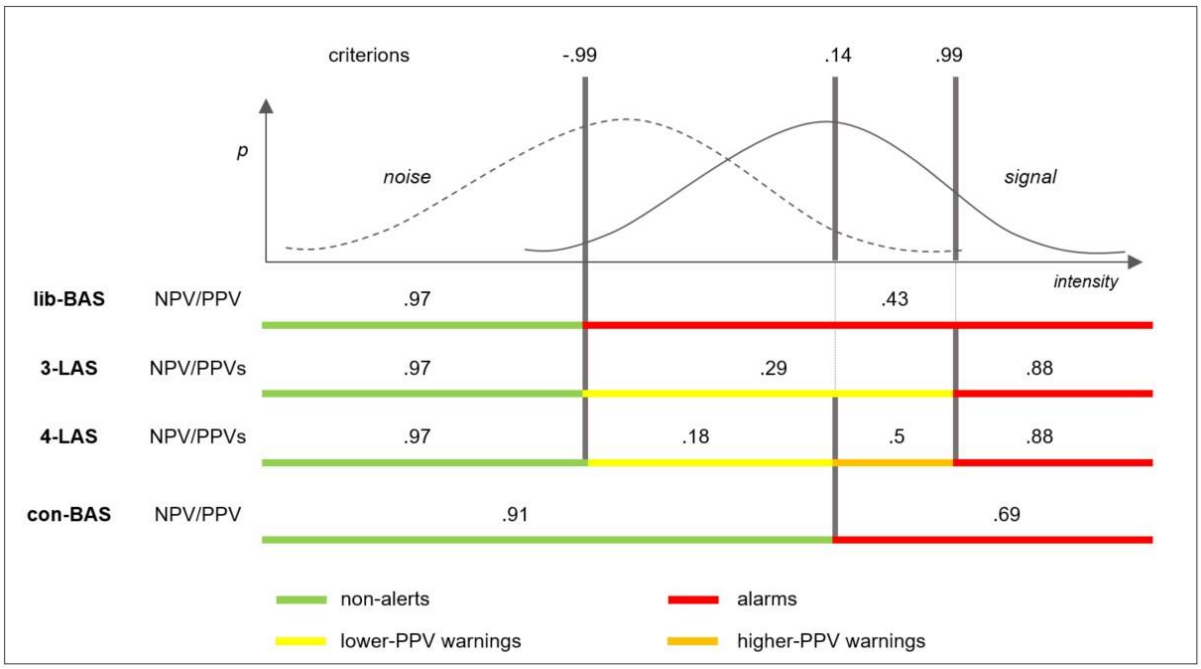

Figure 2. Thresholds and resulting NPVs and PPVs for the four alarm systems used in this study.

The two LASs and the lib-BAS shared the same (first) threshold, separating the non-alert (green) from the alert stage (red). Thus, the overall alert-PPV for these three systems was .43. The 3-LAS had a second threshold, separating the alert stage in an alarm stage (red) with an alarm-PPV of .88 and a warning stage (yellow) with a warning-PPV of .29. For the 4-LAS, this warning stage was further separated by a third threshold, resulting in a PPV of .5 for the higher-PPV warning stage (amber) and .18 for the lower-PPV warning stage (yellow).

In the case of the lib-BAS, probability matching was expected to be the dominant strategy in the alarm stage, resulting in a considerable number of ignored alarms and perhaps missed critical events. For both LASs, however, positive extreme responding was expected to be the dominant response pattern to alarms due to their high PPV. This should lead to more correct responses to true critical events (“hits”) compared to the lib-BAS. The other alert stages of the two LASs were expected to guide behavior in a distinct way, related to the different PPVs with an even better informational basis of proper differentiation between

true and false alerts provided by the 4-LAS compared to the 3-LAS. However, instead of applying different strategies to the different alert levels of the 4-LAS, participants could also reduce the complexity by mentally transforming the 4-LAS into a three-stage or a two-stage system by ignoring one or two of the thresholds. The latter option would then correspond to a sort of mental dichotomization in which the two lower and higher stages would be integrated into one stage, respectively. The resulting mental representation would correspond to a BAS with a more conservative threshold. To investigate this possibility, the con-BAS was included as a fourth alarm system in the current research. The con-BAS's only threshold corresponded to the middle threshold of the 4-LAS (separating lower- and higher-PPV warnings) resulting in an alarm-PPV of .69 for the con-BAS.

METHOD

Participants

Based on a power analysis and the assumption of a large effect ($\eta^2 = .14$), 60 (28 male, 32 female) students were recruited to participate in the study. Their age ranged from 20 to 47 years ($M = 26.27$; $SD = 4.43$). They were randomly assigned to one of the four conditions. This research complied with the tenets of the Declaration of Helsinki. Informed consent was obtained from each participant. For their participation, they received a basic compensation of either €10 or ECTS credits, complemented by an additional reward of up to €8, depending on their performance.

Task environment

The PC-based multi-task operator performance simulation (M-TOPS, Manzey et al., 2014) was used as simulation environment. It represents a dual-task environment requiring concurrent performance of a quality control task and a cognitive task simulating basic operational demands of control room operators. Participants were instructed to keep the

process of the plant running. For this, two tasks had to be executed: the resource ordering task (Figure 3, upper left side) and the alert task (Figure 3, bottom right side).

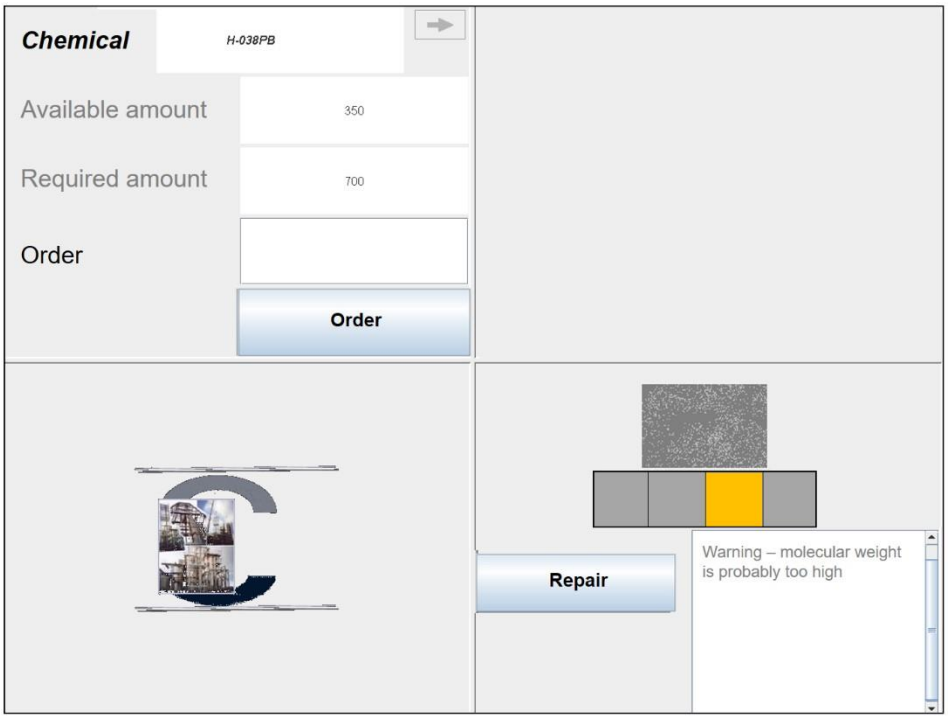

Figure 3. M-TOPS interface with resource ordering task in the upper left side and alarm task with a 4-LAS, indicating a high-likelihood warning, on the bottom right side.

Resource ordering task: Participants are instructed to order chemicals that are needed to maintain the chemical process. In the upper left of the screen, participants see the actual amount and the demand of one chemical at a time. Their task is to calculate the difference (i.e., the required amount), to enter it into the referring field, and to send the order by clicking the 'order' button. After clicking the button, a new task appears. Every ordering task is displayed for a maximum duration of 15 seconds. Participants' responses are logged automatically.

Alert task: Participants are told they are responsible for controlling the quality (i.e., the molecular weight) of the chemical end product. In this task, participants are supported by one of the four alarm systems. They are told that the plant has a control station that checks the containers filled with the chemical product automatically. Every six seconds a new container enters the control station. For each container, a diagnosis is given by the automatic

control system. When the chemical product meets the quality standards (appropriate molecular weight), the alarm system shows a green light. When the quality of the chemical product is not adequate, the alarm system sends off an alert. Each diagnosis is accompanied by a notification as depicted in Figure 4 below. Participants do not receive any alarm validity information.

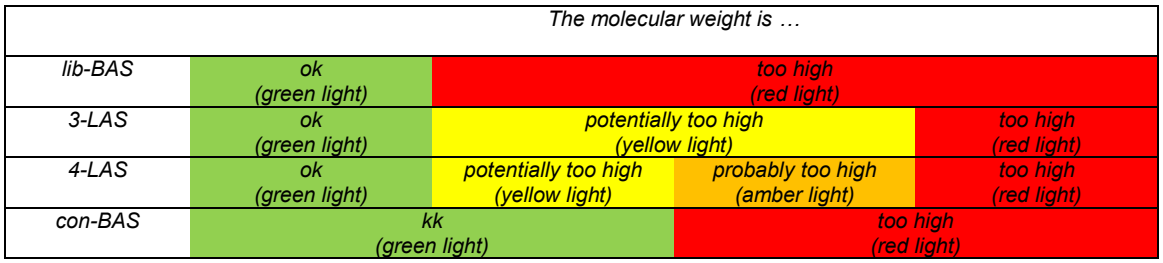

Figure 4. Notifications and colors of the different stages of the four alarm systems.

Containers obtaining a chemical product not meeting the quality standards can be *repaired* by the participant when clicking the 'repair' button within six seconds. Containers that meet the criteria leave the control station automatically after six seconds and no action of the participant is required. Participants' responses are logged automatically.

Payoff

Participants received 1.5 points for every correct order in the resource ordering task. For every wrong decision in the alert task, they lost 2 points. This procedure was chosen to create a competitive situation between both tasks and to ensure that they were considered as equally important by the participants. For each point participants received 2.5 Euro cents.

Dependent measures

Behavior

Response strategies were analyzed for each person and system stage individually based on previous research (Manzey et al., 2014). Response rates of 90% and above were classified as positive extreme responding, response rates of 10% and lower were classified as negative extreme responding. All individual response rates in between were regarded as probability

matching.

Performance

The following measures served as performance indicators of the alert task, reflecting the overall performance of the joint human machine (JHM) system:

(1) number of FAs committed by a participant when supported by a given system, (defined as the number of clicking the repair button when the container was ok),

(2) number of misses committed by a participant when supported by a given system, defined as the number of missing responses when an action was needed (i.e., when the molecular weight was too high),

(3) overall sensitivity of the JHM system corresponding to the d' parameter of the SDT, defined as d'= z[p(JHM hit)] – z[p(JHM FA)], with p(JHM hit) = JHM hits / (JHM hits + JHM misses) and p(JHM FA) = JHM FAs / (JHM FAs + JHM CRs).

In order to assess the performance in the resource ordering task, the total number of correct responses was recorded.

Procedure

The experiment took place at Technische Universität Berlin in groups of up to four people. After signing consent forms and filling in demographic questionnaires, participants navigated through the instruction presentation. They were told they would be operating an industrial plant and were responsible for two tasks – alert task and resource ordering task – which are both equally important and that a reliable but not error-free alarm system would support them executing the alert task. Subsequently, they practiced both tasks as single tasks and in parallel, two minutes each. The alarm system was running during practice sessions (except when practicing the resource ordering task as single task).

After this instruction and practice part, participants conducted a 100-trial alert task block to become familiar with the characteristics of the referring alarm system. During this block

feedback was provided after each trial by an acoustical signal informing participants about committing a wrong decision (i.e., clicking the 'repair' button when the container was intact or not clicking the button when the container was faulty). After this block, participants were informed about the actual system characteristics (NPV and PPV(s)) by showing them the absolute number of correct and wrong diagnoses made by the referring alarm system in order to avoid any biases related to only experience-based vs. description-based information (Hertwig & Erev, 2009). The following experimental block then included a total of 100 trials of the alert task which had to be performed concurrently with the resource ordering task. No feedback was provided during this block. The whole experimental session lasted two hours. At the end of the session the participants were paid and debriefed.

RESULTS

Individual response strategies were only regarded descriptively. The different performance measures (d', number of FAs and misses) of the alert task were analyzed using the non-parametric Kruskal-Wallis test (Kruskal & Wallis, 1952). This test was chosen due to a violated variance homogeneity. Additional pairwise post-hoc contrasts of performances in the different conditions were performed by non-parametric Dunn's test (Dunn, 1961). Since the shape of the distributions of the four groups differed, the Kruskal-Wallis test contrasted the mean ranks of the four groups, which are reported along with the statistical results in the text. Note that small ranks correspond to small variable values. However, for allowing a comprehensive descriptive comparison, medians are depicted in the figures. Performance in the resource ordering task was analyzed using a one-way ANOVA. All analyses were performed with the IBM SPSS Statistics 25 package. Because of missing data, only 59 of the 60 participants were included in the statistical analysis.

Response strategies in interaction with alerts

The response strategies to the different stages differed considerably. Most of the participants

working with the lib-BAS applied probability matching in response to alarms while most of the participants working with one of the two LASs responded to almost all the emitted alarms, i.e., they applied positive extreme responding. The users' main strategy of responding to warnings emitted by the 3-LAS was probability matching. Participants of the 4-LAS showed a distinct pattern of strategies when responding to the two types of warnings, which was more extreme than expected. With the higher-PPV warnings at least half of the participants applied the positive extreme responding heuristic while the dominant strategy for the lower-PPV warnings was negative extreme responding. Finally, the con-BAS system only triggered extreme response strategies, with negative extreme responding to non-alerts and positive extreme responding to alarms.

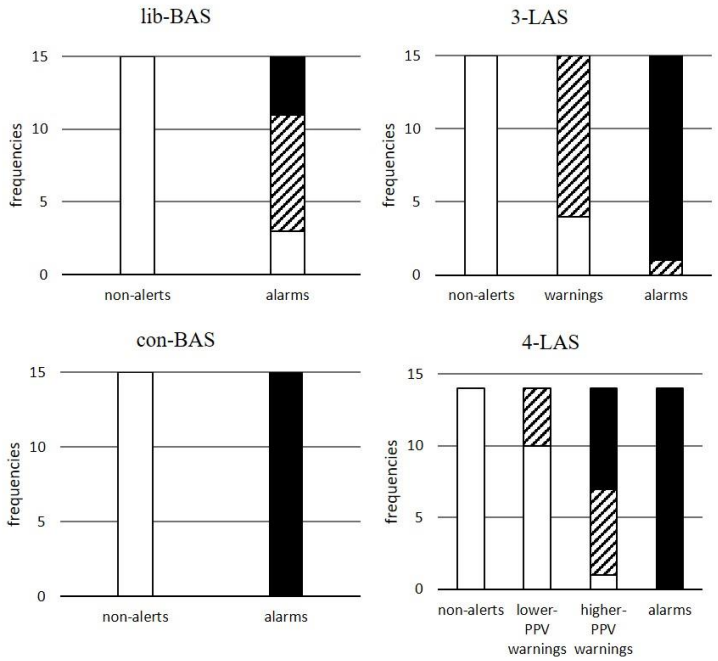

Figure 5. Response strategies applied for all diagnoses of the four alarm systems.

Alert task performance

The distributions of the number of misses and FAs for all four alarm systems are depicted in Figure 6. While the differences between the number of FAs committed by the participants when working with the different systems just failed the conventional level of statistical significance, $\chi^2(3, N = 59) = 7.468$; $p = .058$, $\eta^2 = .13$, a significant effect for alarm system

was found for misses, $\chi^2(3, N = 59) = 19.587$; $p < .001$, $\eta^2 = .34$. The number of misses was lowest when the participants were supported by the con-BAS (mean rank across individuals: 14), followed by the 4-LAS (31.6) and both, the lib-BAS (34.7) and the 3-LAS (39.9). Bonferroni corrected pairwise comparisons based on the Dunn's test revealed the differences between the con-BAS and all other systems as significant (all $p < .04$).

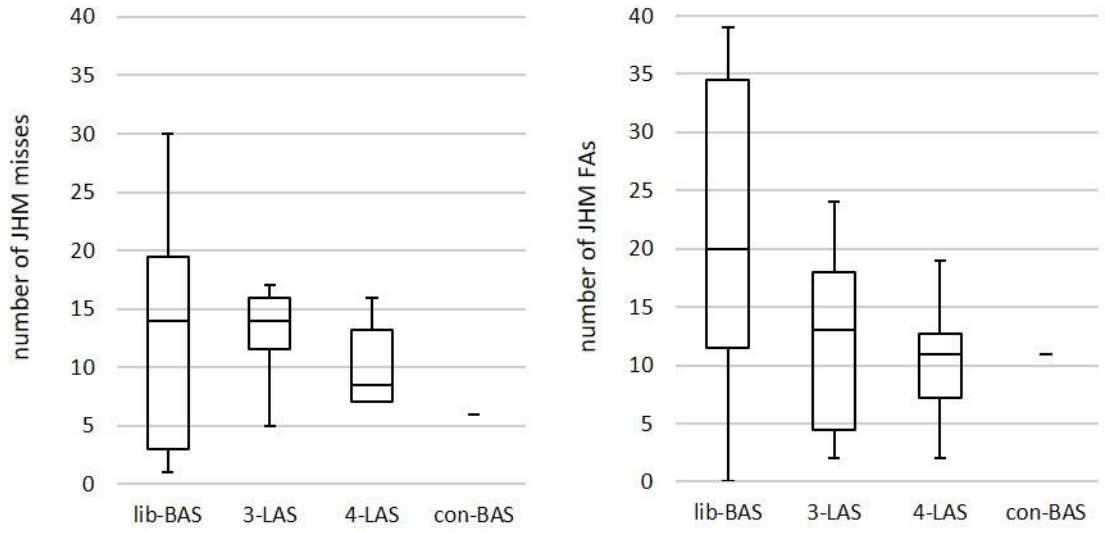

Figure 6. Median, quartiles, minimum and maximum of the number of JHM misses and FAs for the four alarm systems.

Figure 7 shows the d' distributions for the four alarm systems. In line with our expectations, d' was higher when the participants were supported by one of the two LASs (3-LAS: mean rank = 23; 4-LAS: 31.8) compared to the lib-BAS (13.7). However, the highest d' was found for the con-BAS (51.6). Statistically, this was confirmed by a significant main effect for alarm system, $\chi^2(3, N = 59) = 40.376$; $p < .001$, $\eta^2 = .70$.

Bonferroni corrected pairwise comparisons based on Dunn's test revealed significant differences between the con-BAS and all other systems (all $p < .02$), and between the 4-LAS and the lib-BAS, $p = .026$. No significant differences emerged between the lib-BAS

and the 3-LAS, as well as between the two LASs.

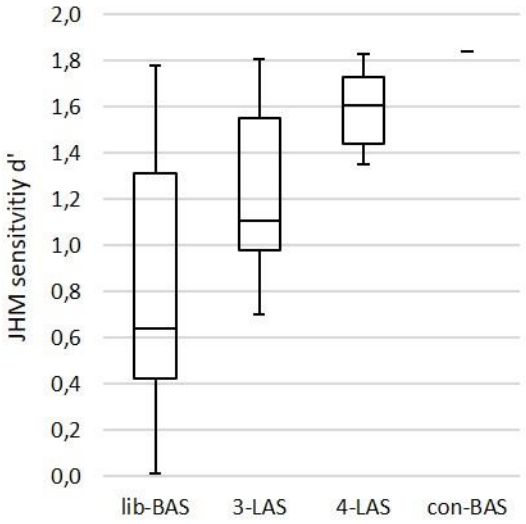

Figure 7. Median, quartiles, minimum and maximum of the JHM d' for the four alarm systems.

Concurrent task performance

No significant differences between the four alarm systems emerged regarding the performance in the resource ordering task, F (3, 55) = .13; p = 945, η^2 = .01.

DISCUSSION

The current study aimed to investigate the possible benefits of LASs compared to BASs in a situation where the emitted alerts represented the only cue to decide whether to intervene in an automated process. For this purpose, we compared the behavioral and performance consequences of two LASs of different complexity (3-LAS; 4-LAS) and a conventional BAS with a relatively liberal threshold setting. In addition, a con-BAS with a conservative threshold was included as control condition to investigate possible strategies of complexity reduction of 4-LAS users.

Surprisingly, the con-BAS yielded the best performance in terms of a significantly improved JHM sensitivity. This superiority was partly related to the (descriptively) lowest number of FAs, but mainly due to the lower number of misses compared to all other systems. The latter finding is particularly interesting because the con-BAS had the highest *a priori* probability by design to commit misses (due to the low NPV resulting from the conservative threshold

406 setting). The analysis of response strategies suggests that participant's high compliance
407 rates to alarms and, thus, the absence of any cry wolf effect, were sufficient to more than
408 compensate for the inbuilt misses of the con-BAS. This result was not expected but might
409 explain previous findings suggesting that humans prefer more conservative thresholds in
410 BASs when they have a choice (Bustamante et al., 2007; Merkel & Wiczorek, 2012).
411 With respect to the behavioral consequences of the 3-LAS, this study confirms the results
412 of previous research by Wiczorek and Manzey (2014). As expected, alarms and warnings
413 induced different response strategies, with positive extreme responding and probability
414 matching being the dominant strategies, respectively. However, in contrast to previous
415 findings (e.g., Bustamante, 2005; Bustamante 2008), the effects of performance
416 improvements over the lib-BAS in terms of reduced FAs, reduced misses, and an increased
417 d' were not strong enough to reach significance. A clearer (significant) advantage of
418 providing graduated alerts compared to the lib-BAS was achieved when participants were
419 supported by the 4-LAS. This is in line with other results of our lab (based on data collected
420 shortly after the one of the present study), which even showed a significantly improved
421 performance of the 4-LAS compared to the 3-LAS with only slightly different threshold
422 settings (Balaud & Manzey, 2014).
423 The analyses of response strategies revealed that the performance advantage of the 4-LAS
424 over the lib-BAS was not due to participants using the more graduated information for a
425 more complex differentiation in responding to the different types of alerts. Actually, ten out
426 of 14 participants in the 4-LAS condition ignored most of the lower-PPV warnings (i.e.,
427 committed a negative extreme responding strategy to this type of alerts) and treated them
428 the same as the non-alerts. For the higher-PPV warnings, half of the participants showed
429 positive responding (i.e., they did not make a difference between the higher-PPV warning
430 stage and the alarm stage). Thus, it seems that providing more graduated information with
431 a 4-LAS caused at least a considerable portion of participants to respond in a way that

reduces the complexity of the system to a sort of BAS with a conservative threshold. This suggests that the 4-LAS induced behavioral strategies like the con-BAS and, thus, might also be effective in countering the cry wolf effect, albeit not as much as the con-BAS.

Implications

The con-BAS appeared to be the most effective system in terms of not only preventing a cry wolf effect in response to alarms, but also in keeping the number of misses low, resulting in the overall best joint human machine sensitivity. Thus, the implication of this research seems to be quite simple: there is no need for additional alert stages in alarm systems. Instead, thresholds in BASs should be set more conservatively. At least this seems to hold true in situations where the cry wolf effect cannot be prevented by other interventions (e.g., availability of alarm verification information; Manzey et al., 2014).

However, from a practitioner's perspective there is a flip side of using conservative BASs. Even though the number of joint human machine misses might be reduced tremendously, implementing a con-BAS would mean to provide a miss prone system. This would directly contradict the common fail-safe engineering approach, and there are only few contexts conceivable where this might be different. One is the medical domain where critical events tend to evolve over time. Here, more conservative thresholds would only introduce a delayed response but don't seem to increase the occurrence of missed events. Thus, the introduction of a con-BAS would not necessarily mean to have a miss prone system in strict sense but might help to reduce issue of alarm fatigue (Welch, 2011). However, for the most contexts it seems highly doubtful that any developers will design miss prone systems when they could be held responsible for critical events not indicated by the system. The current research suggests that the provision of 4-stage LASs constitutes a good compromise here. They guide human behavior towards a very high compliance, and, thus, do not lead to issues of cry wolf and alarm fatigue. At the same time, they allow designers to stick to the fail-safe engineering approach. Another solution to circumvent the problems of alarm fatigue

might be the provision of BASs with adaptable thresholds that leave the threshold setting with the operator. However, thus far, the effects of such adaptable alarm systems have rarely been addressed (e.g., Bustamante et al., 2007; Merkel & Wiczorek, 2012) and more attempts in this direction are eligible.

Key points:

- Behavioral and performance consequences of a conventional binary alarm system with liberal threshold setting (lib-BAS) were compared with three alternative alarm systems, i.e., a three- and a four-stage likelihood alarm system (3-LAS; 4-LAS) and a binary alarm system with conservative threshold setting (con-BAS).
- Compared to the lib-BAS, significant improvements in terms of a reduced cry wolf effect and an increased joint human machine sensitivity d' were found for the 4-LAS and the con-BAS.
- The con-BAS outperformed all other systems with respect to the number of misses.
- Both, 4-LASs and con-BASs provide possible means to counter negative side effects of conventional lib-BASs in terms of the cry wolf effect and the resulting risk of missing critical events.

REFERENCES

Balaud, M., & Manzey, D. (2014). The more the better? The impact of number of stages of likelihood alarm systems on human performance. In *Proceedings of the Human Factors and Ergonomics Society Europe Chapter Annual Conference* (pp. 61 -72).

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38*(11), 2300–2312. https://doi.org/10.1080/00140139508925269

Bliss, J. P. (2003). An investigation of extreme alarm response patterns in laboratory experiments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1683–1687). Santa Monica, CA: HFES https://doi.org/10.1177/154193120304701319

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

Bustamante, E. A. (2005). A signal detection analysis of the effects of workload, task-critical and likelihood information on human alarm response. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1513–1517). Santa Monica, CA: HFES. https://doi.org/10.1177/154193120504901702

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.

Bustamante, E. A. (2008). Implementing likelihood alarm technology in integrated aviation displays for enhancing decision making: A two-stage signal detection modelling approach. *International Journal of Applied Aviation Studies*, *8*(2), 241–262.

Bustamante, E. A., Bliss, J. P., & Anderson, B. L. (2007). Effects of varying the threshold of alarm systems and workload on human performance. *Ergonomics*, *50*(7), 1127-1147.

Clark, R. M., & Bustamante, E. A. (2008). Enhancing decision making by implementing likelihood alarm technology in integrated displays. *Modern Psychological Studies*, *14*(1), 36–49.

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: HFES.

Clark, R. M., Ingebritsen, A. M., & Bustamante, E. A. (2010). Differential Effects of Likelihood Alarm Technology and False-Alarm vs. Miss-Prone Automation on Decision-Making Accuracy and Bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (*pp. 1508–1512). Santa Monica, CA: HFES https://doi.org/10.1177/154193121005401932

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors*, *48*(3), 474-486.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, *56*(293), 52-64.

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.

Goel, P., Datta, A., & Mannan, M. S. (2017). Industrial alarm systems: Challenges and opportunities. *Journal of Loss Prevention in the Process Industries*, *50*, 23-36.

Graham, K. C., & Cvach, M. (2010). Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *American Journal of Critical Care*, *19*(1), 28-34.

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

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517-523.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583-621.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

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

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on 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. https://doi.org/10.1518/001872006777724408

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

McCarley, J. S., Rubinstein, J., Steelman, K. S., & Swanson, L. (2011). Estimating user's preferred response bias in an automated diagnostic aid: A psychophysical approach. In *Proceedings of the Human Factors and Ergonomics Society Annual*

Meeting (pp. 326–329). Santa Monica, CA: HFES https://doi.org/10.1177/1071181311551067

Merkel, C. & Wiczorek, R. (2012). Does higher security always result in better protection? An approach for mitigating the trade-off between usability and security. In D. Waard, K. Brookhuis, F. Dehais, C. Weikert, S. Röttger, D. Manzey, S. Biede, F. Reuzeau, and P. Terrier (Hrsg.), *Human Factors: a view from an integrative perspective* (pp.1-13).

Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(3), 343–353. https://doi.org/10.1518/0018720024497754

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*(4), 563–572. https://doi.org/10.1518/001872001775870395

Meyer, J., Bitan, Y., Shinar, D., & Zmora, E. (1999). Scheduling of actions and reliance on warnings in a simulated control task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 251–255). https://doi.org/10.1177/154193129904300326

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

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). https://doi.org/10.1177/1071181312561458

Sendelbach, S., & Funk, M. (2013). Alarm fatigue: a patient safety concern. *AACN advanced critical care*, *24*(4), 378-386.

Shurtleff, M. S. (1991). Effects of specificity of probability information on human performance in a signal detection task. *Ergonomics*, *34*(4), 469–486. https://doi.org/10.1080/00140139108967330

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 332–336). https://doi.org/10.1177/154193120204600325

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*(4), 445–459.

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

Vargas, J. C., & Bustamante, E. A. (2011). Moderating effects of alarm technology, type of automation, and information processing stage on decision making in UAS operations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 26–30). https://doi.org/10.1177/1071181311551006

Welch, J. (2011). An evidence-based approach to reduce nuisance alarms and alarm fatigue. Biomedical instrumentation & technology, 45(s1), 46-52

Wickens, C., & Colcombe, A. (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*(5), 839–850.

Wiczorek, R. (2017). Investigating users' mental representation of likelihood alarm systems with different thresholds. *Theoretical Issues in Ergonomics Science*, *18*(3), 221–240. https://doi.org/10.1080/1463922X.2016.1207209

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*, *56*(7), 1209–1221. https://doi.org/10.1177/0018720814528534

Wiczorek, R., Balaud, M., & Manzey, D. (2015). Investigating benefits of likelihood alarm systems in presence of alarm validity information. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 850-854).

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* (pp. 380-384).

615 Anna Zirk works as a research assistant in the Junior Research Group FANS at Technische
616 Universität Berlin in Germany. She earned her master's in human Factors in 2013 from the
617 Technische Universität Berlin.
618
619 Rebecca Wiczorek is the leader of the Junior Research Groups FANS at Technische
620 Universität Berlin and earned her PhD in Human Factors from the Technische Universität
621 in Berlin in 2012.
622
623 Dietrich Manzey is professor in the Department of Psychology and Ergonomics at the
624 Technische Universität Berlin. He earned his PhD in psychology from the University Kiel
625 in Germany in 1988.