Magali Balaud, Dietrich Manzey

The more the better? The impact of number of stages of likelihood alarm systems on human performance

Conference paper | Published version
This version is available at https://doi.org/10.14279/depositonce-8650

Balaud, Magali; Manzey, Dietrich (2015), The more the better? The impact of number of stages of likelihood alarm systems on human performance. In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference (pp. 61–72). Downloaded from http://hfeseurope.org (ISSN 2333-4959)

The more the better? The impact of number of stages of likelihood alarm systems on human performance

Magali Balaud & Dietrich Manzey
Technische Universität Berlin
Germany

Abstract

Responses to alarms involve decisions under uncertainty. Operators do not know if an alarm is more likely to be a hit or a false alarm. Likelihood alarm systems (LAS) help reduce this uncertainty by providing information about the certainty of their output. Unlike traditional binary alarm systems, they have three or more stages: each one represents a different degree of likelihood that a critical event is really present. Consequently, the more stages, the more specific is the information provided by the alarm system to reduce uncertainty. A laboratory experiment with 48 participants was conducted to investigate the effect of specificity of information of LAS on performances and responding behaviour. Specifically, a three-stage, four-stage, and five-stage LAS were compared using a multi-task environment. Results show higher percentages of correct decisions in the alarm task when participants used the four- and five-stage LAS than the three-stage LAS but no significant differences were found between the four-and five-stage LAS. Interesting differences in response patterns were also observed. This study suggests that four stages is the best degree of specificity for optimal performance.

Introduction

Alarm systems are extremely useful in multitasking and high workload environments such as aviation cockpits, hospitals and industries. They play a role of mediator between a human operator and a process, receiving information about the current status of a process and informing operators about it so that critical events are not missed. Most of the time operators work with Binary Alarm Systems (BAS) which inform the operator in a binary way: there is a critical event (red) or not (green).

Ideally, an alarm should go off only if there is a critical event. However this is not always the case. Instead alarms systems usually tend to generate a lot of false alarms, i.e. alarms go off even if there is no critical event. This is partly due to the "engineering fail-safe approach" (Swets, 1992): in order not to miss any critical events, engineers design the alarm system so it goes off even if there is little evidence of a critical event. A useful descriptor of the reliability of an alarm system is the Predictive Positive Value (PPV) (Getty et al., 1995). The PPV is the conditional probability that, given an alarm, a problem actually exists. A PPV of 0.3, e.g., means that out of all alarms emitted by the system, 30% are hits and 70% are false alarms. Given that alarm systems in most domains emit a high number of false alarms their PPV is usually low, often less than 0.1 (Parasuraman & Riley, 1997).

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

As a consequence, operators might stop trusting them (Madhavan, Wiegmann & Lacson, 2006). In behavioural terms, this can lead to what has been referred to as the cry-wolf effect (Breznitz, 1984; Getty et al., 1995; Bliss et al., 1995). Operators tend to respond slower or even to ignore the alarm system when it goes off. This can result in dramatic consequences regarding the safety and productivity (Lee & See, 2004).

One possible solution to address this issue is the use of Likelihood Alarms Systems (LAS). This concept was first developed by Sorkin et al. (1988) to constitute an alternative to binary alarm systems. LAS are composed by three or more stages with each stage corresponding to a different likelihood that a critical event is present. In other words, each stage of LAS has a different PPV and communicates it to the operator through the use of different colours, wordings, or sounds.

The goal of LAS is to provide more differentiated information to operators than traditional binary alarm systems so that they can adapt their responding behaviour depending on how likely it is that a critical event is present. By adapting their responding behaviour properly to the PPV of each stage, operators have higher chances to correctly comply with hits and to correctly ignore false alarms produced by the alarm system. Previous laboratory studies have shown that participants respond less to LAS in comparison to BAS but that they are more accurate: operators produce more hits and fewer false alarms with LAS in comparison to BAS (Bustamante & Bliss, 2005; Wiczorek & Manzey, 2014).

This raises the question of what degree of specificity, i.e. number of stages of LAS, is optimal for operators. Two studies (Shurtleff, 1991; Wiczorek et al., 2014) have already investigated this question. Shurtleff compared a BAS, a 4-stage LAS, a 6-stage LAS, an 8-stage LAS, and a control condition in which participants did not get any advice from any alarm system. The difficulty of the decision task was also manipulated. Results show that only when the task is difficult does the number of stages on participant's performance have an effect. Participants showed better performance while using 4-stage LAS and 8-stage LAS than BAS or no alarm. Wiczorek et al. (2014) compared a BAS, a 3-stage LAS, and a 4-stage LAS supporting a monitoring task as part of a multi-task scenario. They found that participants made less incorrect decisions (i.e., misses and false alarms) when they used the 4-stage LAS, followed by the 3-stage LAS and the BAS.

The current study

The current study investigates the optimal number of stages in Likelihood Alarm Systems on participants' responding behaviour, participants' performance and participants' workload. Using the same task environment than Wiczorek et al. (2014), the aim of this study was to replicate their findings using different PPV alarm characteristics and to further investigate the question of the optimal number of stages in LAS by comparing a 3-stage, 4-stage, and 5-stage LAS. The 3-stage LAS was composed by a non-alarm stage, a warning stage, and an alarm stage. Based on that, the 4-stage LAS was created by dividing the warning stage in two stages while the alarm stage was kept constant. The same logic applied in order to make the 5-

stage LAS: the stage of the 4-stage LAS having the lowest PPV (i.e., the yellow-warning stage) was split into two stages.

The following hypotheses were addressed: Firstly, it was hypothesized that participants would adapt their responding behaviour to the PPV of each stage so that participant's response rate in each stage will significantly be different from the others. Secondly, a differentiation in participants' behaviour would be expected between the 3-stage LAS and the 4-stage LAS. Specifically, it was assumed that the cry-wolf effect would be shifted from the warning stage of the 3-stage LAS to the low-PPV warning stage of 4-stage LAS and that participants would comply more with the high-PPV warning stage of the 4-stage LAS than with the warning stage of the 3-stage LAS. A similar effect was expected between the 4- and 5-stage LAS.

Thirdly, regarding participants' performance in the alarm task, a main effect of the number of stages on participants' decision-making performance was expected. The more stages, the better participants' performance would be in terms of the percentage of hits and false alarms. More specifically, participants' percentage of hits would increase with the number of stages and participants' percentage of false alarms would decrease with the number of stages.

Fourthly, with respect to participants' performance in the concurrent tasks, a decrease of performance was expected in the 5-stage LAS condition only. As too much specificity (stages) in the alarm display might increase the workload and time-demands of decision-making in response to the alarm system, it was assumed that increasing specificity might negatively impact operators' ability to deal with concurrent tasks. Since Wiczorek et al. (2014) did not find any difference between the 3-stage LAS and the 4-stage LAS on concurrent tasks performance, a visible decrease of performance was expected only for the most complex 5-stage LAS. Finally, it was expected that the more stages the LAS have, the higher participants' workload would be.

In addition to the hypotheses-driven questions, participants' overall response rate towards alerts (i.e., alarms and warning together) was also investigated in an exploratory manner, in order to know to what extent the number of stages of LAS would impact the cry-wolf effect.

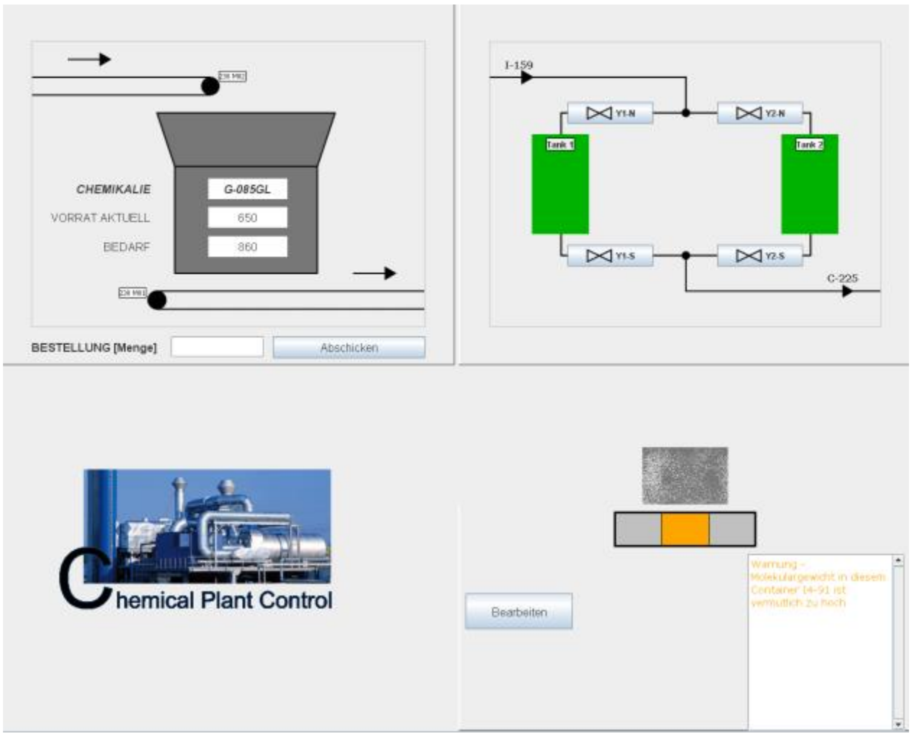
Method

Participants

Forty-eight participants (22 men, 26 women) participated in this study. Participants ranged in age from 18 to 44 years with a mean age of 27.02 years (SD = 5.77). None of them was suffering from any distortion of colour vision which might interfere with the experiment (i.e. red-green colour blindness). Participants were paid 5€ for their participation and they could get an additional bonus of maximum 4€ depending on their performance during the experiment.

Task

The PC-based Multi-Task Operator Performance Simulation (M-TOPS) was used. It simulates in a simplified way typical multi-task demands of operators in a control room. Participants had to accomplish three tasks simultaneously. In one of these tasks, they were assisted by an alarm system. A picture of the M-TOPS interface is shown in Figure 1.



CHEMIKALIE G-085GL
VORRAT AKTUELL 650
BEDARF 860
BESTELLUNG [Menge]

Warnung - Molekulargewicht in diesem Container I4-91 ist vermutlich zu hoch

Figure 1. User interface of M-TOPS

Resource Ordering Task (ROT). This task is a mental arithmetic task displayed in the upper left quadrant of the interface. Participants are instructed that they have to ensure the availability of required chemicals in order to keep the chemical process running. For this purpose, the current and the required value of a chemical is presented. Participants are asked to calculate the arithmetic difference, type the result in the designated ordering field, and initiate the order by clicking a button. They received 1.5 cents for each correctly sent order.

Coolant Exchange Task (CET). This task is displayed on the upper right quadrant of the interface. Participants are responsible for exchanging the coolant in different sub-systems of the plant. To do this they have to open and close a few valves by clicking on them following a certain order. A complete exchange cycle takes about 40 seconds. Participants received 7.5 cents for each refilling cycle successfully completed.

Alarm Task (AT). In this task displayed in the lower right quadrant of the interface, participants have to decide if the final quality of the chemical product has a correct molecular weight. They are assisted by an LAS showing a different *colours* and wordings depending on how likely it is that the chemical product has an improper molecular weight. Based on the diagnostic of the LAS participants choose between sending the container back to the plant (by clicking on the repair button) or letting it go (by doing nothing). Participants have no other cues apart from the output of the alarm system to help them in their decision. They lose 2 cents for each wrong decisions (i.e., repairing a correct container or ignoring an improper container). This pay-off was chosen based on a precise analysis of how much time participants spend on each task. It aims to keep a constant competition between the different tasks so that no task is left out for strategic reasons. The same pay-off was also used in the works of Wiczorek & Manzey (2014) and Wiczorek et al. (2014).

Design and alarm systems characteristics

The experimental design was composed of a single between-subjects factor defined by the number of stages of the likelihood alarm system supporting the alarm task. This factor had three levels: 3-stage (LAS3), 4-stage (LAS4), and 5-stage (LAS5). All alarm systems had the same sensitivity (d = 1.8). The basic characteristics of the three alarm systems used are presented in Figure 2. The first criterion separating the non-alarm stage (“green”) from the other stages was kept constant for all systems (c = -1.05). The numbers reported in the squares correspond to the PPV of each stage and the number reported under each separation corresponds to the criterion. The colours presented in this figure are the colours used for the outputs of the LAS. They were chosen according to findings from previous studies investigating the link between colours and perceived urgency or perceived hazard (Braun & Silver, 1995; Chapanis, 1994; Wolgater et al., 2002).

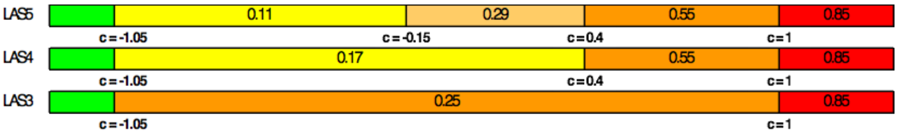


Figure 2. Systems characteristics of the three LAS

Dependent variables

Alarm task response behaviour: Possible differences of participants’ responses to the different stages of the different LAS were assessed by their compliance rates with each stage. Compliance rate was defined as the percentage of alerts emitted by each stage which was responded to by a click on the repair button.

Alarm task performance: Participants’ performance in the alarm task was assessed by the average percentage of hits and false alarms achieved by the participants in interaction with the different LAS. A high percentage of hits as well as a low percentage of false alarms is considered as good performance.

Concurrent tasks performance: Participants' performance in the concurrent tasks was measured by the amount of correctly sent orders in the Resource Ordering Task (ROT) and the amount of refilling cycles successfully completed in the Coolant Exchange Task (CET).

Subjective workload: Participants' perceived workload was assessed using the NASA Task Load Index (Hart & Staveland, 1988). The mean of all six single scales was considered as overall workload measure.

Procedure

Participants first completed an informed consent form and a demographic questionnaire and were then provided with the task instructions on the computer screen. They were told that the experiment was a simulation of a control room of a chemical plant and that they had to perform three tasks concurrently in order to assure the good run of the chemical process and to control the quality of the end-product. Participants had a 2-minute training for each single task. They were then explained that the alarm system was not 100% reliable and that it could sometime provide wrong outputs. This was followed by a 50-trial familiarization session (about 8 minutes) in which participants performed the alarm task only and received an auditory feedback after each decision in response to the outputs of the alarm system they made. The feedback informed them about the correctness of their decision and, thus, implicitly also about the performances of the alarm system. They were told to use this auditory information to get an idea of the reliability of the different stages of the LAS. Participants were then explicitly asked for a subjective assessment of the reliability of each stage of the LAS they had worked with. This was used as a manipulation check to ensure that participants paid attention to the auditory feedbacks in the familiarization session and recognized the differences in PPVs of the different stages. The experimental session finally started. It was composed of 100 containers (about 16 minutes). No auditory feedbacks were provided during this session. Finally participants completed the NASA TLX questionnaire, were thanked for their participation and received a monetary compensation.

Results

Participants' response behaviour

Response rates for the 3-stage LAS

Response rates to the two alert stages of the LAS3 (alarm vs. orange-warning) are shown in Figure 3. As expected participants on average complied more with alarms (98.56%) than warnings (16.51%). This difference was proven to be statistically significant by a two-tailed t-test, *F*(1,15) = 120.58, *p* = .000.

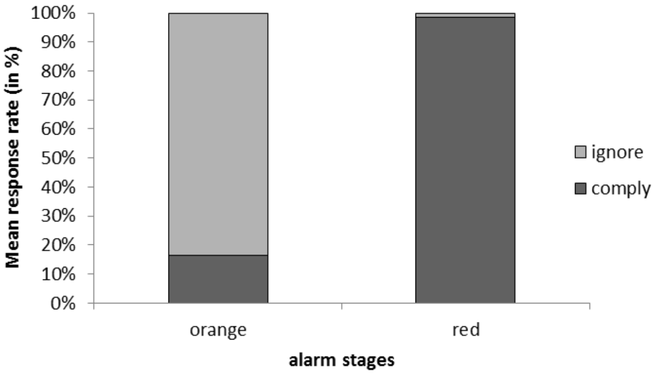


Figure 3. Means of compliance rates and non-compliance rates towards the 3-stage LAS depending on the diagnosis emitted by this LAS.

Response rates for the different stages of LAS4

Mean response rates for the three alert stages of the LAS4 (alarm vs. orange-warning vs. yellow-warning) are displayed in Figure 4. As becomes evident, response rates differed between stages. A one-way ANOVA with stage (red-alarm, orange-warning, yellow-warning) as within factor was used to analyse this effect. This was composed by a linear contrast C1(-1, 0, 1) and a quadratic contrast C2 (-1, 2, -1). The linear contrast was significant suggesting that participants complied more with alarms than yellow-warnings, $F(1, 15) = 111.68$, $p = .00$. However the quadratic trend was also significant showing that participants' compliance rate towards orange-warnings differed from the linear trend, $F(1, 15) = 111.03$, $p = .00$. The significance of the quadratic trend is explained by the high compliance rate observed with orange-warnings (97.16%), which does not significantly differ from participants' compliance rate with alarms (96.63%), $F(1, 15) = .10$, $p = .76$.

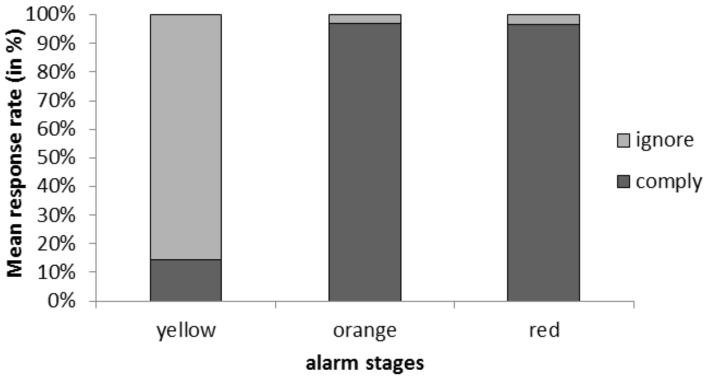


Figure 4. Means of compliance and non-compliance rates toward the 4-stage LAS depending on the diagnosis emitted by this LAS.

Response rates towards different stages of LAS5

Results are displayed in Figure 5. A one-way ANOVA with stage (alarm vs. orange-warning vs. orange-yellow-warning vs. yellow-warning) as within factor was used for the analysis of the response rate toward LAS5. A linear contrast C1 (-3, -1, 1, 3), a quadratic contrast C2 (-1, -1, 1, -1) and a cubic contrast (-1, 3, -3, 1) were used to test how specifically participants' responses to the different stages depends on the PPV of each stage. The linear trend is significant, *F*(1, 15) = 120.34, *p* = .00, as well as the cubic trend, *F*(1, 15) = 5.31, *p* = .04. This means that the pattern of results is not completely linear as expected. The high compliance rate obtained in the orange-warning stage is responsible for the significance of the cubic trend. This was confirmed by the fact that participants' compliance rate did not differ in the orange-warning stage and the red-alarm stage, *F*(1, 15) = 2.46, *p* = .14.

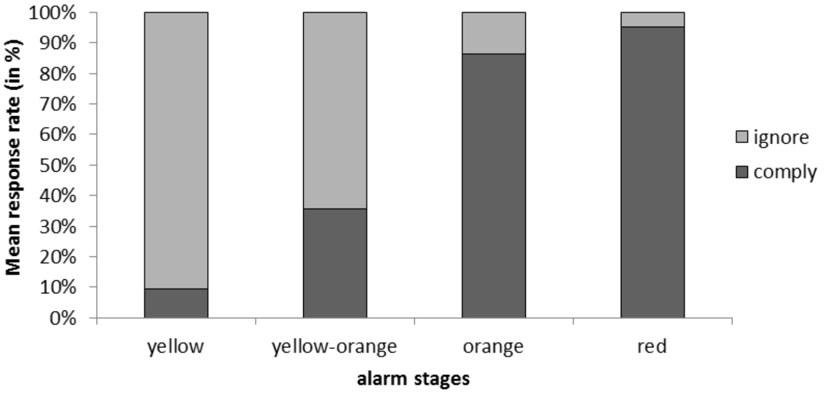


Figure 5. Means of compliance and non-compliance rates toward the 5-stage LAS depending on the diagnosis emitted by this LAS.

Comparisons of response behaviour across different LAS

A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used for the analysis of the response rate toward alerts. Even though participants complied more with LAS4 (44.51%) and LAS5 (44.70%) than with LAS3 (32.67%), these difference were not significant, *F*(2, 45) = 1.7, *p* = .19. This means that the cry-wolf effect, in terms of number of percentage of ignored alerts, was the same among the three LAS.

However a behavioural differentiation was observed as expected in Hypothesis 2. Participants complied significantly more with the orange warning stage of LAS4 (97.16%) than with the warning stage of LAS3 (16.51%), *F*(1, 30) = 107.91, *p* = .00. Moreover, participants complied significantly more with orange warnings of LAS4 than yellow warnings of LAS4, *F*(1, 15) = 116.64, *p* = .00, showing that the cry-wolf effect in LAS4 was reduced to the yellow-waning stage only. A shift of the cry-wolf effect from the warning stage of LAS3 to the yellow warning stage of LAS4 happened.

Regarding LAS4 and LAS5, participants did not significantly complied more with the yellow-orange warning stage of LAS5 (35.71%) than with the yellow warning stage of LAS4 (14.58%), *F*(1, 16) = 2.65, *p* = .11, even though descriptive results show this tendency. A behavioural differentiation occurred still between the yellow warning stage and the yellow-orange warning stage of LAS5. Participants complied significantly more with the yellow-orange warning stage (35.71%) than the yellow warning stage (14.58%), *F*(1, 16) = 7.27, *p* = .02.

Alarm-Task performance

All analyses about participants' performance in the alarm task were performed using a one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor. Two orthogonal contrasts were defined for pairwise comparisons of means: C1 (2, -1, -1) and C2 (0; -1; 1). The first contrast C1 compares the mean performance for LAS3 with the combined mean performances for LAS4 and LAS5. The second contrast C2 tests if performances in conditions LAS4 and LAS5 would differ from each other.

Participants' percentage of hits and false alarms are displayed in Figure 6. Two participants were excluded from the analysis on the percentage of hits based on their outlying SDR and Cook values. One participant was excluded from the analysis on the percentage of false alarms for the same reasons.

Regarding the percentage of hits, results did not show a linear trend as it was predicted. As expected, participants using LAS3 produced significantly less hits (17.64%) than participants using LAS4 (26.33%) but participants using LAS5 (26.42%) did not produce more hits than participants using LAS4. This pattern is also confirmed by the two contrasts, C1: *F*(1, 44) = 52.91, *p* = .00, C2: *F*(1, 44) = 0.01, *p* = .94 (C2).

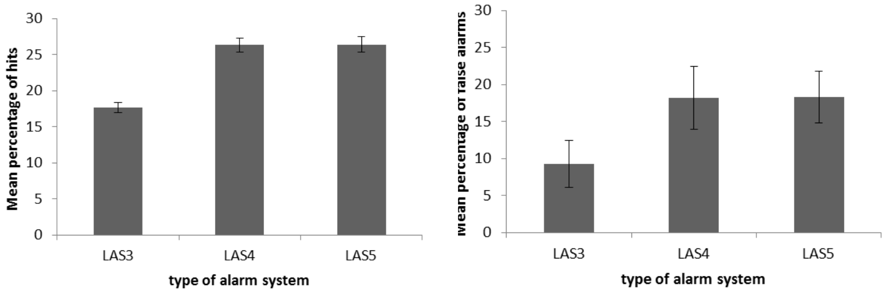


Figure 6. Means and mean standard deviations of participants' percentage of hits (left panel) and false alarms (right panel) in the alarm task depending on the type of LAS.

Regarding participants' percentage of false alarms, the best performance (i.e., the lowest percentage of false alarms) was observed in the LAS3. Participants using LAS3 produced less false alarms (9.29%) than participants using LAS4 (18.18%) and LAS5 (18.27%), *F*(1, 45) = 3.84, *p* = .05 (C1). No difference between the LAS4 and LAS5 condition has been found, *F*(1, 45) = 0.00, *p* = .99 (C2).

Concurrent task performances

A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used to analyse the performance data of the two concurrent tasks. No significant differences between the three conditions were found in both tasks: ROT: *F*(2, 44) = 0.41, *p* = .66; CET: *F*(2, 45)= 0.06, *p* = .943.

Workload

A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used for the analysis of the participants' workload. There is no main effect of number of stages on participants' workload ratings, *F*(2, 45) = 1.05, *p* =.36. No effect was found on any single scale of the NASA TLX.

Discussion

This study aimed to investigate what number of stages of likelihood alarm systems would provide the optimal specificity of information for human performance in interaction with such systems. Specifically, the effect of three different LAS on responding behaviour, performance and workload was investigated. The LAS differed with respect to the number of stages.

Participants adapted only partially their responding behaviour to the PPV of each stage. This means that the pattern of results is not exactly linear but shows a kind of dichotomization. Participants tend to clearly differentiate their responding behaviour depending on the PPV towards stages having a PPV under .5. This tendency of operators to adjust their response behaviour to the PPV of alerts at the lower end of PPVs was also reported by other studies addressing the impact of PPV on responses to alarms of BAS as well as studies investigating different stages of LAS (Manzey et al., 2014; Wiczorek & Manzey, 2014; Wiczorek et al., 2014). However participants tend to consistently comply with alerts emitted by stages having a PPV above .5. Participants complied with more than 93% of orange warnings emitted by the LAS4 and LAS5 even though the PPV is .55. This high compliance rate is actually a rational strategy in order to optimize the amount of correct decisions in interaction with alarm systems and is very surprising, as such high response rates are usually observed in stages having a PPV above .7 (Wickens & Dixon, 2007). Interestingly, adding more stages to LAS does not reduce the cry-wolf effect. However, while participants' overall response rate was the same for the three LAS, their overall decision-making performance in terms of hits clearly benefited from going from an LAS3 to an LAS4. By adding one more stage, thus providing more differentiated likelihood information, participants get more opportunities to differentiate their behaviour. The ignorance of alert, i.e. the cry-wolf effect, still occurs but is shifted to a stage having a lower PPV and thus shifted to a stage where an ignorance of the alert often matches an alert which is false anyway. As a consequence, participants comply more with true alarms and ignore more false alarms even though the overall response rate to alerts stays the same. Studies comparing BAS to LAS3 have even shown that participants' overall response rate is higher with BAS than LAS but performance is still better with the LAS which is attributed to essentially the same effect (Bustamante & Bliss, 2005; Manzey et al. 2014).

Regarding participants' performance in the alarm task, they showed better performance with the LAS4 and the LAS5 than the LAS3 with respect to the percentage of hits. However, no significant differences emerged between the LAS4 and LAS5. Against our expectations, participants had lower performance with the LAS4 and LAS5 than the LAS3 with respect to the percentage of false alarms. This is in contradiction with results reported by Wiczorek et al. (2014) showing that participants produce fewer false alarms with the LAS4 than the LAS3. The high response rate toward orange warnings in the LAS4 and LAS5 might explain these results. By complying with more than 93% of warnings having a PPV of .55, participants produced a great amount of false alarms in comparison to participants in the LAS3 condition who mainly ignored the .25 PPV warnings and produced mostly correct rejections. However the percentage of hits is a more relevant performance indicator to consider than the percentage of false alarms since most alarms systems are used in environment in which misses are more costly than false alarms. From these results, one can draw the conclusion that LAS4 improve performance over LAS3 and that adding one more stage (LAS5) does not improve performance further.

No effect of the number of stages in LAS has been found on participants' performance in the concurrent tasks. This is probably due to the fact that participants' workload did not increase with the greater amount of information provided by the LAS5. Indeed no difference between the three LAS on participants' workload has been found. It would be interesting, however, to know if a higher number of stages affect the workload since alarm systems having more than 5 stages are sometimes used in ecological environments.

Conclusion

Likelihood alarms systems are definitely an option to consider in situations in which the use of a BAS leads to a high cry-wolf effect with the performance effect of decreasing hit rates. This study suggests that a 4-stage LAS provides the optimal degree of specificity and that a higher degree of specificity does not improve performance. However, one limiting factor of the current research was that the participants did not get the opportunity to cross-check the validity of alarms before responding to it. Previous research has shown that providing such an option might significantly impact the response behaviour in interaction with alarms (e.g., Manzey et al., 2014). Further research is needed to investigate if the results reported in this study could be generalized to situations in which operators have access to alarm validity information.

References

Bliss, J., Dunn, M., & 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.

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, NJ: 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.

Getty, D.J., 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.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139-183.

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

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.

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*, 1833-1855.

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

Shurtleff, M.S. (1991). Effects of specificity of probability information on human performance in a signal detection task. *Ergonomics*, *34*, 469–486.

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., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*, 201–212.

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors*, *56*, 1209-1221.

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: HFES.

Wogalter, M.S., Conzola, V.C., & Smith-Jackson, T.L. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, *33*, 219–230.