Benefits of Decision-Support by Likelihood versus Binary Alarm Systems: Does the number of stages make a difference?

Rebecca Wiczorek[1], Dietrich Manzey[2] & Anna Zirk[3]
[1]Tel Aviv University, Tel Aviv, Israel
[2]Technische Universität Berlin, Berlin, Germany
[3]Berlin Institute for Social Research, Berlin, Germany

Recent research has shown that the use of 3-stage likelihood alarm systems (LAS) has the potential to mitigate performance deficits associated with the use of binary alarm systems (BAS). The additional likelihood information can guide operators' behavior and improve their decision-making accuracy. Comparisons of LAS with different numbers of stages are missing so far. Therefore, the current study compared a BAS with a 3-stage LAS and a 4-stage LAS. Participants were found to make significantly fewer wrong decisions with the 4-stage LAS than with the other two systems, and still significantly fewer errors with the 3-stage LAS compared to the BAS. We found that this performance benefit resulted from a reduced number of false alarms, whereas no difference was found with regard to misses. Results are further discussed with regard to their theoretical implications for LAS and threshold setting in BAS.

INTRODUCTION

Work demands of operators in domains like aviation or process control typically include concurrent performance of a number of tasks and supervisory control of different systems in parallel. In order to support performance in such complex work environments, alarm systems are implemented that shall guide operators' decision-making in terms of proper task prioritization and attention allocation. The most basic form of such systems is referred to as binary alarm systems (BAS). BAS consists of two stages. They remain silent (or show a green light) as long as a monitored parameter or assessed state can be regarded as "normal", and provide an alerting signal (e.g. red light) in case of a critical event. Theoretically, these systems allow operators to withdraw attention from the automatically monitored processes without committing any risk of missing a critical state, as long as they run normally (no alarm emitted). As a consequence, they have more resources for working on concurrent tasks. However, in practice, the design and implementation of alarm systems provide a number of human factor issues that can interfere with their efficiency of behavioral guidance. One of these issues is related to the fact that designers of alarm systems often select liberal thresholds for emitting an alarm, i.e. make the alarm system most sensitive, in order to ensure that no critical states or events are missed. However, the reverse side of this is that such alarm systems inevitably generate a large number of false alarms. Thus, their positive predictive value (PPV), i.e. the posterior probability p(E|A) that there really is a critical event (E) in case of an alarm (A) can become relatively low (e.g. Getty, Swets, Pickett & Gonthier, 1995; Parasuraman, Hancock & Olofinboba, 1997).

Research has shown that operators' decision-making in response to BAS is directly governed by this PPV (e.g. Bliss, Gilson & Deaton, 1995; Gérard & Manzey, 2010; Getty et al., 1995). More specifically, it has been shown that operators interacting with alarm systems with a comparatively low PPV < 0.5 tend to respond slower (e.g. Getty et al., 1995) or even ignore the alarms completely (e.g. Bliss et al., 1995). This effect has been referred to as cry wolf phenomenon (Bresnitz, 1984) and presents one of the main human factors issues in relation to alarms because it raises the risk to response too late to or to even miss critical events.

A more sophisticated concept of a multiple-stage likelihood alarm system (LAS) has been proposed to circumvent this problem and to improve decision-making accuracy in response to alarms (Sorkin , Kantowitz & Kantowitz, 1988). The basic idea of LAS is to enlarge the number of stages of the alarm system in order to provide the operator more differentiated information about the relative likelihood of an alarm to truly indicate a critical event. The simplest example of LAS is a 3-stage alarm system with a first threshold separating a non-alert state, i.e. indication of normal operation (green light), from an alert state (red light) in which a critical event is given with a certain probability. Yet, this alert stage is again divided by another threshold; separating the *alarm*-stage from a *warning*- stage (e.g. often indicted by a yellow or amber light). A warning indicates that a critical event *might be* present. The PPV of such a warning-stage is lower, and the PPV of the corresponding alarm-stage is higher than the PPV of a BAS with a comparable first threshold.

It is expected that providing this more differentiated information implicates possible benefits. Wiczorek & Manzey (2014) argue that, compared to typical BAS, LAS do not provide high numbers of false *alarms*, but only generate high numbers of false *warnings*. As a warning just indicates that there *might be* a critical event, the absence of that critical event in case of a given warning does not necessarily prove the system's diagnosis as false. As a consequence false warnings might not be considered as system errors and therefore the risk of the cry wolf effect might be reduced. This assumption is in line with Bustamante's (2008) supposition. He states that, given operators' tendency to match their response behavior to the PPV of the different sorts of alerts, it can be expected that they show a high compliance with alarms and a

considerably less compliance with warnings. Compared to BAS this should result in a more purposeful behavior and, thus, better overall decision-making accuracy in response to alerts.

Thus far only few studies have investigated the human performance consequences of LAS compared to BAS (Bustamante, 2008; Bustamante & Bliss, 2005; Bustamante & Clark, 2010; Clark, Peyton & Bustamante, 2009; Ragsdale, Lew, Dyre & Boring, 2012; Sorkin et al., 1988; Vargas & Bustamante, 2011). Overall they provide empirical evidence that LAS can make human operators respond more often to true alerts and less often to false alerts, compared to BAS. This seems to be particular true if overall workload imposed by concurrent tasks is high and the alarm system is prone to false alarms. Direct evidence for this benefit being related to an appropriate differentiation of behavior in response to warnings and alarms is provided by a recent study of Wiczorek & Manzey (2011, 2014). They investigated the response behavior to alerts emitted by BAS and LAS with same first thresholds as part of a complex dual-task scenario. The BAS had a PPV of .43, the warnings of the LAS had a PPV of .29 and the alarms of the LAS a PPV of .88. Participants of this study behaved in the expected way. Working with the BAS, they showed a cry wolf effect by responding to only about 70% of all alarms. In contrast, with the LAS they responded to nearly all of the alarms emitted by the LAS (99%), but only to 31% of the warnings. This corresponds to an overall rate of responses to alerts in the LAS condition of 47%. These results suggest that LAS might even amplify the cry wolf effect reflected in an increased number of ignored alerts in total, but shift it to the warning-stage where ignorance of an alert has a higher likelihood to be a correct response. This provides evidence that benefits of LAS indeed are due to the provision of more differentiated diagnostic information, triggering more purposeful responses to alerts.

However, more detailed knowledge about most advantageous characteristics of LAS in terms of threshold setting or number of stages is still missing. Apart of the study by Sorkin et al. (1988), most of the studies conducted thus far, have only addressed the simplest type of 3-stage LAS. This leaves the question to what extent decision-making in response to alerts might become even more adequate by adding additional stages to an LAS, i.e. by providing even more specific diagnostic information.

The current study addresses this question by comparing the behavioral and performance consequences of a BAS with a 3-stage and a 4-stage LAS. Using the same dual-task paradigm as Wiczorek & Manzey (2011, 2014) it was investigated how the number of stages of the alarm systems would impact participants' response to alerts and their decision-making accuracy. We predicted that both types of LAS would lead to better decision-making accuracy than the BAS. Furthermore, we predicted that decision-making accuracy would be higher with the 4-stage than the 3-stage LAS. This latter effect was expected because the 4-stage LAS, on the one hand provides an even more differentiated diagnostic information than a 3-stage LAS which in turn might lead to an even better focused response behavior. On the other hand it includes a sufficiently low number of stages to make efficient decision-making still possible (Sorkin et al., 1988).

METHOD

Participants

44 master students (20 female, 23 male, mean age: 26.30 years) took part in the study. They were randomly assigned to one of three conditions and received a basic compensation of €10 or ECTS credits for participation. All of them received an additional payment of up to €8, depending on their performance.

Task Environment

M-TOPS (Multi-Task Operator Performance Simulation), a PC-based laboratory simulation environment was used for the experiment (see Figure 1). It represents a low fidelity simulation of cognitive requirements of control room operators in chemical plants. The two tasks used in the present study were the Ordering Task (upper left side of the experimental screen) and the Alert Task (bottom right side).

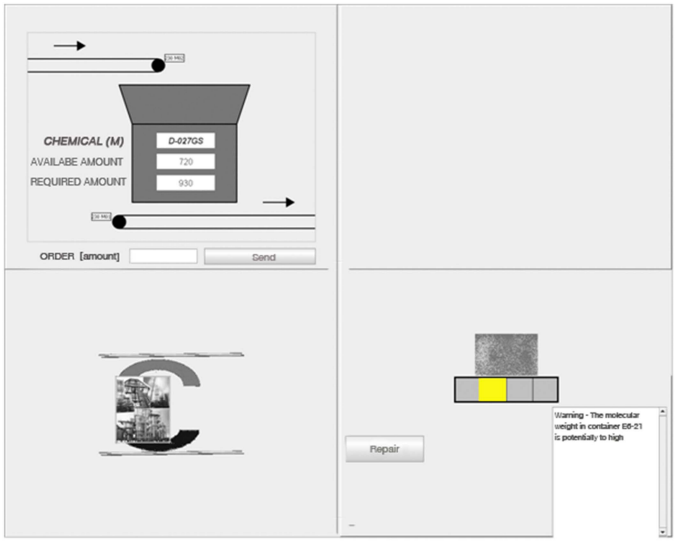

Figure 1. Screenshot of M-TOPS with the Ordering Task on the upper left side and the Alert Task on the bottom right side

Ordering Task

The objective of this task is to order a certain amount of chemicals, depending on the deviation of a given and a required amount (see Figure 1). Participants have to calculate the difference of two 3-digit numbers, type the result in, and send the proper order by clicking the *send* button within 15 seconds. After sending an order, the task is blanked out and a new set of required and available amount of the next chemical appears automatically after one second. Responses of the participants are logged.

Alert Task

The objective of this task is to control the appropriateness of the molecular weight of containers holding the chemical end-product. Participants are instructed that the

plant has a control system which automatically controls and assesses the quality of the chemical end-product filled in single containers. For this purpose every single container passes the control station (represented in the lower right quadrant of Figure 1), where the quality assessment is performed automatically. In case the quality of a given container fulfills a predefined quality criterion ("molecular weight ok"), a green light is emitted indicating that the state of the product is approved. However, in case the system detects an impaired product an alert is emitted. Depending on whether the underlying system represents a BAS, a 3-stage LAS (3-LAS) or a 4-stage LAS (4-LAS), these alerts are coded in different colors and are combined with different messages displayed on an alarm state monitor (see Table 1). After receiving the visual diagnose of the current container state, the operator has five seconds to decide whether or not to respond to the alert by clicking the *repair* button. This decision has to be made without access to any other system information. About six containers are checked per minute. Responses of the participants are logged.

Table 1. Colors and messages emitted by the different alarm systems

Diagnosis of the alarm system				
	The molecular weight is…			
Message on the alarm state monitor	*…ok*	*…potentially too high*	*…probably too high*	*…too high*
BAS	✓	-	-	✓
LAS 3	✓	✓	-	✓
LAS 4	✓	✓	✓	✓

Payoff

Every correct order in the Ordering Task was rewarded by 1.5 points. Every wrong decision in the Alert Task (repairing a container with an appropriate molecular weight or ignoring a container with an inappropriate molecular weight) was penalized by -2 points. This allocation of points was based on an analysis of the time structure and was chosen to produce a competition between both tasks. As performance dependent reward participants received two Euro Cents per point.

Experimental Design

A one-factorial between subjects design was used with the three different alarm systems representing the only factor. All three systems had the same sensitivity d'=1.7, in terms of signal detection theory (cf. Swets, 1964) and the base rate of critical events was the same for all three conditions, p(E)=.3. All systems generated 68 alerts in total out of 100 trials and had the same first threshold, resulting in a PPV of .43 for the BAS. The 3-LAS had one additional threshold leading to an alarm-PPV of .88 and a yellow warning-PPV of .29. The 4-LAS contained two additional thresholds, resulting in an alarm-PPV of .88, an amber warning-PPV of .5 and a yellow warning-PPV of .18.

Procedure

Participants completed the experiment in groups of up to four. All tasks were performed at a personal computer with a 17" screen. After providing written consent participants read a standardized instruction. In this it was said that they would be responsible for the Monitoring and the Ordering Task simultaneously and both tasks would be equally important. It was clarified that the alarm systems wouldn't work perfectly but very precise. Furthermore, subjects were acquainted with the rewarding system. Participants started with a training period. At first, the two tasks were separately practiced for two minutes each. Subsequently, the two tasks were practiced for two minutes in parallel. Afterwards, participants performed a practice block (ca. 20 min.) with 100 containers fulfilling only the Alert Task in order to build up a proper mental representation about the actual reliability of the current alarm system, which was the same as in the experimental block. Only in the practice block they got an acoustic feedback via headphone after every wrong decision. After the practice block they were delivered the true PPVs to avoid biases regarding reliability estimation. Afterwards, participants completed the experimental block (ca. 20 min) with 100 containers, during which Ordering and Alert Task had to be performed concurrently (no on-line feedback was provided). Finally, participants were paid and debriefed.

Measures

Alert response frequencies: Number of clicks on the repair button in response to red diagnoses (BAS), to red and yellow diagnoses (3-LAS), or to red, amber, and yellow diagnoses (4-LAS).

Performance in the Alert Task: Number of the two types of wrong decisions:

- misses – not repairing a container whose molecular weight was inappropriate;
- false alarms (FAs), i.e. unnecessary action – repairing a container whose molecular weight was appropriate.

Performance in the Ordering Task: Number of correct orders.

RESULTS

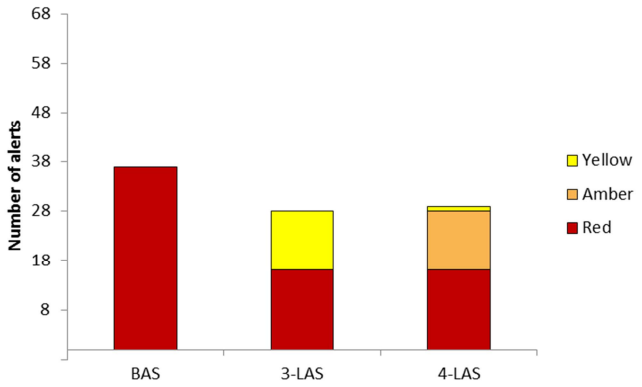

Figure 2. Response frequencies to the different types of alerts generated by the three alarm systems

Responses to Alerts

Response frequencies to the different types of alerts generated by the three alarm systems are shown in Figure 2. Participants interacting with the BAS exhibited the well-known cry wolf effect and only responded to 38 of the 68

alarms emitted by the system. Participants interacting with the two LAS matched their responses to the diagnostic value of the different types of alerts. In the 3-LAS condition, they responded to 16 of the 16 alarms but only to 12 of the 52 warnings. This equals a total response frequency to alerts of 28/68. The participants interacting with the 4-LAS responded to 16 of the 16 alarms, to 12 of the 18 amber warnings and only to 1 of the 34 yellow warnings, which corresponds to an overall response frequency of 29/68. However, inferential analyses by means of a one-way ANOVA with "type of alarm system" as between-subjects factor did not reveal a significant difference in the overall response rates to alerts between the three systems, $F(1, 41)=1.38$, NS.

Alert Task Performance

A significant difference between the three alarm systems emerged with respect to the number of wrong decisions made in response to alerts, $F(1, 41)=50.41$, $p<.0001$, $\eta^2_p=.71$. Post hoc Scheffè pairwise comparisons showed a significant difference between the BAS and both the 3-LAS and the 4-LAS with, $p<.0001$ and $p<.0001$, respectively. Furthermore, also the comparison of the 3-LAS and the 4-LAS revealed a significant difference, $p<.01$. As can be seen from Figure 3, the use of 4-LAS led to fewer wrong decisions than the use of 3-LAS, which still led to fewer wrong decisions than did the BAS.

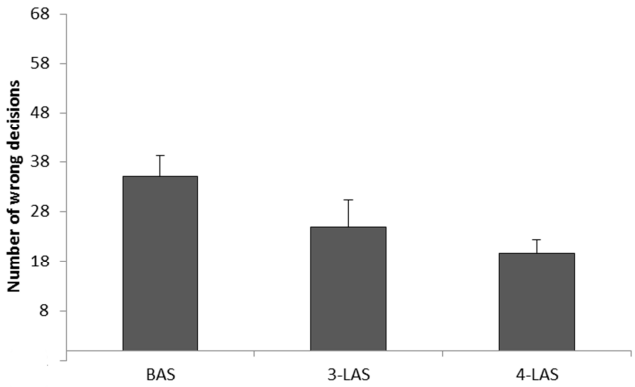

Figure 3. Wrong decisions in response to the three different alarm systems

A more detailed analysis suggests that this performance benefit of the two LAS was mainly due to a fewer number of FAs, i.e. unnecessary actions. Figure 4 shows that participants using the BAS committed more FAs than did the users of the 3-LAS and the 4-LAS. A separate one-way ANOVA comparing the number of FAs revealed a significant effect, $F(1, 41)=5.79$, $p<.01$, $\eta^2_p=.22$. Post hoc Scheffé comparisons revealed significant differences for the number of FAs only between the BAS and both, the 3-LAS, $p<.05$, and the 4-LAS, $p<.05$, whereas a single comparison between the two LAS failed to show significance. No such differences between the three systems were found for misses, $F(1, 41)=1.58$, NS.

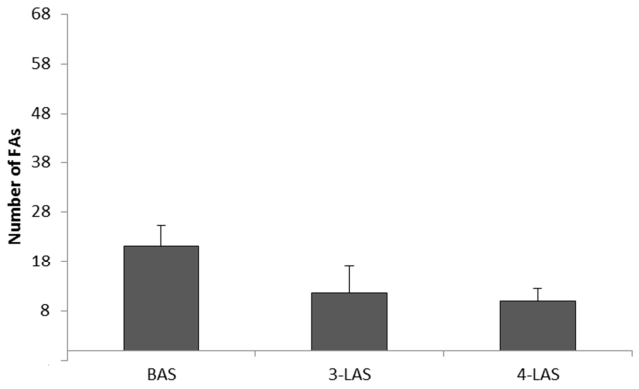

Figure 4. False Alarms (FAs) in response to the three different alarm systems

Concurrent Task Performance

No significant differences between the three types of alarm systems emerged with regard to ordering task performance, $F(1, 41)=.12$, NS. Neither the provision of a 3-stage nor 4-stage LAS provided any benefits for concurrent task performance as compared to the BAS. This might be due to the fact that the overall response rates to alerts did not differ significantly between the three systems, i.e. all three systems demanded a similar amount of attentional resources.

DISCUSSION

The aim of the current study was the comparison of a standard BAS with one LAS with 3 stages and one LAS with 4 stages. We investigated effects of the different alarm systems on participants' response behavior, their performance in the alert task and concurrent task performance. Moreover we had a closer look at the different types of errors, the users made in interaction with the different alarm systems.

It was predicted that the more differentiated diagnostic information provided by the two LAS would improve decision-making accuracy compared to BAS. Furthermore, it was expected that the 4-LAS would lead to even more improvements than the 3- LAS.

In line with our assumptions and prior research (e.g. Bustamante, 2008) we found performance with the 3-LAS to be better than with the BAS. In addition we could also find an advantage of the 4-LAS over the 3-LAS, as participants made fewer wrong decisions in interaction with the 4-stage system. This improvement of alert task performance was mainly due to the reduction of unnecessary actions in terms of FAs which took place in both LAS conditions, whereas a reduction of misses could not be found as a consequence of LAS use.

Finally, no differences in concurrent task performance were found related to the different alarm systems. This is in contrast to earlier findings with the same paradigm where benefits for concurrent task performance were found for a 3-stage LAS compared to a BAS (Wiczorek & Manzey, 2014). However, in this latter study the compliance rates to the alarms of the BAS for some reason were considerably higher than in the present research (70% vs. 55%), leading to a more pronounced difference with respect to the resources invested in dealing with alerts between the two types of alarm systems.

In summary, two main conclusions can be drawn from the results of the current study. First, it has been shown that not only LAS with 3 stages has advantages over a classical BAS but that the same holds true also for a 4-stage LAS. Second, the 4-LAS has also been found to have additional benefits over the 3- LAS in terms of improved decision-making accuracy. However, it is important to point out that performance improvements with both LAS were due to the reduction of FAs only, whereas the number of the more safety-relevant errors – the misses of critical events – could not be reduced by the use of LAS.

Whereas these findings in the first place suggest the interpretation that the number of stages of LAS plays a crucial role for the performance in the alert task, a closer look at the response rates for the different types of alerts also offers an alternative explanation. As can be seen in Figure 2, 4-LAS users responded to most of the alarms and amber warnings, whereas they did not respond to most of the yellow warnings. That could be interpreted as a dichotomization of behavior, where participants treated the yellow warnings like the green signals, interpreting them as a cue for a normal operation state rather than as a cue for a critical situation. In other words, it is possible that participants reduced the four stages into two, resulting in an underlying mental representation of a classical BAS.

Therefore it could be argued that participants only made use of the middle threshold for the adjustment of their own response criterion, while ignoring the other two system thresholds. In this case the benefit of the 4-LAS found in the present study would not have resulted from the additional stage, but rather as an effect of more conservative threshold setting. In this case the improved alert task performance would be due to an increase in compliance with alerts. For example, Merkel & Wiczorek (2012) provided evidence that a more conservative threshold setting in BAS can improve compliance with alarms without affecting the overall number of misses.

If this interpretation holds true, the same results as with the 4-LAS in the present study should be found by using a BAS with a liberal threshold setting, corresponding to the middle one of the 4-LAS system. Studies in our lab are currently being conducted to further investigate this hypothesis. However, if it turns out that these findings are not due to threshold setting but rather a result of the additional stage, future research should address whether more stages can lead to further improvements.

LITERATURE

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics, 38*, 2300–2312.

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

Bustamante, E.A. (2008). Implementing Likelihood Alarm Technology in Integrated Aviation displays for Enhancing Decision-Making: A Two-Stage signal Detection Modeling Approach. *International Journal of Applied Aviation Studies, 8*, 241-261.

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

Bustamante, E. A., & Clark, R. M. (2010). Differential Effects of Likelihood Alarm Technology, Type of Automation, and Type of Task on Decision Making as Applied to Aviation and UAS Operations. *International Journal of Applied Aviation Studies*, 10(1), 51.

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* (Vol. 53, pp. 349–353). Sage Publications.

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

Getty, D., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1(1)*, 19–33.

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 (Ed.), *Human Factors: a view from an integrative perspective* (pp.1-13).

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* (Vol. 56, No. 1, pp. 2167-2171). Sage Publications.

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.

Swets, J.A. (1964). *Signal detection and recognition by human observers*. New York: John Wiley & Sons.

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* (Vol. 55, No. 1, pp. 26-30). Sage Publications.

Wiczorek, R. & Manzey D. (2011). Evaluating Likelihood Alarm Systems as an Alternative to Binary Alarm Systems. In D. Waard, N. Gérard, L. Onnasch, R. Wiczorek & D. Manzey (Ed.), *Human centred automation* (pp.69-83). Maastricht: Shaker Publishing.

Wiczorek, R. & Manzey D. (2014). Supporting attention allocation in multi-task environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 0018720814528534.*