Linda Onnasch, Stefan Ruff, Dietrich Manzey

Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance

Subtitle

Journal article | **Accepted manuscript (Postprint)**
This version is available at https://doi.org/10.14279/depositonce-10989

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. International Journal of Human-Computer Studies, 72(10–11), 772–782. https://doi.org/10.1016/j.ijhcs.2014.05.001

Operators' Adaptation to Imperfect Automation – Impact of Miss-Prone Alarm Systems on Attention Allocation and Performance

Linda Onnasch, Stefan Ruff, Dietrich Manzey

Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

Corresponding Author: Linda Onnasch, Sekr. F 7, Marchstr. 12, D - 10587 Berlin, phone +49 30 314 – 25997, fax +49 30 314 – 25434, email: linda.onnasch@tu-berlin.de

Stefan Ruff, Sekr. MAR 3-1, Marchstr. 23, 10587 Berlin, Germany, phone +49 30 314 – 79518, fax +49 30 314 – 72 581, email: stefan.ruff@mms.tu-berlin.de

Dietrich Manzey, Sekr. F7, Marchstr. 12, 10587 Berlin, Germany, phone +49 30 314 – 21340, fax +49 30 314 – 25434, email: dietrich.manzey@tu-berlin.de

Abstract: Operators in complex environments are often supported by alarm systems that indicate when to shift attention to certain tasks. As alarms are not perfectly reliable, operators have to select appropriate strategies of attention allocation to compensate for unreliability and to maintain overall performance. This study explores how humans adapt to differing alarm reliabilities. Within a multi-task simulation consisting of a monitoring task and two other concurrent tasks, participants were assigned to one of five groups. In the manual control group none of the tasks was supported by an alarm system, whereas the four experimental groups were supported in the monitoring task by a miss-prone alarm system differing in reliability, i.e. 68.75%, 75%, 87.5%, 93.75%. Compared to the manual control group, all experimental groups benefited from the support by alarms, with best performance for the highest reliability condition. However, for the lowest reliability group the benefit was associated with an increased attentional effort, a more demanding attention allocation strategy, and a declined relative performance in a concurrent task. Results are discussed in the context of recent automation research.

Keywords: alarm systems, reliability, miss-prone automation, attention allocation, adaptive behaviour

1. Introduction

1.1 Alarm systems

Alarm systems represent a very basic form of automation, typically implemented to gather and analyse information on a certain task in order to inform a human operator about critical states or events, and to support the operator's attention allocation and decision-making. According to Parasuraman et al. (2000), this kind of automation represents the first two stages of their framework model, i.e. automation of information acquisition and information analysis. Information acquisition is automated when an alarm system monitors a single parameter and alerts the operator when critical thresholds are exceeded. If the alarm system is more complex, i.e. if it integrates different variables to detect a possible hazard, it involves both, automation of information acquisition and analysis (Pritchett, 2001). The common characteristic of these two types of information automation is that only cognitive functions related to the sensory perception and evaluation of environmental information are delegated to the automation whereas processes of decision-making and response selection as well as response execution are still left to the human (stages 3 and 4, Parasuraman et al., 2000).

Binary alarm systems are a stereotypical realisation of this widespread technology. The objective of these alarm systems is to support complex supervisory control tasks of operators. Typically, they are implemented in domains like aviation or the process industry where the monitoring of underlying system states and process information constitutes just one of several tasks that have to be performed by operators at the same time. The support provided by alarm systems is mainly enabled by the

attention-grabbing properties of alarms which relieve operators from continuous monitoring of a given process while still staying in the loop as alerts inform them when to shift attention to a critical system state (Pritchett, 2001).

Benefits of this type of automation can be described in terms of more efficient task management and prioritisation, as well as reduced operator workload. This in turn leads to a better performance in the task and improved performance in concurrent tasks as operators gain more spare capacities, which can be re-allocated (e.g. Bustamante et al., 2004; Meyer and Bitan, 2002).

However, the proposed benefit of this kind of automation can be off-set when alarm systems do not function properly. The reason for such alarm failures can be found in imperfect sensors and algorithms as well as in a noisy and uncertain world that cannot be interpreted distinctively by the alarm system. Generally, the performance of alarm systems can be described in the framework of signal detection theory (Green and Swets, 1966; Swets, 1964). Following this framework, there are two different errors that can occur and have to be differentiated dependent on the response criterion of the system. First, an alarm system can be miss-prone, i.e., the alarm system can fail to alert the operator by missing critical events. Second, an alarm system can be false-alarm prone. This is the case if it alerts an operator too often as not every alert corresponds to a critical event (Green and Swets, 1966; Swets, 1964). Given these possible failures, operators' responses to alarms always imply a decision under uncertainty. This decision reflects their assessment of how much they can rely on the alarm function.

1.2 Reliance vs. compliance

According to Meyer (2001, 2004), the explicit distinction between the two kinds of unreliability in human-alarm interaction is important because of their exclusive

behavioural consequences on the human part. False alarms may lead to delayed responses towards an alarm as operators know from experience that many of the alarms provided by the system do not correspond to actual malfunction (Getty et al., 1995). In extreme cases, i.e. in cases of high frequencies of false alarms, operators even refuse to respond to an alarm at all (Breznitz, 1984). Misses on the other hand affect operators' monitoring strategies in non-alarm periods. The more critical events are missed by the alarm system, the more operators must shift attention to the alarm-supported task and the raw data to compensate for this unreliability.

Meyer (2001, 2004) therefore characterises operators' behaviour as dependent on the alarm systems' state, i.e. if an alarm is present or not. In this context, compliance refers to operators' response to an alert that indicates a malfunction of the system and is mainly affected by the number of false-alarms emitted by a system. In contrast, reliance describes operators' tendency to rely on the alarm system when it indicates that the monitored process runs properly and the operators accordingly do not have to take evasive action. This latter behavioural tendency represents the major focus of the present paper and shall be addressed in some more detail in the following.

1.3 Operators' adaptation to imperfect alarm systems

According to Lee and See (2004), one of the most important perceivable characteristics for the calibration of reliance on automation (like alarm systems) is the system's reliability. With respect to miss-prone alarm systems, reliability can be described as the percentage of critical events that are correctly indicated by the alarm system. The higher the alarm system's reliability in this respect, the more operators can rely on the alarm and the less they are required to monitor the underlying data by themselves. In contrast, when reliability is low and the occurrence of misses cannot be excluded, operators have

to monitor relevant process data more frequently in order to compensate for the alarm system's imperfection and to keep overall monitoring performance high.

Calibration of reliance and compliance therefore can be considered as the result of an adaptive process which develops over time in interaction with an automated system, dependent on the user's experience with the automation's reliability (Lee and See, 2004; Parasuraman and Manzey, 2010).

How and to what extent operators adjust their own monitoring behaviour in case of the availability of (imperfect) alarm systems or other decision support has been addressed in several studies (e.g. Parasuraman et al., 1993; Wickens and Dixon, 2007). However, the results are mixed and provide a somewhat inconsistent pattern of effects. For example, Bailey and Scerbo (2007) examined operators' adaptation to a highly reliable support system. In three sessions, each lasting approximately 100 minutes, participants had to work on a manually controlled flight task while monitoring several simulated aircraft displays for failures. The monitoring tasks were supported by an alarm system that automatically indicated and resolved critical system states. Results indicated that participants' monitoring of the supported task decreased as a function of increasing system reliability, which was set to 87%, 98% and 99.7%, respectively. Participants who were supported by a highly reliable but still not perfect alarm system did detect fewer automation misses and showed increased response latencies to critical events when not alerted by the system, compared to participants who worked with an alarm system with lower reliability. Time-on-task had no effect on these results, i.e. even participants with more system experience and supported by a highly reliable alarm system could not appropriately adapt to automation's imperfection. These findings supported earlier results by Molloy and Parasuraman (1996) who also reported degraded monitoring performance in terms of less miss detection when participants interacted

with a highly reliable alarm system. However, they are in contrast to a number of other studies which suggest that operators indeed are very well capable of adapting their own monitoring behaviour to changing reliability levels, suggesting nearly optimum calibration of their reliance on automation reliability (e.g. Parasuraman et al., 1993; Sharma, 1999; Singh et al., 2005; Singh et al., 1997; Wiegmann et al., 2001).

In most of these studies however, the evaluation of monitoring performance was solely based on operator's performance (Bailey and Scerbo, 2007; Parasuraman et al., 1993; Wiegmann et al., 2001). This does not seem to be appropriate as the concept of an automated assistance or alarm system is to support the operator and to resume parts of the task; i.e. the task is performed jointly. As a consequence it is considered important to always respect the joint human-automation performance while evaluating overall performance benefits or costs associated with this sort of automated support.

In accordance with this approach, Wickens and Dixon (2007) conducted a meta-analysis consisting of 22 studies with varying reliabilities. In contrast to most interpretations of the aforementioned research, they found a positive linear relation between automation's reliability and the joint human-automation performance. That is, even though operators may have tended to miss more critical events when working with alarm systems of high reliability compared to systems with lower reliability, the overall number of *jointly* detected critical events was still higher with highly reliable systems than with lower ones. However, below an alarm system's reliability of 70%, accompanied by a 95% confidence interval, which brackets 65% and 75%, this compensation was associated with disproportional effort, and joint performance even got worse than working with no automation at all. Thus, compensation for unreliability seems to be possible to a certain level only.

This finding is supported by several other studies like, for example, a series of studies conducted by Dixon et al. (e. g. Dixon et al., 2004; Dixon et al., 2007; Dixon and Wickens, 2006).

In these studies, Dixon et al. (e. g. Dixon et al., 2004; Dixon et al., 2007; Dixon and Wickens, 2006) compared different levels of reliability of an alarm system supporting monitoring performance in a multi-task environment. They also found certain cost effects on concurrent task performance for alarm system reliabilities at least below 70%. When imperfect alarm reliability was realised by an increased number of misses, operators re-allocated their attention to the alarm-supported task to such extent that a high performance level in the alarm-supported task was maintained. However, concurrent task performance even dropped below the performance of a manual control group without automation support. This drop of performance was explained by a sort of overcompensation effect. The low reliability of the alarm system led to such a decrease in reliance on alarms that participants started to shift more attention than necessary to the alarm-supported task in order to compensate for the imperfection of their system.

Finally, the assumption that operator's adaptation to imperfect alarm systems might not be perfect - particularly for low reliability systems - is also supported by a study conducted by Wickens et al. (2005). In contrast to the aforementioned studies, Wickens et al. (2005) did not just evaluate possible costs of imperfect reliability on the performance level but also used eye-tracking data to directly evaluate the impact of different reliabilities on visual attention allocation. This additional evaluation level, i.e. eye-tracking data for attention allocation, complies with Moray's and Inagaki's (2000) assertion to evaluate operators' performance not only by fault detection but first and foremost by an analysis of their attention allocation strategies. Participants were required to work on a multi-task scenario based on demands of unmanned air vehicle

(UAV) control and several UAV-mission-related tasks that had to be performed concurrently. One of these latter tasks was supported by a binary auditory alarm system that was either perfectly reliable, 60% reliable in terms of misses (miss-prone) or 60% reliable in terms of false alarms (false-alarm prone). Additionally, these groups were compared to a baseline condition in which no automation support for any task was available. Most interesting to the current study was the result that working with the miss-prone automation removed visual attention from the concurrent tasks to the alarm-supported task. In the attempt to maintain adequate performance, participants drew even more attention to the alarm-supported task than in the baseline condition without automation support. Yet, even with this strategy, performance in the alarm-supported task dropped below the baseline condition level.

Summarizing the scope of this research it can be assumed that human operators adapt their behaviour to the characteristics of the automation they are working with. However, there is evidence that this adaptation might not always be appropriate. Studies focussing on human monitoring performance alone suggest that particularly highly reliable alarm systems might lead to miscalibrations of behaviour in terms of an inappropriate withdrawal of attention from the alarm-supported tasks, and an elevated risk of missing critical events. Studies focussing on joint human-system performance specifically point to issues related to low reliable systems (i.e. reliability < .70) which might reduce reliance levels to an extent that it becomes even more detrimental for concurrent task performance than working without any automation support.

However, there are two common drawbacks of most of the studies conducted thus far. The first one concerns the relatively extreme levels of automation reliability that were usually compared in those studies, and thus failed to describe the characteristics of adaptation across a whole range of reliability levels. Second, most

studies that explicitly varied reliability only concentrated on the state manifestation of reliability effects on human performance, hence excluding the adaptation process itself (some exceptions are Parasuraman et al.,1993 or Bailey and Scerbo, 2007). Although, researchers in the early 90s already argued that system experience has substantial impact on how operators interact with and monitor automation (e.g. Lee and Moray, 1992; Muir, 1987, 1994), only few studies have picked up this claim and focused on reliance development since then. What is known to date is that the adaptation to automation's characteristics seems to proceed fast, and that already single automation failures can have a detrimental impact on users' trust and behaviour (e.g. Bahner, Hüper, Manzey, 2008; Lee and See, 2004; Parasuraman and Manzey, 2010; Manzey, Reichenbach, Onnasch, 2012). Beyond that, only little is known about how these effects develop dependent on different reliability levels, to what extent they are reflected in changes of monitoring strategies, and what the performance consequences are in multi-task environments.

Based on these findings, the goal of the current study was to gain further insight into possible adaptation strategies to alarm systems with respect to different levels of alarm reliability. In contrast to numerous other studies that have concentrated on false alarm-prone automation (e.g. Bliss and Dunn, 2000; Bliss et al., 1995; Lees and Lee, 2007; Wickens et al., 2009), the focus of our study was on miss-prone alarm systems. Even though this kind of error seems to occur less often because designers tend to set sensor thresholds at a very low level (engineering fail safe approach; Swets, 1992), the consequences of missing critical events in safety-related domains are usually more severe than consequences of false alarms. For this reason, it was of special interest if and how operators would compensate for this kind of diagnostic failure.

The task used for the experiment was a multi-task simulation, including three different subtasks. One of these tasks involved a system monitoring task where participants had to monitor different engine gauges for possible failures with or without support of a binary visual alarm system of different reliability. To evaluate participants' monitoring effectiveness, we considered the joint human-automation performance as well as participants' performance in concurrent tasks. In addition, eye-tracking analyses were performed in order to directly assess the impact of alarm system's reliability on participants' attention allocation. By separate analyses of eye-tracking data for periods where alarms were emitted vs. non-alarm periods it was further possible to distinguish between effects of alarm reliability on the level of participant's reliance and compliance.

For the impact of alarm reliability on **performance** we hypothesised:

(1) There is an automation benefit in the alarm-supported task in terms of a superior joint performance of human and alarm system compared to no automation support at all.

(2) Automation benefits in terms of a superior joint performance of human and alarm system compared to no automation support at all are positively related to the alarm system's reliability (Wickens and Dixon, 2007).

(3) Concurrent task performance benefits from highly reliable automation support compared to the manual control condition. However, these benefits decrease with decreasing alarm reliability over time because participants start to reallocate attention to the alarm-supported task to compensate for automation's imperfection.

In extreme cases, i.e. interacting with an automation with a reliability below the critical cut-off of 70%, this adaptation of attentional reallocation should even lead to

cost effects in terms of a degraded performance compared to working with no automation support at all (Dixon et al., 2007; Dixon and Wickens; 2006; Rovira et al., 2007; Wickens and Dixon, 2007; Wickens et al., 2005).

For participants' **visual attention allocation**, operationalised by eye-tracking measures, we expected:

(4) Participants supported by an alarm system of sufficient reliability invest less attentional resources in system monitoring compared to working with no automation support.

(5) Participants adapt their own monitoring of engine gauges to the alarm systems' reliability over time.

Participants working together with relatively reliable automation support should decrease their own monitoring with growing system experience whereas participants supported by an unreliable automation should increase monitoring of the underlying data (engine gauges).

(6) In interaction with alarm reliability below 70% participants' attention allocation is not distinguishable from attention allocation when working manually on this task as compensation for unreliability becomes inefficient (Wickens et al., 2005).

(7) Because we operationalised reliability only by misses of the alarm system, differences in participants' attention allocation primarily emerge during non-alarm periods, reflecting effects on participants' reliance.

No or only little differences were expected for visual attention effects in direct response to alarms, which would reflect the level of compliance and which was expected to be high for systems that did not commit false alarms.

2. Method

2.1 Participants

The number of participants was defined based on a power analysis (GPower 3.1, for details see e.g. Buchner et al., 1997). A total of 65 students from the faculty of mechanical engineering and transport systems (18 female, 47 male) ranging in age from 19 to 32 (M = 23.6, SD = 2.3) participated in partial fulfilment of course requirements. None of the participants had prior experience with the flight simulation task used in the study. Participation was voluntary (other alternatives for fulfilment of course requirements were available) and could be cancelled anytime.

2.2 Task and apparatus

As experimental task the most recent version of the *Multi-Attribute Task Battery* (MATB; Miller, 2010) was used. It was directly based on the original version developed by Comstock and Arnegard (1992) which was used in previous research (e.g. Parasuraman et al., 1993). All main functionalities including the interface corresponded to the original version. Only the programming environment has been changed (MatLab instead of QBasic) which made it easier to implement experimental modifications.

The MATB is a multi-task flight simulation consisting of a two dimensional compensatory tracking, engine-system monitoring, fuel resource management, communications, and scheduling. In the present study, only the compensatory tracking, the resource management, and the system monitoring were implemented and had to be performed concurrently. The user interface of the MATB used in the present study is shown in Figure 1.

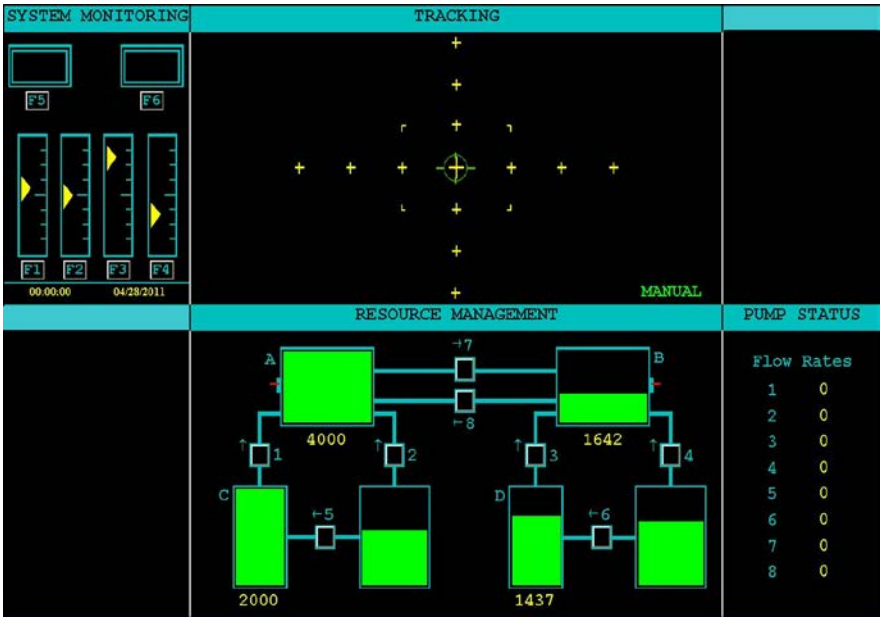

Figure 1. MATB as used in the current study with the compensatory tracking in the upper middle position, the resource management beneath and the system monitoring in the upper left display corner.

In the two-dimensional compensatory tracking task participants are required to keep a randomly moving cursor in the centre target position by applying appropriate control inputs via joystick. In the resource management task participants must compensate for fuel depletion by pumping fuel from four supply tanks into two main tanks.

The system monitoring task was most important for the current research. It consists of four vertical engine gauges with moving pointers that participants must monitor for abnormal values that occur randomly. As long as all engines function properly, the pointers fluctuate by chance within a fixed range around the centre value of the gauges. However, in case of a malfunction the pointer of the gauge for the affected engine suddenly shifts upwards or downwards by two gauge units and starts to

fluctuate around this new position. These deviations must be detected by participants and reset by a corresponding key press. If a malfunction is not detected within 10 seconds the gauge resets automatically and the event is logged as an event missed by the participant.

Dependent on the task configuration, this system monitoring task has to be performed manually or with support of a binary master alarm system. In the latter case, a visual red alert appears above the gauges whenever the alarm system detects a parameter deviating from its nominal value. Nevertheless, the identification of the affected gauge and the corresponding reset of the parameter still have to be performed manually by participant. According to the stages and levels framework of automation proposed by Parasuraman et al. (2000), this type of alarm system can be classified as a stage 1 automation (information acquisition).

The MATB was presented in front of the participant on a 20 inch monitor that was equipped with a remote eye-tracking system (RED system, SensoryMotoric Instruments, Germany). This latter system enabled to sample gaze movements during task performance with a sampling rate of 120 Hz. Based on these data, gaze fixations in different areas of interest (AOI, see definition below) were automatically recorded.

2.3 Design

The study used a two factorial design. The first factor (Group) was defined as a between-subject factor and consisted of four experimental groups and one manual control group. The four experimental groups differed with respect to the reliability of the alarm system participants worked with in the monitoring task. The alarm reliabilities were set to 68.75%, 75%, 87.5%, and 93.75% by varying the number of critical signals that were missed by the alarm system. The two lowest reliability levels (68.75% and

75%) were chosen in reference to the result of the meta-analysis of Wickens and Dixon (2007) which suggests that a reliability level around .70 represents an important cut-off value which needs to be exceeded before automation support might become beneficial for joint human-system performance compared to conditions without automation support. The two highest reliability levels were realised to compare the results to findings from previous studies and to include reliability levels quite close to realistic scenarios (Bagheri and Jamieson, 2004; Parasuraman et al., 1993). In the manual control group there was no automation support at all, i.e. participants had to detect all malfunctions reflected by parameter deviations in one of the four gauges without the support of an alarm system.

The second factor (Block) was defined as a within-subject factor and was included to gain further insight on how participants' adapt their attention and performance over time in response to the alarm system's reliability they were working with. Every participant had to perform the three concurrent tasks of the MATB for three 10-minute blocks. A total of 16 critical events occurred in the monitoring task during each block which had to be detected by the alarm system or the participant, respectively. The resulting 5 (Group) x 3 (Block) design is shown in Figure 2.

		Block		
		block #1	block #2	block #3
Group	manual		n = 13	
	68.75%		n = 13	
	75.00%		n = 13	
	87.50%		n = 13	
	93.75%		n = 13	

Figure 2. 5 (Group) x 3 (Block) study design.

A somewhat more complex design was used for supplementary analyses of effects of reliability on visual attention allocation in phases where alarms were present vs. phases where alarms were not present. The beginning of alarm phases could be identified by the visual red alert that appeared to inform participants about an abnormal system state. The end of these phases was defined by participants' appropriate reaction to the alarm or, if participants did not react, the maximum time the failure was present, i.e. 10 seconds. These supplementary analyses involved the four alarm-supported groups as between-subjects factor, the block factor (within-subject) and a third factor representing alarm vs. non-alarm periods (within-subject). The resulting 4 (Group) x 3 (Block) x 2 (Alarm State) design allowed a test of the hypothesis that differences in reliability of the alarm system would affect attention reallocation during non-alarm periods only, reflecting effects on reliance on the automation but not compliance (Hypothesis 7).

2.4 Dependent measures

To investigate the impact of the experimental factors on the perceived alarm reliability (manipulation check) as well as on performance and visual attention allocation, three different categories of dependent measures were sampled and analysed.

A visual-analogue scale assessed the perceived reliability. Participants provided ratings to the question "How reliable was the system you worked with on a scale ranging from 0% to 100%.

Performance measures were defined for all three tasks of the MATB participants had to perform concurrently and collected for each 10 minute block separately. For the system monitoring task, *percentage of detected system failures* was defined as the

percentage of all engine failures detected correctly by the human operator (control condition) or the human and alarm system together (joint performance in the alarm conditions).

For the tracking task as well as the resource management task the *root mean squared errors* (RMSE; Parasuraman et al., 1993; Prinzel et al., 2001; Singh et al., 1997) were calculated. The RMSE for the tracking task was calculated as a measure of mean deviation from the central target position, based on deviation data sampled at an interval of 5 seconds. The RMSE for the resource management task was calculated in relation to an optimal tank level, which had to be maintained in both main tanks. Fuel levels were sampled and RMS errors computed for each 5-second period.

Visual attention allocation was measured by means of eye-tracking. Specifically, the *relative fixation time* for different pre-defined areas of interest (AOI) was assessed. For this purpose, three different AOIs (specified by pixel areas) were defined before the experiment started. These AOIs corresponded to the three different tasks participants had to perform: compensatory tracking, resource management, and system monitoring (see Figure 1). Fixations were defined by a minimum duration of 80 ms and a maximum dispersion in this time of 100 pixel. Relative fixation time was defined as the time participants fixated an AOI relative to the overall fixation time, i.e. sum of times any one of the AOIs was fixated.

2.5 Procedure

Following a demographic questionnaire, an instruction on the MATB, and an initial calibration of the eye-tracking system, participants were familiarised with performing the three different tasks in a 10 minute practice block. They were instructed that all three tasks would be of equal importance, and that they should work on all tasks

concurrently with equal priority. Afterwards, they were randomly assigned to one of the five groups. Participants in the four experimental groups were introduced to the function of the alarm system. Specifically, they were told that the alarm system would not be perfectly reliable and that therefore, they may not fully rely on it. However, no concrete reliability information was provided. Then, the experiment started consisting of three 10 minute blocks. Prior to each block the eye tracker was re-calibrated. The perceived reliability of the alarm system was assessed in the experimental groups after the second block. The experiment ended with the debriefing of participants.

3. Results

In the following, the results are presented separately for subjective measures, performance, and eye-tracking data. The description of results focusses on effects of reliability (factor Group) and / or possible interactions with time-on-task (factor Block) indicating adaptive processes.

3.1 Perceived reliability

A univariate between-subjects ANOVA contrasting the four experimental conditions with automation support of different reliability revealed that mean ratings of perceived reliability differed between these experimental groups in a meaningful manner ($M_{68.75\%}$ = 66.77%, $M_{75\%}$ = 72.38%, $M_{87.5\%}$ = 80.08%, $M_{93.75\%}$ = 87.08%), $F(3, 51) = 6.11$, $p < .002$.

Further t-tests were performed in order to analyse whether perceived ratings differed from the actual reliability. Because no differences were expected, α was adapted to a 20% level for these analyses (null-hypothesis testing). Results showed that participants in the two highest reliability conditions systematically underestimated the

actual reliability (87.5%: $t(12) = -3.29$, $p < .007$; 93.75%: $t(12) = -3.09$, $p < .01$). No differences between actual and perceived reliability were found for the 68.75% and 75% reliability condition (68.75%: $t(12) = -.48$, $p = .63$; 75%: $t(12) = -.52$, $p = .61$). This finding is in line with previous research (Wiczorek and Manzey, 2010; Wiegmann et al., 2001; Wiegmann and Cristina, 2000) indicating a systematic bias of under- and overestimation, respectively, for extreme levels of reliability. Nevertheless, the overall pattern of results confirms that our manipulation had worked successfully as the perceived reliabilities were systematically related to the actual ones and significantly differed between the experimental conditions.

3.2 Performance measures

3.2.1 Monitoring task

Performance measures were analysed in two steps according to the different hypotheses. The first step addressed the testing of our hypothesis which postulated an alarm-support benefit in the monitoring task compared to no alarm-support at all (Hypothesis 1).

For this purpose, the *percentage of detected system failures* was analysed with a 5 (Group) x 3 (Block) ANOVA. The corresponding data, i.e. detection rates for all experimental groups and the manual control group across blocks, are shown in Figure 3. As expected, there was a clear alarm-support advantage reflected in a higher percentage of detected system failures by human and automation together in all alarm-supported groups, compared to the manual control group ($F(4, 60) = 10.36$, $p < .001$, $\eta^2 = .40$). Averaged across blocks, participants of the control group only detected 73.23% of all failures. In contrast, participants in the experimental group with the least reliable alarm system already detected 90.70% of all failures, and this number increased systematically with increasing reliability of alarms ($M_{75\%} = 92.46$, $M_{87.5\%} = 93.26$, $M_{93.75\%} = 95.83$).

This difference between automation-supported groups and the manual control group was statistically supported by post hoc analyses using Scheffe' tests. Analyses revealed that the manual control group detected significantly less system malfunctions compared to any of the alarm-supported groups ($p_{manual\text{-}68.75\%} < .003$; $p_{manual\text{-}75\%} < .001$; $p_{manual\text{-}87.5\%} < .001$; $p_{manual\text{-}93.75\%} < .001$). No differences occurred between the alarm-supported groups (all $p > .05$). Additionally, an interaction of reliability with participants' time-on-task was found, Group x Block interaction effect, $F(8, 120) = 2.37$, $p < .03$, $\eta^2 = .13$. Whereas all conditions showed an improved performance across blocks, the extent of this performance increase was different for the five groups. The largest increase in detected system failures over time was observed for the manual control group. In this condition, no alarm system support was available. Still, participants had to adapt to the underlying system characteristics and get familiar with the error rate in the monitoring task to perform adequately. As becomes evident from Figure 3, this form of adaptation was comparable to a similar, albeit weaker trend of participants' behaviour in the group working with the least reliable alarm system. Compared to the other conditions with alarm support, this group showed the worst performance at the beginning, but participants adapted their behaviour to the characteristics of the alarm system over time and were able to compensate effectively for its unreliability. However, this latter difference between the alarm-supported groups did not become significant in an additional 4(Group) x 3(Block) ANOVA, comparing the alarm-supported groups only. For this analysis neither the expected effect of Group ($F = 1.69$), nor a Group x Block interaction effect ($F = 1.16$) emerged (contradicting Hypothesis 2).

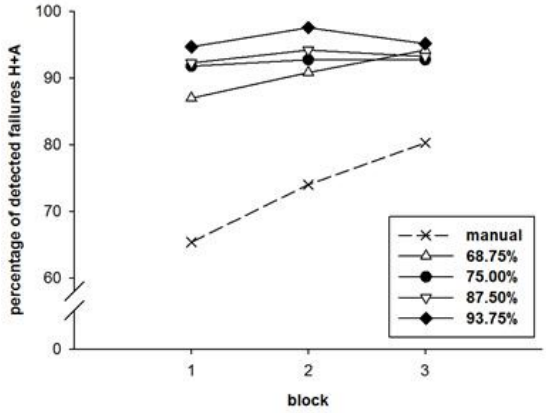

Figure 3. Effect of alarm reliability on detected system failures - human + alarm system.

3.2.2 Concurrent tasks

Following the same statistical approach as for the monitoring task, performance in the concurrent tasks was analysed in two steps. We expected that compared to higher reliability levels, working with the least reliable alarm system would negatively affect concurrent task performance because participants would rely to a lesser extent on the proper functioning of the alarm support (Hypothesis 3). More specifically, it was expected that concurrent task performance of the 68.75% reliability group would not be better than performance in the manual control group, i.e. a condition with no automation support at all.

For concurrent *tracking task* performance the 5 (Group) x 3 (Block) ANOVA revealed a significant Group x Block interaction, $F(8, 120) = 3.59$, $p < .002$, $\eta^2 = .19$. Essentially the same pattern of effects was also observed when comparing the alarm-supported groups only by a 4 (Group) x 3 (Block) ANOVA, with a significant interaction effect of Group x Block, $F(6, 96) = 4.98$, $p < .001$, $\eta^2 = .23$.

As can be seen in Figure 4, contrary to our expectations, participants in the 68.75% reliability group started at a very high performance level reflected in a smaller mean tracking error than in all other groups ($M_{manual} = 131.78$, $M_{68.75\%} = 117.58$, $M_{75\%} =$

136.05, $M_{87.5\%}$ = 144.57, $M_{93.75\%}$ = 137.76). However, whereas participants of the other groups showed a considerable performance improvement over time, mean performance of participants in the 68.75% reliability condition declined across the three blocks. This eventually led to comparable performance levels for all groups in block #3 (M_{manual} = 124.62, $M_{68.75\%}$ = 126.94, $M_{75\%}$ = 126.78, $M_{87.5\%}$ = 127.58, $M_{93.75\%}$ = 129.55). This finding provides some indirect support for our hypothesis. In contrast to all other alarm-supported groups, participants working with the lowest reliable alarm system were only able to protect their performance in the monitoring task across time at the expense of compensatory decrements in concurrent task performance.

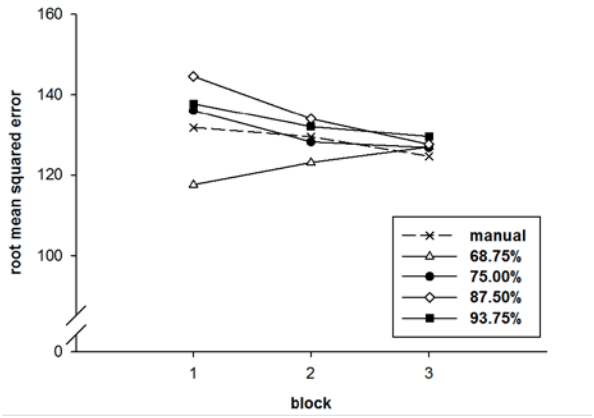

Figure 4. Effect of alarm reliability on performance in the concurrent tracking task (higher values represent greater deviations).

For the *resource management task* neither a main effect of Group nor a Group x Block interaction emerged (all F<1.0). Only a Block effect became significant independent of whether all groups were considered in a 5(Group) x 3(Block) ANOVA, $F(1.2, 75.69) = 5.02$, $p < .03$, $\eta^2 = .07$, or the analysis was only conducted for the four experimental groups with alarm support, $F(1.5, 73.47) = 4.31$, $p < .03$, $\eta^2 = .08$. With

increasing time-on-task all groups achieved better results reflected in a decreased mean RMSE.

3.3 Visual attention allocation

3.3.1 Overall monitoring effects for the different AOIs

Figure 5 illustrates the results for the mean *relative fixation times* on the three different AOIs, i.e. monitoring task (left panel), tracking task (middle panel) and resource management task (right panel).

For the *monitoring task*, participants in the two highest groups (93.75% & 87.5%) showed relatively short but stable mean *fixation times* across blocks. This effect was expected because these participants could rely to a high degree on the alarm system. Stable mean fixation times across blocks also were found for the 75% reliability group, albeit on a somewhat higher level. In clear contrast to these three groups, a considerable increase of mean fixation time through blocks was found for both, the manual control group as well as the group working with the lowest reliable alarm system (Figure 5, left panel). Analysed by a 5 (Group) x 3 (Block) ANOVA these findings were statistically supported by a significant Group x Block interaction, $F(7.14, 107.13) = 2.46$, $p < .03$, $\eta^2 = .14$.

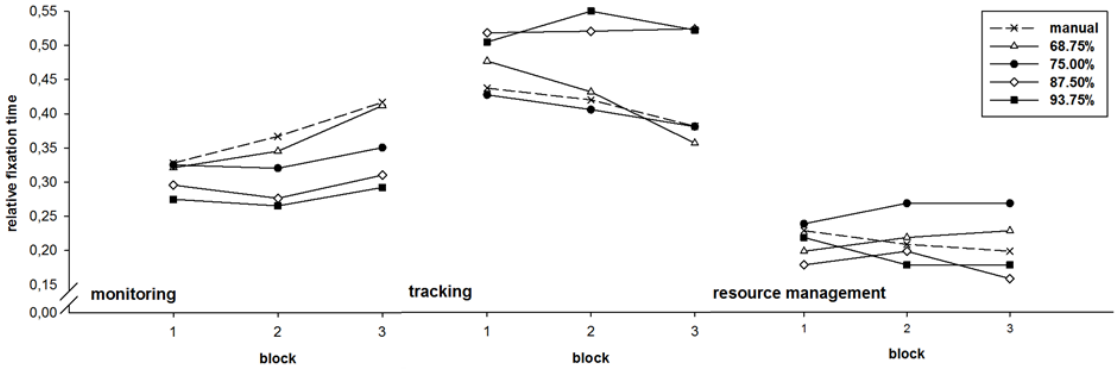

Figure 5. Effect of alarm reliability on the relative fixation time; AOI from left to right: monitoring, tracking and resource management.

Results for the monitoring task were mirrored in the *relative fixation times* for the *tracking task* (Figure 5, middle panel). Directly inverse to the findings for the monitoring task, the 93.75% and the 87.5% reliability groups had the longest fixation times on tracking which only marginally changed over time. For the other groups, a considerable decrease of fixation times across blocks was found which was most substantial for the 68.75% reliability condition and indicated a successive re-allocation of visual attention away from the tracking task over time. The 5(Group) x 3 (Block) ANOVA revealed a significant main effect of Group ($F(4, 60) = 2.64$, $p < .05$, $\eta^2 = .15$), moderated by a Group x Block interaction effect, $F(6.75, 101.29) = 3.62$, $p < .003$, $\eta^2 = .19$.

Finally, mean relative times of fixation for the *resource management task* did not show a clear pattern of effects. The 5 (Group) x 3 (Block) ANOVA did not reveal a main effect of Group ($F= 1.55$), however, the Group x Block interaction became significant, $F(7.16, 107.39) = 2.14$, $p < .05$, $\eta^2 = .12$. As becomes evident from Figure 5 (right panel), relative fixation times showed a slight increase across blocks for the two

conditions with the lowest reliable alarm systems, and a reverse trend for the other three groups.

In summary, the pattern of effects for relative fixation times on the three different tasks is in accordance with our hypothesis that alarm reliabilities affected the allocation of visual attention. Specifically, the results point to a successive re-allocation of attention over time, away from the tracking task to the monitoring task. Re-allocation emerged in a very similar way in both, the control condition without automation support and the condition with support of the lowest reliable alarm system.

3.3.2 Specific effects for alarm and non-alarm periods

As our alarm systems were miss-prone it was expected that they would primarily affect the reliance of participants in the alarm systems' function but not their compliance. Accordingly, it was expected that possible effects of alarm reliability on visual attention allocation would only emerge during periods when no alarm was present (non-alarm periods). During these non-alarm periods participants should allocate more attention to the monitoring task, the less they relied on the proper functioning of the alarm system. I.e., if participants expected that the alarm system could miss critical system states they should reallocate their attention from the other two concurrent tasks to the alarm-supported monitoring task. In contrast, no differences were expected for visual attention allocation in direct response to alarms which never represented false alarms. For the analysis of this presumed effect only the alarm-supported groups were considered, as a differentiation of these periods was not possible for the manual control group who worked without alarm system.

Figure 6 shows mean relative fixation times for all groups across blocks, separated for the three tasks (from left to right), and periods with and without alarm (upper vs. lower panel).

Results for the *monitoring task* revealed that the pattern of effect found in the overall analysis reported above, i.e. an increase in *relative fixation time* across blocks only in the control group and the group working with the lowest reliable alarm system, was exclusively related to non-alarm periods (Figure 6, upper left panel). In contrast, a decrease of mean fixation times across blocks emerged in all groups during alarm periods (Figure 6, lower left panel). In the analysis of these data by a 4(Group) x 3(Block) x 2(Alarm State) ANOVA this was reflected in a significant main effect of Alarm State, $F(1, 48) = 52.44$, $p < .001$, $\eta^2 = .52$, which was moderated by a significant Alarm State x Block interaction, $F(1.74, 83.81) = 27.86$, $p < .001$, $\eta^2 = .36$. Furthermore, the significant main effect of Alarm State indicated that mean relative fixation times for the monitoring task were higher during alarm vs. non-alarm periods, i.e. higher when an alarm prompted the participants to visually analyse which of the four gauges indicated a failure.

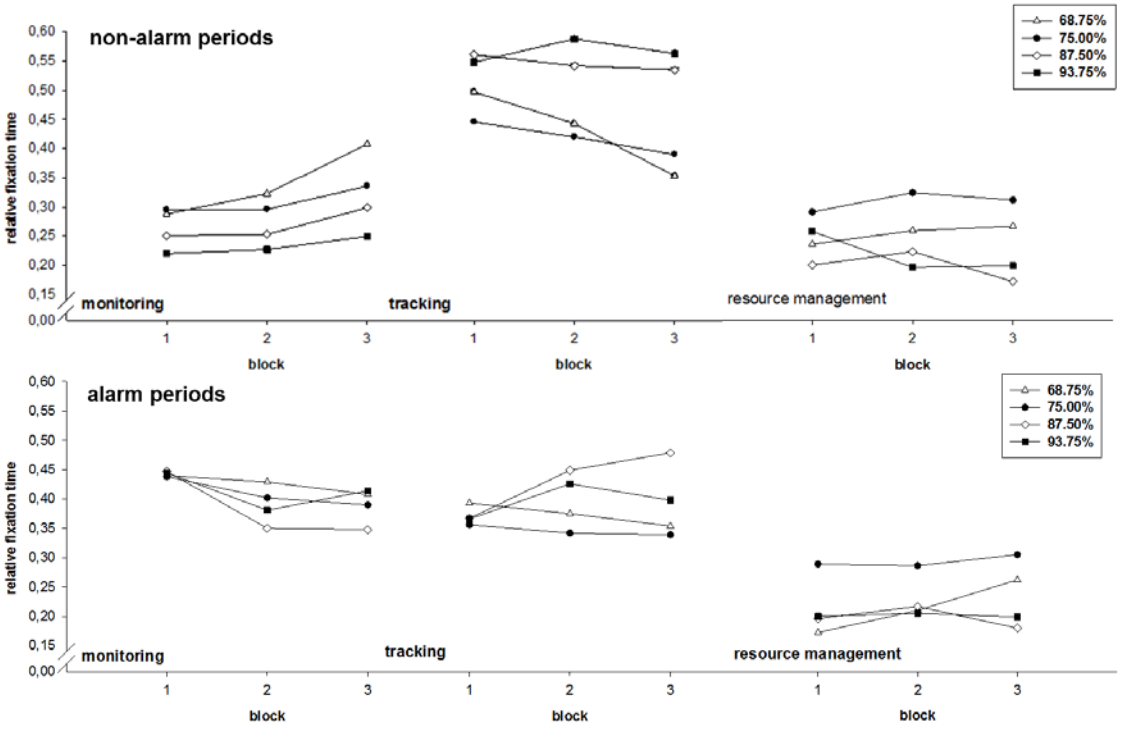

Figure 6. Effect of alarm reliability and alarm state (upper panels non-alarm periods, lower panels alarm periods) on the relative fixation time; AOI from left to right: monitoring, tracking and resource management.

The results for the *tracking task*, separated for alarm and non-alarm periods, are illustrated in the middle panel of Figure 6. Again the effects for non-alarm periods equalled the effects reported above for the overall analysis. In these periods the participants of the two groups with highest alarm reliability had the longest mean fixation times and showed a constant monitoring pattern. In contrast, a continuous decrease of relative fixation times was found for participants in the two lower reliability conditions (see upper panel). This separation of groups became also evident in alarm periods, although in a slightly different pattern (see lower panel). All groups spent comparable time looking at the tracking task in block #1. However, with on-going adaptation to the reliability of the alarm system, the 93.75% and the 87.5% groups spent

more time on this task even when a failure in the monitoring task was present. The two groups working with the less reliable alarm systems slightly decreased their monitoring time on tracking in alarm periods. The statistical equivalents to these findings are presented in Table 1.

Table 1. Results of the three-factorial ANOVA for the AOI Tracking Task.

	F-statistic	**p-value**	**Partial Eta-Squared**
Block x Reliability	$F(4.76, 76.30) = 4.21$	$p < .003$	$\eta^2 = .20$
Alarm-State	$F(1, 48) = 62.30$	$p < .001$	$\eta^2 = .56$
Alarm-State x Reliability	$F(3, 48) = 3.66$	$p < .02$	$\eta^2 = .18$
Block x Alarm-State	$F(1.76, 84.78) = 15.10$	$p < .001$	$\eta^2 = .23$
Block x Alarm-State x Reliability	$F(5.29, 84.78) = 2.58$	$p < .03$	$\eta^2 = .13$

Results for the *resource management* task again revealed that the pattern of effects found in the overall analysis, i.e. an increase in fixation time across blocks for the two lowest reliability conditions, and a reverse effect for the other two conditions, was exclusively related to non-alarm periods, $F(4.70, 75.28) = 2.77$, $p < .03$, $\eta^2 = .14$ (Figure 6, upper right panel). Moreover, when an alarm was present, the resource management was less fixated than in non-alarm periods, $F(1, 48) = 4.39$, $p < .05$, $\eta^2 = .08$; $M_{alarm} = 0.205$, $M_{no\ alarm} = 0.222$. This separation was enforced with ongoing time-on-task, $F(2, 96) = 6.48$, $p < .003$, $\eta^2 = .11$ (Figure 6, right panel).

In summary, results from the monitoring AOI supported the assumption that the attention re-allocation, related to different reliability levels, was only observable in non-alarm periods (hypothesis 7). However, for the tracking task the alarm system's reliability not only affected attention allocation in non-alarm periods but also in alarm

periods. Eye-tracking data from the AOI resource management revealed a difference between participants' attention allocation in non-alarm and alarm periods but no interaction of Alarm State and Group. Therefore, results did not fully support Hypothesis 7.

4. Discussion

The main objective of this study was to investigate to what extent human operators adapted their visual attention allocation and multi-task performance to different reliability levels of miss-prone alarm systems.

First hypotheses (1-3) were stated with regard to effects of different alarm reliabilities on performance in both, the alarm-supported task as well as other concurrent tasks. Based on previous studies we specifically assumed that participants' performance as well as attentional demands would benefit from an automation support that is fairly reliable, i.e. at least 70% (e.g. Dixon et al., 2007; Rovira et al., 2007; Wickens et al., 2005). Below this reliability, automation support is not expected to be helpful as there is some evidence that a reliability of approximately 70% (accompanied by a 95% confidence interval) represents a critical boundary below which manual compensation strategies would not be effective any more (e.g. Wickens and Dixon, 2007). In this case, we supposed at least performance in concurrent tasks to suffer because participants would start to re-allocate attention to the automation-supported task and monitor the underlying data by themselves to compensate for unreliability. The results of the present study support most of these assumptions.

Considering the results for the performance data first, we found a clear automation benefit in the alarm-supported task in terms of joint human-automation performance compared to working with no automation support at all supporting

Hypothesis 1. This was true for all groups that worked with alarm system support. Whereas the manual control group only detected around 70% of engine malfunctions, detection rates increased with alarm-support even in the lowest reliability condition up to 90%. This effect seemed to be an overall automation benefit as differences between the alarm-supported groups did not become significant (contradicting Hypothesis 2). Therefore, the automation benefit was only attributable to the difference between alarm-supported groups on the one hand and the manual control group on the other. This result revealed that all participants in the alarm-supported groups adapted to differing reliability levels in a very effective way. This (non-)finding indicated that participants' adaptation was even more successful than we would have assumed based on findings by Wickens and Dixon (2007) which showed that higher reliability levels still led to significantly improved performance compared to lower reliability levels in the automation-supported task. One reason for these deviant findings might be due to the operational definition of reliability in our study. Whereas Wickens and Dixon (2007) included studies to the meta-analysis that operationalised (un-)reliability by misses and/or false alarms, we defined reliability by misses only. According to Meyer (2001, 2004) false alarms mainly impact participants' compliance with the automation. As a consequence, false alarms lead to a degraded performance in the automation-supported task as participants start to ignore alarms (Meyer, 2001, 2004). This could explain why Wickens and Dixon (2007) found performance decrements in the automation-supported task when reliability was low. Misses, on the other hand, affect participants' reliance on the alarm system. Because of the frequently missed critical states participants start to monitor the underlying data to compensate for the alarm system's unreliability. This adaptation should not and in fact did not affect performance in the alarm-supported task. However, following Meyer (2001, 2004),

concurrent task performance should be negatively affected by this change of attentional focus to the aided task. Therefore, participants' adaptation of monitoring strategies to compensate for the unreliability of the alarm systems was expected to lead to differences in concurrent task performance between the different groups.

This was addressed by our third hypothesis. According to previous studies (e.g. Dixon et al., 2007; Dixon and Wickens; 2006; Rovira et al., 2007; Wickens and Dixon, 2007; Wickens et al., 2005) we expected that, compared to higher reliability levels, lower alarm reliabilities should result in significant performance decrements in the concurrent tasks because participants' reliance on the alarms would decline and induce a re-allocation of attention. Following Wickens and Dixon (2007), we especially assumed that working with an alarm system with a reliability of less than 70% might be even more detrimental to performance than working with no automation support at all. This assumption was at least indirectly supported by our findings.

Although concurrent tracking task performance in the 68.75% group was better than the one of all other groups in the first block, participants in this group were the only ones who could not protect their performance over time but showed a considerable decline across blocks. As we only told participants that the alarm system would not be perfectly reliable, but gave no precise information, the first block was especially important to participants to gain experience with the system and to start to adapt their behaviour to the alarm system's (un)proper functioning. It seems that participants working with the least reliable alarm system initially spent more time on the concurrent tracking task than on the alarm-supported monitoring task, which resulted in superior results compared to the other conditions. However, with increasing experience they started to realise the limitation of their alarm system and changed their behaviour accordingly by re-allocating their attention away from the tracking task. The other aided

groups also started to adapt to the alarm system's characteristics. Because these alarm systems were more reliable, adaptation proceeded the other way, i.e. in favour of the concurrent tracking task, as these groups recognised that they could rely more on their automation support. These diverging adaptation characteristics of the lowest alarm-supported group and conditions with a more reliable alarm-support eventually led to comparable performance levels in the last block.

However, our far-reaching hypothesis that working with the least reliable alarm system would impair concurrent task performance even more than working without automation support was not supported by the data. This might be related to the fact that, contrary to our expectations, the provision of alarm support did not lead to obvious benefits in concurrent task performance in any of the alarm-supported groups. That is, even in the groups with highly reliable alarm systems, the participants were not able to make use of this support in terms of improved concurrent task performance.

One reason for this finding could be the overall high task load involved in performing the MATB. In contrast to, for example, Dixon et al. (2007) or Rovira et al. (2007) who have reported automation benefits for concurrent task performance, participants had to work on three instead of two concurrent tasks. Additionally, the MATB compensatory tracking has high visual attentional demands as it needs continuous control inputs since even short interruptions of control lead to great deviations from the centre target position. Given this, it might not be too surprising that even a reliable alarm support for the monitoring task has not led to better concurrent task performance in our study because participants already performed at their maximum; the tasks were not sensitive to changes in attention allocation. Yet, this is a post-hoc explanation and cannot be fully proved by the present data.

The most direct insights in the nature of adaptation to alarm systems of different reliability are provided by the effect of alarm reliabilities on participants' attention allocation strategies reflected in the eye-tracking data. These data were collected in order to directly capture possible effects of the experimental conditions on allocation of visual attention which might help understanding effects on performance. Indeed the analyses of eye-tracking data suggest that the effects of alarm support first and foremost become evident in their effects on attention allocation (supporting Hypothesis 4).

As expected, participants in the groups with the highest reliable alarms allocated least attention to the monitoring task, followed by the two groups with the less reliable alarm systems and the manual control group (supporting Hypothesis 5).

A comparison of the eye-tracking pattern between the alarm-supported groups and the manual control group further revealed, that the participants of the 68.75% group allocated as much visual attention to the supported task as the manual control group, i.e. behaved as if no automation support were available (supporting Hypothesis 6). It reveals that participants working with the least reliable alarm system were able to compensate for the imperfection of their alarm system on a performance level but only at the expense of a highly demanding attention allocation strategy and a reallocation of attention away from the concurrent tasks which eventually led to the relative performance decline for the tracking. These results are in line with previous findings by Wickens et al. (2005) who also showed that miss-prone automation led to a reallocation of visual attention away from other tasks to the raw data in order to compensate for unreliability. Furthermore, our findings provide some more support for the assumption of a critical reliability cut-off around 70% below which automation support cannot be considered as helpful anymore (Wickens and Dixon, 2007). Albeit we could not entirely confirm a detrimental effect of reliability below 70% on the performance level, the costs

for compensation became directly evident when considering the distribution of visual attention. Although the least reliable alarm system still detected 68.75% of all system malfunctions it obviously was not considered to be of much help and did not reduce participants' attentional demands of this task compared to performing it with no automation support at all.

Our last hypothesis (Hypothesis 7) was based on Meyer's assumption (2001, 2004), that misses of an alarm system mainly affect participants' attention allocation in non-alarm periods and have no effect on their visual attention in alarm periods. Regarding the non-alarm periods, this assumption was completely confirmed. The effects found in the overall analysis exactly mirrored participants' attention allocation in non-alarm periods, i.e. the overall effects were mainly due to these periods. This was true for attention allocation on all three concurrent tasks. We could confirm that working with the least reliable aid in terms of misses led to a reallocation of attention away from the tracking and resource management to the alarm-supported monitoring task in the attempt to compensate for unreliability and to maintain performance on this task. In contrast, groups working with more reliable alarm systems maintained the initial level of attention to the supported task and overall focused more on the resource management and tracking.

In alarm periods attention allocation to the supported monitoring task did not seem to be much affected by reliability of the alarm systems. All groups slightly reduced their attention to this task over time but no impact of different reliability levels became evident. This was in line with our assumption, which assumed that only reliance on automation would be affected by a miss-prone alarm system and not compliance (Meyer, 2004).

In conclusion, the current study provides further insights in the adaptation strategies of humans in relation to automation's reliability, one of the most important perceivable characteristics of automation (Lee and See, 2004). The additional value compared to previous studies originates from the level of detail in design and analysis as most of the previous studies only compared two very extreme reliability levels (e.g. Dixon et al., 2007, 2006; Rovira et al., 2007). Furthermore, the analysis of eye-tracking data provided more detailed insights into the impact of alarm systems on attention allocation as compared to the consideration of just performance measures in previous studies.

With regard to practical implications, results are certainly not applicable to high risk work domains like aviation where only alarm systems are used that are optimized in reliability with respect to avoidance of misses and, thus, if ever typically are false-alarm prone. But in other domains like quality control inspection in the manufacturing industry comparable reliability levels even in terms of miss-prone alerting systems, can be found. In this case the finding of a critical reliability cut-off should be taken into account when considering the implementation of such systems. Even though consequences might not be apparent in the beginning, the cognitive effort of operators needed to compensate for the imperfect reliability of such systems could lead to severe problems in the long term, like complete performance breakdowns in the automation-supported task or an overall performance decrease when operators are responsible for multiple concurrent tasks.

5. Limitations

Regarding possible limitations of the current study, two aspects should be discussed which might limit the generalisation of results. First, given the fact that

participants in the current study only had to work for 30 minutes on the tasks, the results could possibly underestimate some of the observed effects. Especially, the compensation strategies for the alarm's unreliability in order to maintain a high monitoring performance may be difficult to maintain over prolonged periods of time. Ultimately, in terms of cognitive exhaustion, this overexertion might even lead to a complete performance breakdown (Hockey, 1997). Therefore, more research, especially longitudinal studies addressing long-term effects of imperfect alarm-support on operators' behavioural adaptation, is needed.

A second possible limitation concerns the lack of feedback when participants failed to detect a critical event they were not alerted for by the alarm system. In the present study, critical events were reset automatically if a malfunction remained undiscovered for 10 seconds; no consequences became apparent in this case. However, feedback is critically important for operators to get a clear picture of the level of reliability of a system and to adapt their behaviour accordingly. In a lot of systems, when feedback is not provided or evident, the operator does not know that s/he has failed to detect an alarm system's failure. In real life, misses committed by alarm systems often are linked to severe consequences, albeit these might be delayed somewhat in time (e.g. an overheating of an engine that only after some time leads to a breakdown or engine fire). Nonetheless, in the current study participants still adapted to the alarm systems' reliabilities even without feedback as became evident in the increasing performance in the monitoring task.

References

Bagheri, N., Jamieson, G.A., 2004. The impact of context-related reliability on automation failure detection and scanning behaviour, in: Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, 212-217.

Bahner, J.E., Hüper, A.-D., Manzey, D., 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experiences. International Journal of Human-Computer Studies 66, 688-699.

Bailey, N.R., Scerbo, M.W., 2007. Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. Theoretical Issues in Ergonomics Science 8, 321-348.

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.

Bliss, J.P., Dunn, M.C., 2000. Behavioural implications of alarm mistrust as a function of task workload. Ergonomics 43, 1283-1300.

Breznitz, S., 1984. Cry-wolf: The psychology of false alarms. Hillsdale, NJ: Erlbaum.

Buchner, A., Erdfelder, E., Faul, F., 1997. How to use G*Power [Computer manual]. Available at http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html.

Bustamante, E.A., Anderson, B.L., Bliss, J.P., 2004. Effects of varying the threshold of alarm systems and task complexity on human performance and perceived workload, in: Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA: Human Factors Society, 1948–1952.

Comstock, J.R., Arnegard, R.J., 1992. The multi-attribute task battery for human operator workload and strategic behavior research (Technical memorandum No. 104174). Hampton, VA: NASA Langley Research Center.

Dixon, S.R., Wickens, C.D., Chang, D., 2004. Unmanned aerial vehicle flight control: False alarms versus misses, in: Proceedings of the 48th Annual Meeting of the Human Factors & Ergonomics Society, Santa Monica, CA: Human Factors Society, 152-156.

Dixon, S.R., Wickens, C.D., 2006. Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high workload. Human Factors 48, 474–486.

Dixon, S.R., Wickens, C.D., McCarley, J.S., 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? Human Factors 49, 564-572.

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.

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

Hockey, G.R.J., 1997. Compensatory control in the regulation of human performance under stress and high workload. Biological Psychology 45, 73–93.

Lee, J.D., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243-1270.

Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. Human Factors 46, 50-80.

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

Manzey, D., Reichenbach, J., Onnasch. L., 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. Journal of Cognitive Engineering and Decision Making 6, 57-87.

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

Meyer, J., 2004. Conceptual issues in the study of dynamic hazard warnings. Human Factors 46, 196-204.

Meyer, J., Bitan, Y., 2002. Why better operators receive worse warnings. Human Factors 44, 343-354.

Miller, W.D., 2010. The U.S. air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior (Technical report No. AFRL-RH-WP-TR-2010-0133). Wright-Patterson, OH: Air Force Research Lab. Retrieved from http://dodreports.com/pdf/ada537547.pdf.

Molloy, R., Parasuraman, R., 1996. Monitoring an automated system for a single failure: Vigilance and task complexity effects. Human Factors 38, 311-322.

Moray, N., Inagaki, T., 2000. Attention and complacency. Theoretical Issues in Ergonomic Science 1, 354-365.

Muir, B.M., 1987. Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies 27, 527-539.

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37, 1905-1922.

Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. Human Factors 52, 381-410.

Parasuraman, R., Molloy, R., Singh, I.L., 1993. Performance consequences of automation induced “complacency”. The International Journal of Aviation Psychology 2, 1-23.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 30, 286-297.

Prinzel, L.J., De Vries, H., Freeman, F.G., Mikulka, P., 2001. Examination of automation-induced complacency and individual difference variates (Tech. Memo. No. TM-2001-211413). Hampton, VA: NASA Langley Research Center.

Pritchett, A., 2001. Reviewing the role of cockpit alerting systems. Human Factors and Aerospace Safety 1, 5-38.

Rovira, E., McGarry, K., Parasuraman, R., 2007. Effects of imperfect automation on decision making in a simulated command and control task. Human Factors 49, 76–87.

Sharma, H.O., 1999. Effects of training, automation reliability, personality and arousal on automation-induced complacency in flight simulation task. PhD diss. (unpublished), Banaras Hindu University.

Singh, I.L., Molloy, R., Parasuraman, R., 1997. Automation-induced monitoring inefficiency: Role of display location. International Journal of Human-Computer Studies 46, 17–30.

Singh, I.L., Sharma, H.O., Singh, A.L., 2005. Effect of training on workload in flight simulation task performance. Journal of the Indian Academy of Applied Psychology 31, 81-90.

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

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., Goh, J., Hammer, B., 2005. Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. Paper presented at the 13th International Symposium on Aviation Psychology, April 18–21, Oklahoma City.

Wickens, C.D., Dixon, S., 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science 8, 201-212.

Wickens, C.D., Rice, S., Keller, D., Hutchins, S., Hughes, J., Clayton, K., 2009. False alerts in air traffic control conflict alerting systems: Is there a “cry wolf” effect? Human Factors 51, 446-462.

Wiczorek, R., Manzey, D., 2010. Is operators’ compliance with alarm systems a product of rational consideration?, in: Proceedings of the HFES 54th Annual Meeting, Santa Monica: Human Factors and Ergonomics Society, 1722-1726.

Wiegmann, D.A., Rich, A., Zhang, H., 2001. Automated diagnostic aids: The effects of aid reliability on users’ trust and reliance. Theoretical Issues in Ergonomic Science 2, 352-367.

Wiegmann, D.A., Cristina Jr, F.J., 2000. Effects of feedback lag variability on the choice of an automated diagnostic aid: a preliminary predictive model. Theoretical Issues in Ergonomics Science 1, 139-156.