Likelihood alarm systems: The impact of the base rate of critical events, the cost of alarm validity information, and the number of stages on operator's performance

VORGELEGT VON

M.Sc.

MAGALI BALAUD

geb. in Chambéry

von der Fakultät V - Verkehrs- und Maschinensysteme
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Philosophie
- Dr.-Phil. -

genehmigte Dissertation

Promotionsausschuss

Vorsitzender:	Prof. Dr. Nele Rußwinkel
Gutachter:	Prof. Dr. Dietrich Manzey
Gutachter:	Prof. Dr. Hartmut Wandke

Tag der wissenschaftlichen Aussprache: 11. November 2015

Berlin 2015
D83

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, den

Magali Balaud

ACKNOWLEGMENTS

My sincere thanks go to my PhD supervisor Prof. Dr. Dietrich Manzey for offering me the opportunity to pursue my PhD among Prometei and for offering valuable guidance for the achievement of this PhD project. I have learnt a lot from working with him over the past three years. I would like to thank him for being supportive and for devoting so much time to his PhD students.

I would also like to thank Prof. Dr. Hartmut Wandke. It was always a pleasure to discuss my work with him and I am very grateful for the valuable insights I received from him.

I would also like to thank Prof. Dr. Nele Rußwinkel for accepting being a member of my thesis committee.

Many other people contributed to and facilitated this work. Achieving this PhD would have been much longer and difficult without the financial, intellectual, technical, organizational and emotional support I received.

I gratefully acknowledge the DFG, the Zentrale Frauenbeauftragte and the Fachgebiet AIO for offering the financial support that made this PhD work possible. Thanks to Ulrike Wiedensohler and Sandra Widera for offering the organizational support needed to facilitate the successful achievement of a PhD.

A PhD work is also not possible without technical supports. Thanks Marcus Bleil, for programming M-TOPS and for making all our research wishes possible to investigate.

I am very grateful to Maximilian Kraft and Frederik Görtelmeyer for helping conducting the experiments and collecting the data of the experiments presented in this PhD Dissertation.

I was lucky enough to have a mentor. Thanks Dr. Rebecca Wiczorek for the guidance and for being available when I needed it. Thanks for the valuable advices during the last weeks of writing.

When one starts the PhD, one becomes a new family. Thanks to my PhD sisters and brothers: Jessika Reissland, Dr. Linda Onnasch, Simon Müller, Helene Cymek and particularly Torsten Günzler and Maria Luz. It has always been nice and helpful to discuss methodological or theoretical issues with them. Thanks for giving valuable inputs during the research meetings or various papers.

I had the chance to have a second new family: Prometei. This group has been a source of friendships and interdisciplinary exchange. I would like to give a special thanks to Ekaterina Ivanova for the amazing friendship, the emotional and motivational support along these three years. I am grateful to Stephan Lindner for the patience he had with me, helping me from the very beginning to translate and write documents in German. Thanks Sebastian Werk for being my PhD writing marathon partner. They have been extremely useful and productive.

There are many more people who have helped me in various ways. Thanks to Lucie Colpaert, Oulmann Zerhouni, Dr. Marie-Pierre Fayant, Prof. Dr. Dominique Muller and Prof. Dr. Michel Dubois for supporting my application to Prometei. Nothing would have happen without the help of Pawel. I am very grateful for the long hours he spent discussing my application and for the help proof-reading the first research proposal I wrote and my application to Prometei.

A lot of friends helped me proofreading abstracts and articles in English or German along these three-years. Thanks to Alex, Elgin, Kap, Mirko, Franziska and John. Thanks Alice, Robert and Elsa for the inputs in this final document.

Last but not least, I would like to express my profound gratitude to Tobias for proofreading the entire PhD and for the loving support.

ABSTRACT

Traditional alarm systems emit two types of outputs, namely alarms and non-alarms. They are therefore called binary alarm systems (BAS). Ideally, an alarm should go off only if a critical event occurs. This is unfortunately not the case and they tend to generate a lot of false alarms. This means that an alarm goes off even if there actually is no critical event. False alarms affect operators, which start reacting slower to emitted alarms or even start ignoring them. This has been referred to as the cry-wolf effect that can result in a loss of safety and productivity.

A possible solution to the cry-wolf effect is the use of Likelihood Alarm Systems (LAS). Contrary to BAS, LAS emit different types of alerts depending on the likelihood that a critical event occurs. Two main benefits of LAS over BAS had been suggested in the literature. First, LAS might better support operators to differentiate between true alarms and false alarms and therefore make their decisions more accurate. Second, LAS might allow for a better allocation of operators' attention between the alarm task and other concurrent tasks they are responsible for.

Despite these assumptions of benefits, studies investigating the benefit of LAS over BAS have provided very inconsistent results. These inconsistent results are likely to be due to the presence of moderating variables, which affect the potential benefit that LAS have over BAS. The base rate of critical events and the cost and ease of consulting to Alarm Validity Information (AVI) might be among these moderators.

This PhD Dissertation aims clarify these inconsistent results. It does so by investigating two of these potential moderating variables as well as the characteristics of LAS that lead to the best possible performance. Three laboratory experiments have been conducted for this purpose that put participants in the role of operators working in control rooms of a chemical plants. Participants were concurrently responsible for three tasks and in one of these tasks, participants were assisted by an alarm system. In the first experiment the effect of LAS and BAS on participants´ behavior and performance was investigated while manipulating the base rate of critical events. In the second experiment the cost of consulting additional information

was manipulated. In the third experiment the effect of the number of stages of LAS on performance was investigated.

Results show that the base rate of critical events is an important moderator of operators' behavior and performance. If the base rate is low, LAS improve performance compared to BAS, but only regarding to the number of false alarms emitted. If the base rate of critical events is high, BAS support operators better to detect critical events. Furthermore, the findings show that when participants have access to additional information, LAS do not help participants to detect more critical events than BAS, whatever the cost of consulting additional information is. The results also show that 4-stage LAS support operators´ performance in the best way and that more than four stages (e.g., 5-stage LAS) do not further improve performance.

The results clearly suggest that LAS should not be implemented in high base rate environments and in settings in which additional information is available to participants. The key takeaway of this PhD Dissertation is that there is no optimal alarm system design or characteristic that can be applied across the diversity of existing task settings and environments. One useful future line of research on alarm systems is the development of adaptive alarm systems, which would constantly display the optimal characteristics to any evolving situations in order to improve the best safety and productivity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AT: Alarm Task

AVI: Alarm Validity Information

BAS: Binary Alarm System

CT: Computed Tomography

CET: Coolant Exchange Task

EV: Expected Value

EICAS: Engine Indication and Crew Alerting System

FMV: Fragebogen zur Mehrdimensionalen erfassung von Vertrauen

LAS: Likelihood Alarm System

MAT-Battery: Multi-Attribute-Task-Battery

M-TOPS: Multi-Task Operator Performance Simulation

NASA-TLX: NASA Task Load Index

NPV: Negative Predictive Value

PhD: Doctor of Philosophy

PPV: Positive Predictive Value

ROT: Resource Ordering Task

SafAS: Safety Augmentation System

SDR: Standard Deviation of Residuals

SDT: Signal Detection Theory

TCAS: Traffic alert and Collision Avoidance System

1. Introduction

More and more technologies are present in our daily life. They direct a car driver's attention to the presence of an obstacle when going backward, analyze what is missing in the fridge before going to the grocery store, or help to decide what is the best itinerary to drive from Berlin to Paris by car depending on user needs such as fast, economic, ecological, etc. These technologies support humans in performing tasks that humans used to perform entirely on their own: acquiring information about the environment, analyzing information, taking a decision and in some cases implementing the action. These technologies are called automation: technologies which perform functions previously performed by humans (Sheridan & Parasuraman, 2006).

Automation can be also found in work environments such as cockpits of airplanes (Bliss, 2003a), nuclear power control rooms (Roth & O´Hara, 1999), or intensive care units (Borowski et al., 2011). In these very complex, multitask, and high stress working environments automation is extremely important. Their goal is to enable the human-automation system to perform better than either the human or the automation would perform it alone.

However the use of automation does not always make performances error-free and new errors might even be introduced in some cases (Bliss, 2003a; Coiera, Westbrook & Wyatt, 2006; Garot & Durand, 2005; Sarter, Woods & Billings, 1997). A lot of problems that result from the human-automation interaction have already been identified. These include automation bias, mistrust, out of the loop unfamiliarity, cry-wolf effect, and others (Bliss & Fallon, 2006; Parasuraman & Riley, 1997; Sheridan & Parasuraman, 2006). The design of human-automation interaction in order to make performance optimal is a complex issue that cannot be resolved only by improving the performance characteristics of the automation (e.g., quality of the sensors and quality of the algorithm). The human operator (e.g., cognitive limitations and decision biases) has to be systematically taken into account for a better design of the human-automation interaction.

This PhD Dissertation focus on one kind of automation, namely alarm systems. Alarm systems are a very basic form of automation. They are implemented to support humans in

supervising complex systems or environments where a human's attention and vigilance might be too limited to detect potential problems. They are designed to direct user attention to critical events (Woods, 1995). Alarm systems are mostly used in critical environments where missing a critical event can have very serious consequences regarding safety and productivity. In order to not miss any critical events, designers of alarm systems follow the so-called engineering fail-safe approach (Swets, 1992). In this approach the alarm system is designed to go off even with very little evidence of a critical event. As a consequence, alarm systems produce, in most contexts, false alarms regularly (Barnes, Gruntfest, Hayden, Schultz & Benight, 2007; Parasuraman & Riley, 1997; Wickens et al., 2009). For example, 80% to 99% of alarms produced by medical alarm systems are false alarms (Chambrin et al., 1999; Lawless, 1994).

This high false alarms rate is not without consequences for the human-alarm system performance. Bliss (2003a) reports that between 20% (according to the National Transportation Safety Board's aviation database) and 90% (according to the U.S. Army Safety Center database) of alarm-related accidents and incidents are caused by false alarms. Operators might lose trust in the alarm system (Madhavan, Wiegmann & Lacson, 2006), react slower to emitted alarms (Getty, Swets, Pickett & Gonthier, 1995), or even ignore them (Bliss, Gilson & Deaton, 1995). This has been referred to as the cry-wolf effect (Breznitz, 1984) and results in a loss of safety and productivity (Lee & See, 2004). The cry-wolf effect is one the biggest problem caused by the use of alarm systems (Bliss & Fallon, 2006).

This PhD Dissertation investigates one of the solutions that have been proposed to address the cry-wolf effect issue: the use of Likelihood Alarm Systems (LAS). The concept of LAS was first developed in 1988 by Sorkin, Kantowitz, & Kantowitz to constitute an alternative to Binary Alarm Systems (BAS). Contrary to their binary cousins, likelihood alarm systems are composed by three or more stages. Each stage display a different color or/and wording characteristics corresponding to a different likelihood that a critical event is present. For example, a three-stage LAS could use the color 'red' and the label 'alarm' to indicate alerts having a high likelihood to truely indicate a critical event and the color 'amber' and the label 'warning' for alerts having a low likelihood to truely annonce a critical event. LAS provide more differentiated information to operators than BAS so that operators know which alerts are more likely to be a true alarm vs. a false alarm. LAS has in theory two benefits over BAS.

First, in comparison to BAS, it might better support operators to differentiate between true alarms and false alarms. If LAS users adapt their responding behavior to the LAS stages, they might miss fewer critical events than BAS users. Secondly, LAS might better support the allocation of attention between all tasks participants are concurrently responsible for. The different alerts act as cues, which inform operators if an immediate switch of attention is required or not. As a consequence, operators can optimize their performances in the ongoing tasks.

Despite these assumptions of possible advantages, studies investigating the benefit of LAS over traditional BAS have provided very inconsistent conclusions. Some studies reported an improvement of correct response selection in interaction with alarms (Bustamante & Bliss, 2005; Bustamante, 2008; Clark & Bustamante, 2008; Clark, Ingebritsen & Bustamante, 2010; McCarley, 2009; Wiczorek & Manzey, 2014 – condition without alarm validity information). Other did not find such improvement in the alarm task but reported that LAS improve performance only in other tasks that had to perform concurrently by the operator (Wiczorek & Manzey, 2014 – condition with alarm validity information). The latter effect was attributed to the fact that operators supported by LAS got more precise information that helped them to better allocate their attention. However, still other studies did not find any difference between BAS and LAS on correct response selection (McCarley, 2009; Sorkin et al., 1988; Swanson, 2010; Wickens & Colcombe, 2007) and one study even reported a benefit of BAS over LAS (Wiczorek, Manzey & Zirk, 2014).

This PhD Dissertation aims to get a better understanding of these contradictory results by identifying, on the one hand, some of the conditions in which LAS lead to better performance than BAS and, on the other hand, the characteristics of LAS that lead to optimal performance. Three laboratory experiments have been conducted for this purpose. Participants had to accomplish three tasks that, in a simplified way, represent typical multitask demands of operators in a control room of a chemical plant. In one of these tasks, participants were assisted by an alarm system. In a first experiment, the influence of the base rate of critical event as a moderator of the beneficial effect of LAS over BAS was investigated. In the second experiment, the cost of accessing additional information in order to check the validity of the output of the alarm system was a variable of interest. Finally, the effect of the number of stages of LAS on performance was investigated in a third experiment. The performance of the

human-automation system, the responding behavior of participants, the workload and the trust were measured.

This PhD Dissertation is divided in three main parts: the theory (Chapter 2-4), the experiments (Chapter 5-8) and the discussion (Chapter 9). The Chapter 2 first introduces the reader to the concept of automation and then more specifically to alarms systems. The Chapter 3 presents the main concepts and findings about decision-making in human-alarm system interaction. In this chapter, the cry-wolf effect, one of the main issue resulting from the human-alarm system interaction, will be presented. In Chapter 4, the concept of likelihood alarm system will be described. This chapter also reviews and discusses previous works dealing with such systems. The second part of this PhD Dissertation first introduces objectives and goals to readers (Chapter 5) and then describes the methods and results of the three experiments that have been conducted (Chapter 6-8). The results of each of these experiments are also discussed regarding a priori hypotheses and previous research. Finally, the last part (Chapter 9) summarizes the results of the three studies together. It discusses to what extent the combined results of these studies contribute to the research question investigated here, as well as their relevance for the current state of the research about LAS. Practical applications resulting from these experiments are also presented.

2. Alarm systems

Alarm systems represent the most basic form of automation (Parasuraman & Manzey, 2010). Automation is defined by Moray, Inagaki and Itoh (2000) as "any sensing, detection, information-processing, decision-making, or control action that could be performed by humans but is actually performed by machine". Automations can be described using a 4-stage model of levels of automation (Parasuraman, Sheridan & Wickens, 2000). This model is inspired by the 4-stage model of human information processing (Wickens, 1984). According to this model, automation can be described regarding to what extent they support the different stages of information processing, namely information acquisition, information analysis, decision selection, and action implementation (see Figure 1).

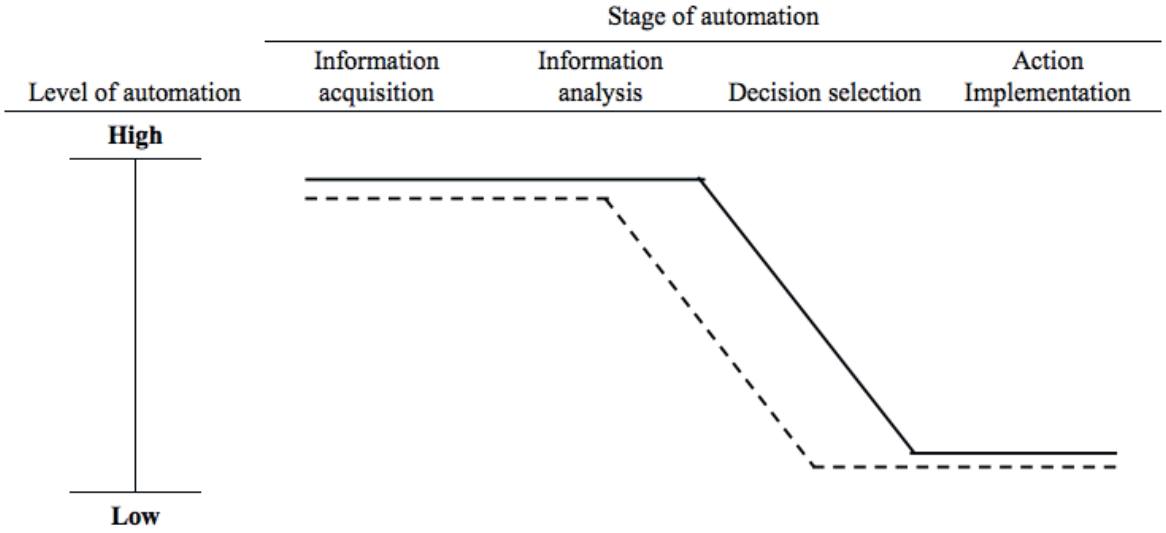


Figure 1. Model of levels of automation (after Parasuraman et al., 2000)

According to this model, alarm systems carry out the information acquisition and information analysis. More precisely, they are considered as diagnosis aids (Rice, 2009), which receive information about the actual status of a process, analyze this information, and transmit it to a human operator. The dotted line on Figure 1 represents such alarm systems. Some alarm systems also support operators in the selection of an action. This is the case for the Traffic alert and Collision Avoidance System (TCAS) that suggest operators which action should be implemented: "increase climb", "reduce descent", "climb; climb now", etc. They are

represented by the continuous line in Figure 1. Other systems are even more automated and can implement actions. This is the case for the Safety Augmentation System (SafAS), which prevents aircraft from entering into prohibited airspace. In the case of emergency situations SafAS can execute a last-resort maneuver to avoid the intrusion in prohibited airspace.

Another way to describe alarm system is regarding their degree of complexity (Pritchett, 2001). According to this categorization three main kinds of alarm systems have been identified: signal detectors, hazard detectors, and hazard resolvers. Signal detectors are the simplest kind of alarm systems. Signal detectors monitor only one parameter, and if this parameter exceeds a predetermined threshold an alarm goes off. Signal detectors are present in a lot of fields. These include intensive care units of hospitals where they monitor heart rate, oximetry, or oxygen saturation of patients. Hazard detectors are more complex than signal detectors. Contrary to signal detectors, data from a few parameters are used in order to calculate a single measure of hazard. This measure of hazard is then compared to a predetermined hazard threshold. If the measure exceeds the predetermined threshold, an alarm is released. Traffic alert and Collision Avoidance Systems (TCAS) implemented in cars are hazard detectors. TCAS use information about the speed and the distance to objects in order to calculate a projected time-to-impact and releases an alarm if the time-to-impact is critical. The most complex kind of alarm system according to the categorization of Pritchett (2001) is called hazard resolver. They work similarly to hazard detectors, with the only difference that they provide users with concrete feedbacks of which actions should be implemented. Aircraft TCAS, as described above, are hazard resolvers because they provide pilots clear indications such as “increase climb”, “reduce descent”, “climb, or climb now”.

The next chapter of this PhD Dissertation aims to present models of human-alarm interaction, describe the alarm systems characteristics in this interaction, and report main findings about human behavior in interaction with binary alarm systems.

3. Interaction with alarms as a decision-making issue

Decision-making in interaction with alarms is a complex topic because it often involves two decisions taken under uncertainty: one from the alarm system and another one from the human operator. Alarm systems take decisions about the presence of a critical event based on data, which can be ambiguous (Swets, 1992) or evolving in times (Thomas, Wickens & Rantanen, 2003) by using sensors which vary in their ability to detect and analyze a critical event (Sorkin & Woods, 1985). For these reasons, the diagnosis provided by the alarm system to the human operator can be erroneous. Based on the output of the alarm system, the operator has to decide whether or not to follow the diagnosis of the alarm system and initiate the appropriate action to deal with the critical event. This is a decision taken under uncertainty because the diagnosis of the alarm system is not 100% reliable and not all settings offer operators access to additional reliable information about the real presence of a critical event. Such additional information can take two forms depending on the settings. It can either be directly readily available to the outputs of the alarm system (e.g., the CT scan that radiologist examined to detect the presence of a tumor) or it can be actively seek by the operator in case he/she wants to cross-check a given alarm (e.g., deploying a small fire patrol to cross-check the real presence of a fire before deploying more resources). In continuation these two forms of access to additional information will be called (1) available raw data and (2) Alarm Validity Information (AVI) respectively.

In this chapter we will review the literature about human-alarm interaction. In a first part, a representation of the human-alarm interaction as a two- or three-stage decision-making process is presented. The second part of this chapter focuses on the alarm system as a decision maker. Theories and models used by human-automation researchers to describe the decision-making process and characteristics of the alarm system are outlined. The third part presents the main findings about the human decision-making process in interaction with alarm systems. Finally, the last part aims to illustrate one of the challenges resulting from the human-alarm interaction: the cry-wolf effect.

3.1. Human-alarm systems interaction

Sorkin and Woods (1985) have been the first ones to propose a model of the human-alarm system interaction. The model consists of two decision-making levels. On the first decision-making level, the alarm system makes a binary decision about the presence vs. absence of a critical event based on the current state of a process. The diagnosis of the alarm system is then emitted to a human operator in a binary way: alarm vs. non-alert. As it has been described previously, alarm systems are not 100% reliable. The accuracy of their diagnosis depends on the state of the monitored process and on the technical characteristics of the alarm system. At a second level, a human operator decides whether or not he confirms the output of the alarm system. The decision of the operator depends on the output emitted by the alarm system and on directly readily available raw data. Following this decision, the operator will initiate (or not) appropriate actions to cope with the critical event. Both decisions taken by the alarm system and the human operator resemble a signal detection task. Therefore Sorkin and Woods (1985) proposed that the decisions made by the BAS and the human operator can be modeled using the signal detection theory (Green & Swets, 1966).

One critic, which has been addressed to this model, is that it does not apply to all settings (Bustamante, 2005). For example, it does not apply to settings, in which participants can consult additional information by initiating an additional action after encountering an alarm only (i.e., settings with alarm validity information). In settings providing AVI to participants, the first decision made by the human operator after encountering an alarm is not "Is there a critical event?" but rather "Do I comply with the output of the alarm system or do I want additional information to decide?". Bustamante (2005) illustrates his critic addressed to Sorkin and Wood's model (2005) by presenting the glass cockpit settings. In commercial aviation, due to display integration no information about the critical event is readily available to pilots when an alarm is emitted. When pilots notice an alarm, they can decide to acknowledge the diagnosis of the alarm by navigating through different layers of a display in order get more information about the real presence of a critical event.

More recently Allendoerfer, Pai, and Friedman-Berg (2008) adapted the model of Sorkin and Woods (1985) to propose a model that can, among others, be applied to settings providing AVI. This model has two or three levels of decision-making depending on the decision of the

operator on the second decision-making level and on the availability of AVI. The first level of decision-making is the same as the one described by Sorkin and Woods (1985); i.e., the alarm system decides if a critical event is present vs. absent. On the second level, the operator decides if they trust (i.e., comply with the alarm and initiate the appropriate action vs. ignore the non-alert) vs. distrust the output of the alarm system. If operators distrust the output of the alarm system, they should either look for further information or wait that further information is available. If operators cross-checked AVI, then a third level of decision-making takes place in which operators decide about the presence vs. absence of a critical event. This decision can be in agreement or disagreement with the alarm system. Figure 2 depicts this model.

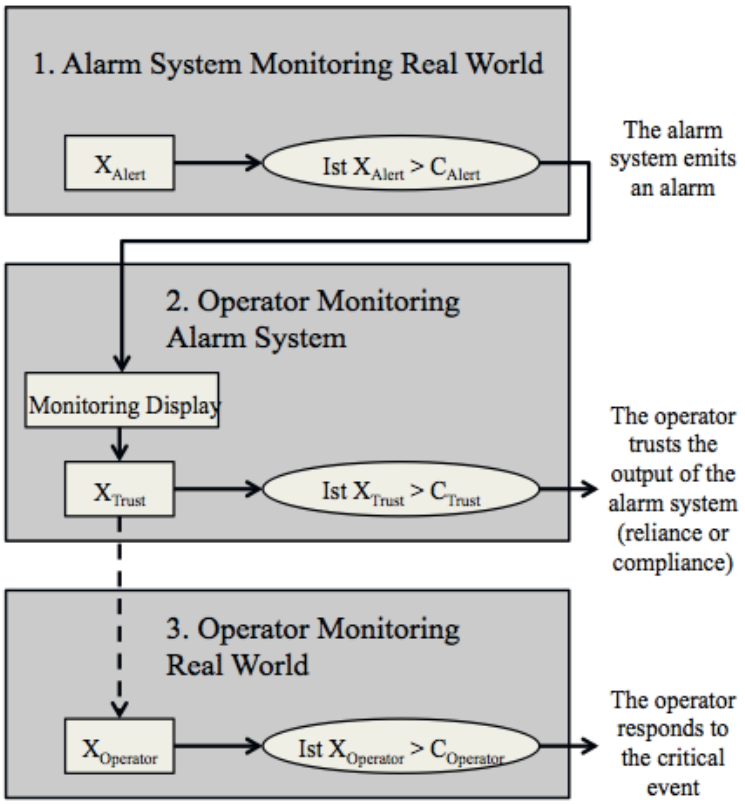


Figure 2. Three-stage human-alarm system interaction model from Allendoerfer, Pai, and Friedman-Berg (2008)

The model proposed by Allendoerfer, Pai, and Friedman-Berg (2008) is the most suitable for the research presented in this PhD Dissertation given that investigations only take place using

experimental settings representing situations in which no raw data is directly readily available for participants.

3.2. Decision-making at the alarm level

3.2.1. The signal detection theory

The first level of decision-making in human-alarm interaction, the decision made by the alarm system, can be represented within the framework of the signal detection theory (Green & Swets, 1966; Wickens, 2002). The Signal Detection Theory (SDT) was developed in the 1950s and represented a major theoretical advance to model decisions under uncertainty at its time. Indeed, this theory allows a distinction between the information processing (e.g., the alarm system aggregates sensory evidence about the presence of a critical event) and the decision process (i.e., the alarm system decides based on this evidence if a critical event is really present or not) (Egan, 1975). This distinction was not possible in classical psychophysical methods, such as the method of limits, the method of adjustments, and the method of constant stimuli. The SDT can be applied to any situations in which there are two discrete states of the world and in which a responder, in this case the alarm system, has to decide which state has occurred. These two states correspond either to the presence of the absence of a critical event. The state of the world in which the critical event is absent is also called noise. According to this theory, the noise corresponds to the random variations of the environment. These variations of the environment are present even if the signal is absent.

This theory is extremely useful to describe binary alarm systems because their function is precisely to detect the presence of a critical event (i.e., deviation from the normal state of the world) and to inform the operator about it in a binary way (no-alarm vs. alarm goes off). For example, an in-flight icing alarm system assists a pilot by detecting significant ice buildup on the wings of the plane (i.e, presence of a critical event) or not (i.e., absence of a critical event) and will inform the pilot about it in a binary way.

The Figure 3 illustrates the signal detection theory model. The x-axis represents the intensity, or strength, of the sensory evidence (e.g., quantity of ice on the wings). The y-axis indicates

how likely this intensity of evidence is to occur. The curve on the left represents the noise distribution and the curve on the right the signal-plus-noise distribution. The SDT assumes that the signal is added to the noise. As a consequence, the signal-plus-noise distribution has the same shape than the noise distribution but is shifted to the right.

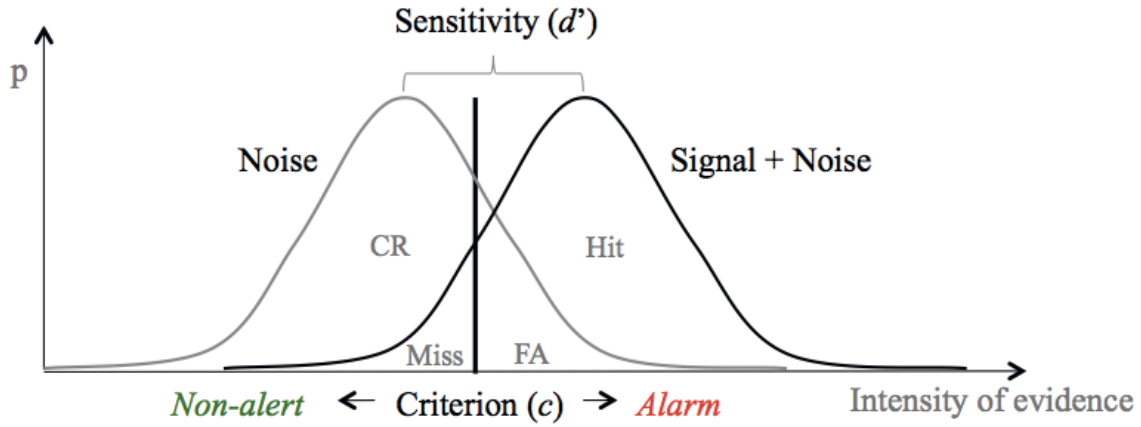


Figure 3. The model of SDT

Sometimes, it is difficult for the alarm system to discriminate if the evidence comes from the noise or from the signal. This is the case for example, when the noise alone produces more evidence than the signal-plus-noise. This uncertainty is represented in the SDT by the overlap of the two curves. If both curves were completely overlapping, this would mean that the alarm system would not able to differentiate at all between noise and signal. The distance between the two curves is called the sensitivity (*d'*). The larger this distance, the better an alarm system can differentiate between noise and signal. The sensitivity of an alarm system depends on the quality of its algorithms and sensors, and of the quality of the data that the alarm system becomes (Wickens & Colcombe, 2007). In most alarm systems, the sensitivity is not perfect because of technical limitations. Moreover, even in the cases in which it would be possible to implement extremely reliable sensors, their production would be so expensive that the technology could not be put on the market. The consequence of these technical and financial limitations is that most alarm systems are imperfect and are sometimes not able to clearly differentiate signal from noise.

Even in cases where the alarm system is uncertain about the presence of a critical event, it has to make a decision and transmit it to a human operator. In these cases the alarm system will

interpret the evidence as noise or signal depending on the placement of the response criterion (*c*), also called decision threshold. It divides the x-axis in a dichotomous way. If the intensity of the evidence is above the criterion, the alarm system will go off, signaling that a critical event is occurring. If the intensity of the evidence is below the criterion, no alarm will be released.

The sensitivity (1) and the criterion (2) can be calculated using the following formula:

(1) $d' = z(pHit) - z(pFA)$

(2) $c = -0.5 * z(pHit) + z(pFA)$

Four different combinations are derived from the state of the world and the output of the alarm system: hit, miss, false alarm, and correct rejection. These four different combinations of the state of the world and the response emitted by the alarm system are presented in the table 1. Hits and correct rejections are correct decisions from the alarm system and misses and false alarms are wrong decisions.

Table 1. 2x2 matrix of the four possible combinations resulting from the SDT.

		Output of the alarm system	
		Alarm	Non-alert
State of the world	Signal	Hit	Miss
	Noise	False alarm	Correct rejection

Knowing the distribution of these four different combinations one can calculate the hit rate (pHit) and false alarm rate (pFA) of an alarm system. The hit rate (3) is the ratio of critical

events successfully detected while the false alarm (4) rate indicates the ratio that a noise has been wrongly identified as being a signal.

$$(3)\quad pHit = \frac{hits}{hits + misses}$$

$$(4)\quad pFA = \frac{false\ alarms}{false\ alarms + correct\ rejetions}$$

While the sensitivity of an alarm system remains static, the placement of the response criterion is usually chosen by the engineer and depends on the cost associated with each decision outcome. If a miss is more costly than a false alarm, a liberal criterion is usually preferred. This placement of the criterion will maximize the hit rate but will also increase the false alarm rate of the alarm system. If a false alarm is more costly than a miss, then a conservative criterion should be preferred. This placement of the criterion will decrease the false alarm rate but will also inevitably decrease the hit rate so that more critical events will be missed. Engineers will usually prefer a liberal setting for in-flight icing alarm system given that the cost of not detecting ice buildup on the wings of a plane is very high. On the opposite, engineers will prefer a conservative criterion for an alarm system to be used in the army and that has the goal of detecting military buildings. In such an example the cost of a false alarm (destroying a civil building) is very high.

3.2.2. The positive predictive value and negative predictive value

A very useful descriptor of the alarm system in the context of an alarm-human interaction is the Positive Predictive Value (PPV) and Negative Predictive Value (NPV). They are of greater interest for human-alarm interaction researchers because they correspond to the mental model that operators have of an alarm system's reliability (Getty et al., 1995). The PPV is the conditional probability that, given an alarm, a critical event actually exists. In other words, this is the ratio between the hits and the number of times that an alarm went off (hits and false alarms). The NPV is the conditional probability that, given a non-alert, only noise is present. This is the ratio between the correct rejections and the number of times that an alarm system emits a non-alert (correct rejections and misses).

The PPV (5) and the NPV (6) can be calculated using the following formula:

$$(5)\quad PPV = \frac{hits}{hits + false\ alarms}$$

$$(6)\quad NPV = \frac{correct\ rejections}{correct\ rejections + misses}$$

For example, a PPV of .3 means that out of all alarms emitted by the system, 30% are hits and 70% are false alarms. In other words, the alarm system is correct 30% of the time when it goes off. A NPV of .9 means that when the alarm systems stay silent (i.e., non-alert), 90% were correct rejections and 10% were misses (i.e., the alarm system was correct 90% of the time in signaling that no critical event is present).

Two parameters can influence the PPV of an alarm system: the placement of the criterion and the base rate of critical events (Getty et al., 1995; Meyer & Bitan 2002; Parasuraman, Hancock & Olofinboba, 1997). Concerning the placement of the criterion, the more liberal it is set, the lower is the PPV of the alarm system. Because of the "engineering fail-safe approach" (Swets, 1992), engineers tend to set the criterion liberal in order not to miss any critical events. As a consequence, alarm systems produce a lot of false alarms. Concerning the base rate of critical events, the lower the base rate of critical event is, the lower the PPV is. So even if an alarm system has a high sensitivity, its PPV can end up being low because of the low base rate of critical events (Parasuraman et al., 1997). The relation between the alarm system´ PPV and the base rate of critical events is displayed in Figure 4.

As evocated previously, the PPV is of great interest for human-alarm interaction researchers because it corresponds to users' mental representation of the system reliability. For that reason operators tend to adapt their responding behavior to the PPV of the alarm system. The PPV is thus a useful predictor of operators' reaction time (Getty et al., 1995) and compliance rate toward the alarm system (Bliss, Gilson et al., 1995; Manzey, Gérard & Wiczorek, 2014).

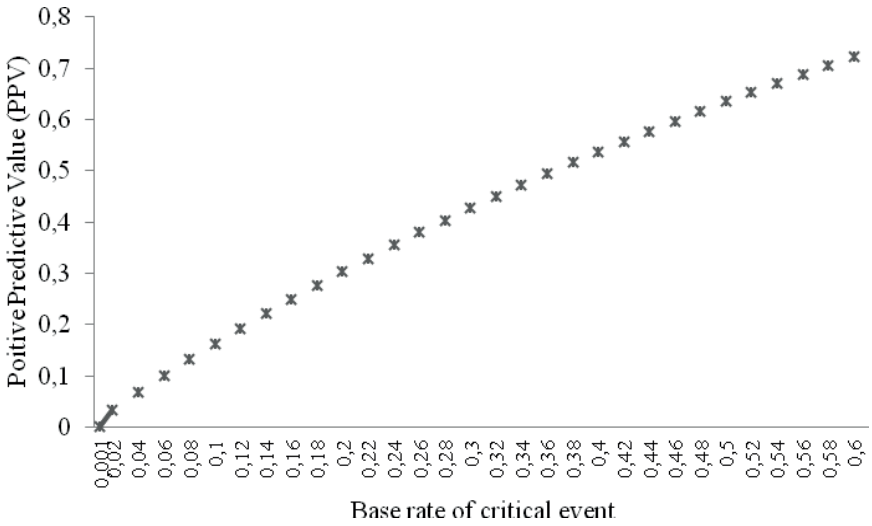


Figure 4. The positive predictive value (PPV) as a function of the base rate of critical event for an alarm system having a sensitivity (d') of 1.8 and a criterion (c) of -1.05

3.3. Decision-making at the human level

Supposing that the alarm has been perceived, the operator has to decide if a critical event is present. Depending on the presence of AVI, this decision takes place in the second or third level of decision-making according to the model of Allendoerfer et al. (2008). Depending on the decision made by the operator, four overall (i.e., human and alarm together) performance outcomes can occur when an alarm goes off:

- Hit: an alarm correctly went off (hit) and the operator initiated reparative actions;
- Miss: an alarm correctly went off (hit) and the operator ignored the alarm;
- False alarm: an alarm wrongly went off (false alarm) and the operator initiated reparative actions;
- Correct rejection: an alarm wrongly went off (false alarm) and the operator ignored this alarm.

Following the same logic, four overall performance outcomes are possible when the alarm system emits a non-alert:

- Correct rejection: the alarm system correctly emitted a non-alert (correct rejection) and the operator ignored it;
- False alarm: the alarm system correctly emitted a non-alert (correct rejection) and the operator initiated reparative actions;
- Miss: the alarm system wrongly emitted a non-alert (miss) and the operator ignored it;
- Hit: the alarm system wrongly emitted a non-alert (miss) and the operator initiated reparative actions.

According to Meyer (2004), these eight possible performance outputs, which results from the human-alarm interaction, can be described through the combination of three factors: 1) the presence of a critical event, 2) the output of the binary alarm system, and 3) the operator's decision (i.e., ignore the alarm vs. initiate reparative actions). The table 2 resumes the eight possible outputs of the combination of these three factors.

Table 2. Overall performance outcomes (i.e., alarm system and human together) depending on the state of the system, the alarm system's output, and the operator's action.

		Output of the alarm system			
		Alarm		Non-alert	
State of the world		Critical event	Noise	Critical event	Noise
Action from the operator	Initiate reparative actions	Hit	False alarm	Hit	False alarm
	Ignore the alarm system	Miss	Correct Rejection	Miss	Correct Rejection

This decision-making process has very different characteristics depending on the presence of Alarm Validity Information (AVI). AVI helps operators to identify possible errors made by the alarm system (i.e., misses and false alarms). Without access to AVI, it is very difficult, in

some cases impossible, for operators to differentiate between a hit and false alarm when an alarm goes off, or between a miss and a correct rejection when the alarm system stays silent. Numerous studies have already investigated the psychological processes of the human decision-making at the second level of decision-making (Bliss 2003b; Bliss, Gilson et al., 1995; Manzey et al., 2014; Meyer, 2001). The next part reports the main findings from the literature. Findings about operators' behavior toward alarm systems when AVI is available are presented in Part 3.3.3.

3.3.1. Reliance and compliance

The concepts of reliance and compliance (Meyer, 2001) describe two cognitive states of the operators interacting with alarm systems. They have also been described as two types of trust in alarm systems (Meyer, Bitan, Shinar & Zmora, 1999). The degree of compliance of an operator in a system affects to what extent an operator will follow the alarm system's output when an alarm goes off (i.e., the operator initiates reparative actions against the critical event). In contrast, the degree of reliance of an operator in a system affects to what extent the operator will ignore the non-alerts emitted by the alarm system. Reliance and compliance are normal operators' states to alarm systems that are 100% reliable. However, they can lead to a loss of performance if an alarm system produces either false alarms or misses. For example, if an operator complies with an alarm system that produces false alarms, he will unnecessarily interrupt the concurrent tasks he was working on. This will result in a decrease of performance in the concurrent tasks. On the opposite, if an operator relies on the non-alerts emitted by an alarm system which produce misses, the operator is less likely to detect critical events.

The best decision that an operator can make in order to optimize performance can be modeled using the expected value theory (Meyer, 2004) assuming that the operator knows a priori the PPV and NPV of the alarm systems as well as the value associated to each performance outcome (i.e., hit, miss, false alarm and correct rejection made by the operator). According to this theory, each action (initiate reparative action vs. do nothing) can be associated to an Expected Value (EV). The EV is the "sum of the products of the probabilities of the possible outcomes following this action and the value of these outcomes" (Meyer, 2004).

$$(7) \qquad EV = \sum_{all\ x} p_x V_x$$

In formula (7), p_x and V_x are the probability and the value, i.e., pay-off, of outcome x. Based on this formula, one can calculate the EV of responding that a critical event is present (8) vs. of responding that there is no critical event (9).

$$(8) \qquad EV_{CE} = P_{CE} V_{hit} + (1\text{-}P_{CE}) V_{false\ \text{alarm}}$$

$$(9) \qquad EV_{N} = P_{CE} V_{miss} + (1\text{-}P_{CE}) V_{correct\ rejection}$$

The best decision is the decision that maximizes the EV. In settings in which false alarms and misses have the same value, the EV of a decision is directly dependent on the PPV and NPV of the alarm system. If the PPV of an alarm system is lower than .5 (i.e., more than 50% of alarms are false alarms), the best decision to optimize performance is to reduce the compliance with alarms. In behavioral terms, this means that operators respond slower (Dixon & Wickens, 2006) or less often to alarms (Meyer, 2004). If the operator has to perform other tasks concurrently to the alarm task, then a reduction of compliance correlates with an improvement of performance in the concurrent tasks (Manzey et al., 2014; Wickens & Colcombe, 2007). If the operator does not reduce his compliance, the risk is that the concurrent tasks suffer of allocating too much unnecessary attention to the alarm task (Wickens, Dixon, Goh & Hammer, 2005). This can be particularly the case if additional information is available for the operators to analyze. Moreover, another factor of cost is operator needing to reorient after they return to the concurrent tasks (Bailey, Konstan & Carlis, 2003; McFarlane & Latorella, 2001).

On the opposite, if an alarm system produces a lot of misses, the best decision to optimize performance is to reduce reliance to non-alerts. In settings in which additional information is available, participants will pay more attention to them in order to detect alarm system's misses and this will degrade concurrent tasks performance (Dixon & Wickens, 2006; Dixon, Wickens & Chang, 2005). If the operator does not reduce its reliance to an alarm system producing misses, there is a risk of missing critical events that the alarm system did not detect.

Meyer (2004) did not explicitly make any statement about the independence vs. dependence of compliance and reliance. However, the description made by Meyer (2004) suggests that these two states are independent. In other words, that compliance is not affected by the NPV and that reliance is not affected by the PPV. Some studies support this statement (Dixon & Wickens 2006; Meyer, 2001; Meyer & Bitan, 2002; Wickens & Colcombe, 2007) but more recent studies have shown that a reduction of PPV affects operators' reliance whereas a reduction of NPV do not affect operators´ compliance (Meyer, Wiczorek & Guenzler, 2013; Rice & McCarley, 2011).

3.3.2. Response strategies

Participants adopt different response strategies, more or less rational, depending on the PPV value of the alarm system. Two decision strategies have been identified in interaction with alarm systems (Bliss et al., 1995; Bliss, 2003b). The first one is called extreme responding and is the most rational strategy that operators can adopt in order to optimize their performance. As it has been explained previously, assuming that the value of a miss and a false alarm is the same, the best strategy to adopt is extreme responding: comply with all alarms when the PPV is above .5 and ignore all alarms when the PPV is below .5. What has been reported so far in studies is that operators tend to adopt these strategies with alarm systems having a very high PPV or extremely low PPV. If participants tend to comply with all alarms, it is called positive extreme responding and if participants tend to ignore all alarms, it is called extreme negative responding.

Even if operators know the alarm system's PPV and the value of each performance outcome, they do not always follow this optimal solution to optimize their performance. This is caused by decisions biases when making decision under uncertainty. In this case operators tend to follow decisions heuristics instead of rational decisions. For moderate PPVs, operators tend to adopt a heuristic behavior called probability matching. Operators tend to match their responding behavior to the PPV of the alarm system. For example, if the PPV of an alarm system is .6, operators tend to follow the output of the alarm system 60% of the time although the best strategy to optimize performance would be to follow the output of the alarm system

all of the time. This decision heuristic has also been reported in other fields from the psychology (Herrnstein, 1961; Vulkan, 2000).

3.3.3. The influence of alarm validity information

Alarm Validity Information (AVI) offers participants the opportunity to access additional information in order to check the diagnosis of the alarm system. If AVI is available the decision-making model of the human-alarm interaction (Allendoerfer et al, 2008) has three levels. Operators' reliance and compliance, as well as, operators' response strategies are influenced by the presence of AVI. For example, operators are less likely to adopt an extreme response strategy when AVI is available (Bliss, 2003b). They however tend to check almost all outputs emitted by the alarm systems having a PPV below .5 and decrease their cross-checking behavior in favor of a compliant behavior for alarm systems having a PPV above .7 (Manzey et al., 2014). Manzey et al. (2014) named this strategy over-checking. The authors interpreted this behavior as a wish from participants to reduce the uncertainty in their decision even if the time investigated in the alarm task to process this cross-checking can have important consequences regarding the performance in the concurrent tasks. Interestingly, participants' cross-checking behavior is only slightly reduced if the cost of cross-checking AVI is increased but cross-checking remains the most chosen strategy for PPV under <.5. Only an increase of the workload in the concurrent tasks leads to significant changes in the pattern of participants' response behavior. In this case, participants start to adopt more extreme response strategies in the low and high-PPV conditions, expect in the PPV = .3. For such a PPV, an increase of workload in the concurrent tasks did not affect participants' behavior.

3.4. Problem with binary alarm systems: the cry-wolf effect

It has been mentioned previously that in contexts in which missing a critical event can lead to catastrophic consequences regarding safety and/or productivity, engineers tend to follow an approach called "fail-safe engineering" (Swets, 1992). In this case the criterion of alarm systems is set low enough so that an alarm goes off even with very little evidence of a critical event. As a consequence alarm systems tend to produce a lot of false alarms (Edworthy,

2013; Parasuraman & Riley, 1997; Pritchett, 2001). This false alarm rate is often magnified by the fact that in most settings the base rate of occurrence of a critical event is very low (Parasuraman et al., 1997). The consequence of this high false alarm rate is a loss of operators' trust in the alarm system (Madhavan, Wiegmann & Lacson, 2006). In behavioral terms, this can lead to what has been referred to as the cry-wolf effect (Breznitz, 1984). Most of the time the cry-wolf effect is described as a decision-making issue caused by different mechanisms like decision biases (i.e., probability matching) or classical conditioning (Bliss, Gilson et al., 1995). The term 'cry-wolf effect' has been taken from the fable "The Boy Who Cried Wolf". According to the story, a young shepherd boy felt bored of looking after the sheeps and sang out "Wolf! Wolf! The Wolf is chasing the sheep!". The village came and figured out that no wolf was there: the boy had lied. Amused by the angry faces of the inhabitants, the young boy reiterated his trick once more and triggered the same reactions than before. Suddenly a real wolf showed up. The young boy informed the village by singing up "Wolf! Wolf!". This time nobody believed him and the wolf devoured the sheeps.

In the same way, alarm systems often wrongly inform operators about the presence of a critical event. As a consequence, operators tend to respond slower or even partially (e.g., if the probability matching heuristic is adopted) or completely (e.g., if the extreme negative response strategy is adopted) ignore the alarms (Bliss, Gilson et al., 1995; Getty et al., 1995). Some evidence of cry-wolf effect can usually be observed for alarm systems having a PPV bellow .7 and the cry-wolf effect becomes stronger and stronger the lower is the PPV (Getty et al., 1995; Manzey et al., 2014). The cry-wolf effect becomes also stronger if operators' workload increases because high levels of workload negatively affect operators´ ability to respond to alarms (Bliss & Dunn, 2000). Bliss and Dunn (2000) have shown this effect both by manipulating the workload caused by the concurrent tasks (by increasing the number of tasks participants had to work on) and by manipulating the workload induced by the alarm task (by increasing alarm frequencies). Different methods have been used to measure the cry-wolf effect. These include physiological responses (Breznitz, 1984), response time to alerts (Bliss, Jeans, & Prioux, 1996; Getty et al., 1995) as well as response rates of the operators (Bliss, Gilson et al., 1995; Manzey et al., 2014; Wiczorek & Manzey, 2014).

Even though a lot is known about the way human operators behave with alarm systems, the cry-wolf issue still persists. As it was well illustrated by the fable, the cry-wolf effect is the

consequence of the uncertainty present in the human-alarm system interaction. Under such uncertainty levels, operators might adopt responding behavior, which can degrade performance. The solution to reduce or eliminate the cry-wolf effect is to decrease the uncertainty occurring in the decision-making process. There are a few solutions to do it. The first solution consists to increase the PPV of the alarm system by setting a more conservative criterion. The inconvenient being that it decreases slightly the NPV of the alarm system so that the alarm system is more susceptible to produce misses. Zirk (2013) conducted an experiment in which participants had to interact either with a high-PPV BAS (i.e., conservative criterion) or a low-PPV BAS (i.e., liberal criterion). Even if the goal of this experiment was not investigate possible effects of BAS-criterion settings on performance, Zirk (2013) figured out that participants' performance was better with the high-PPV alarm system than with the low-PPV alarm system. No cry-wolf effect has been observed toward the high-PPV BAS so that even if this BAS produced more misses than the low-PPV BAS, the performance of the overall system (i.e., human operator and alarm system together) was better.

However, for legal reasons such solutions cannot be easily deployed in most settings. Given that misses are very costly or can lead to dramatic consequences in most real settings, it is not conceivable to design an alarm system which misses some critical events in order to higher the overall PPV of the alarm system. One other countermeasure to the cry-wolf effect is to provide operators with AVI (Manzey et al., 2014). This should reduce or even eliminate the cry-wolf effect because participants tend to over-check the outputs emitted by the alarm system, as it has been described in the Part 3.3.3. This works at least for setting with a moderate workload level because increasing operators' workload diminishes operators' cross-checking behavior (Manzey et al., 2014). However even though the over-checking of AVI leads operators to detect more critical events, it can also cause a decrease of performance in the other tasks operators are concurrently responsible for. Such countermeasure to the cry-wolf effect is thus not effective in high-workload multitask environment.

Another solution to reduce or eliminate the cry-wolf effect would be to give operators cues about which alarms are more likely to truly announce the presence of a critical event. This concept already exists and has been called likelihood alarm systems (Sorkin et al., 1988). Such a solution allows, on the one hand, to conserve a high NPV and, on the other hand, to

preserve participants' performance in the concurrent tasks in multitask settings. LAS copes successfully with the issues raised by setting a conservative criterion and by the provision of AVI. Operators can use the likelihood cues to adapt their behavior; e.g., ignore only alerts having a low likelihood to announce a critical event and initiate appropriate actions in case of high likelihood alerts. Even though LAS present very promising characteristics to be the perfect countermeasure to the cry-wolf effect, this solution has so far not been much considered and only a few studies have investigated the benefits of LAS over BAS. In the next Chapter, a more detailed presentation of LAS is proposed as well as a review of the main findings about LAS reported so far.

4. Likelihood alarm systems

The concept of likelihood alarm systems was first developed by Sorkin et al. (1988) to constitute an alternative to binary alarm systems. They are composed of three or more stages

and each stage correspond to a different likelihood that a critical event is present. In other words, each level of LAS has a different positive predictive value or negative predictive value and communicates it to the operator through the use of different colors, wordings, or sounds.

The goal of LAS is to help operators to take a decision about the state of the environment. Likelihood information provided by LAS about the real presence of a critical event has a direct influence on operators responding behavior so that they tend to respond more often to high-likelihood alarms (i.e., high-PPV alarms) and less often to low-likelihood alarms (i.e., warnings or low-PPV alarms).

According to Bustamante (2008), the basic of LAS development is based on two human-automation interaction principles: probability matching and urgency mapping. Probability matching describes the tendency of people to match their responding behavior with the probability that there really is a signal when the alarm goes off (Bliss, Gilson, et al., 1995). Urgency mapping describes the tendency of people to respond more often to signals that they perceive as more urgent (Haas & Casali, 1995). These two principles allow alarm designers to expect that operators would respond more to high-likelihood-urgent signals and less to low-likelihood-non-urgent signals. It is important to notice that even though Bustamante (2008) did not say that one interaction principle is more important than the other one, prior research suggest that the probability matching principle overshadows any effect which could be attributed to the perceived urgency principle (Burt, Bartolome-Rull, Burdette, & Comstock, 1999).

In case the LAS has more than three stages, the use of colors or symbols should help the operator to know which stage is more likely to present a critical event. There are a bunch of studies investigating the effects of colors, sounds, or wording on operator's perceptions of the hazardousness of an alarm (Braun & Silver, 1995; Chapanis, 1994; Edworthy, Hellier, Walters, Clift-Matthews & Crowther, 2003; Edworthy, Stanton, & Hellier, 1995; Wogalter et al., 2002).

One can distinguish LAS from what is usually called graded alarms by the nature of the signals. LAS refer to many-stages alarm systems treating discrete signals. Graded alarm systems refer usually to analog signals. An alarm system detecting the presence of a weapon

in a luggage or of a malignant tumor on a CT-scan are examples of alarm systems working with discrete signals. A vehicle-safety warning system advising a driver about the presence of an obstacle while driving backwards or a smoke alarm detector in a house are examples of alarm systems treating analog signals. A graded alarm system has a temporal dimension in comparison to LAS: a warning is followed by an alarm if no precautious actions are done or if the danger does not go away by itself. In a LAS, there is no temporal relationship between the different stages; they only indicate a different level of likelihood that a critical event is really present.

4.1. Definition of LAS in terms of the signal detection theory

In term of signal detection theory, LAS differ from BAS by having a multiple response criterion. The evidence axis x is divided in three or more regions in contrary to BAS which only have a single response criterion dividing the evidence axis into the alarm stage and the non-alert stage.

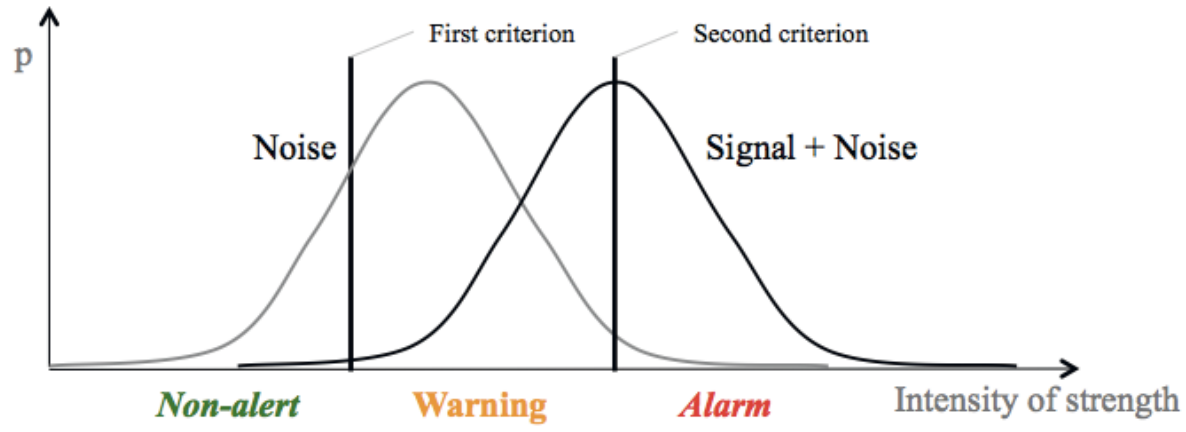


Figure 5. Representation of LAS within the framework of the signal detection theory

In the example displayed in Figure 5, a second criterion was added in the alarm stage of a BAS having a liberal criterion. As a consequence, the alarm stage has a great proportion of hits in comparison to false alarms and, in turn, the warning stage has a great proportion of false alarms in comparison to hits. The alarm stage has a very high PPV, i.e., when an alarm goes off, the likelihood that a critical event is present is very high. The warning stage has a

low PPV, i.e., when a warning goes off, the likelihood that a critical event is present is low. The red color of the alarm indicates that the likelihood that a signal is present is bigger than the warning stage displayed by the orange color (Edworthy, Stanton, & Hellier, 1995). However, some other configurations of LAS could be imagined as presented in Figure 6. The figure 6a represents a binary alarm system having a neutral criterion. As shown in Figure 6b, it is possible to set a new criterion in the alarm part of the evidence axis. The new LAS has then the same NPV as the BAS, a new warning-PPV and a new alarm-PPV. It is also possible to set the new criterion in the non-alert part of the evidence axis (see Figure 6c). The new LAS has thus the same alarm-PPV than the BAS and a new warning-PPV and NPV. A LAS can be composed of more than two criterions (Figure 6d) or the new criterions can be completely different from the ones of the BAS (Figure 6e).

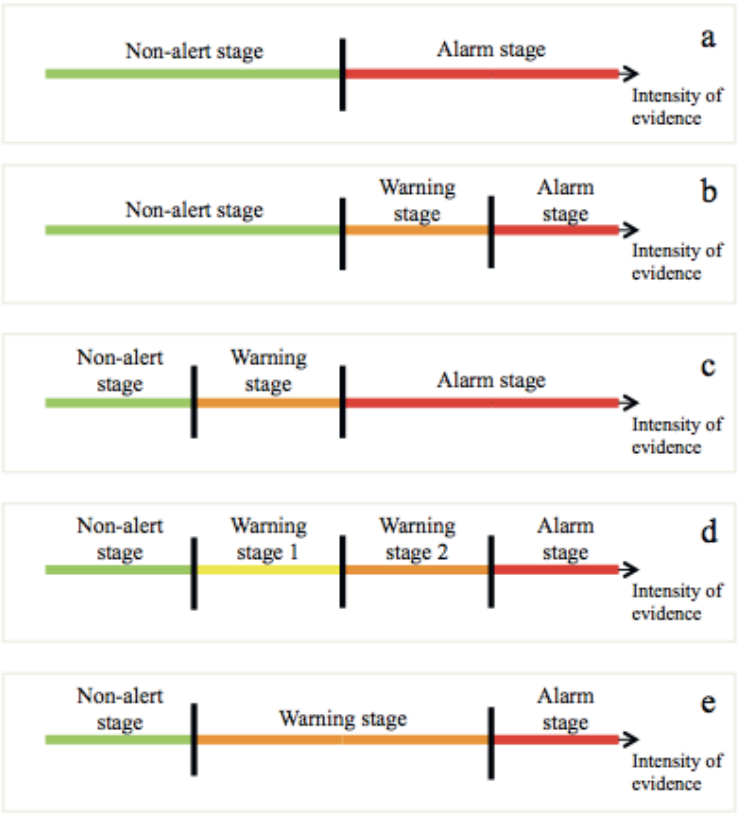


Figure 6. Example of different LAS criterion settings in comparison to a BAS having a neutral criterion

4.2. Goal of LAS

According to Sorkin et al. (1988) LAS present two main benefits over BAS. First, operators should achieve better performance in the alarm task. Because of the probability matching and urgency mapping principles, operators will respond more often to high-likelihood alarms and less often to low likelihood alarms. The cry-wolf effect is therefore limited to low-likelihood alarms only. By adapting their responding behavior to the PPV of each stage, participants have higher chance to comply with hits and to ignore false alarms produced by the alarm system. In terms of performances, operators using LAS produce more hits and fewer false alarms in comparison to operators using BAS.

Secondly, this extra information about the likelihood of a signal can be extremely useful for a good attention management. Operators using alarm systems must sometimes take care of several tasks at the same time or have high workload. Woods (1995) explains that knowing how urgent an alarm is, is important information to allow a good attention management of the attention and to protect performances in the ongoing concurrent tasks. If an alarm is less important than the ongoing concurrent tasks, an interruption of the concurrent tasks is thus an inappropriate behavior because it will degrade the concurrent task performances for no reason. For an operator, being able to assess the urgency of an alarm will enable an optimal performance in the alarm task and the concurrent tasks. Woods (1995) calls it preattentive referencing.

4.3. Literature review about LAS

4.3.1. A comparison between LAS and BAS

A few studies have compared BAS and LAS using different kinds of experimental paradigms. This part proposes an analysis of the main findings, a presentation of the most important experimental paradigms used and a review of some factors investigated as potential moderators of the beneficial effect of LAS over BAS.

As far as I am aware of, the first study which compared BAS to LAS was conducted by Sorkin, Kantowitz and Kantowitz (1988) using a dual-task paradigm. In the experiment, six participants had to perform a diagnosis decision task (alarm task) and a tracking task (concurrent task) simultaneously. The presence and the difficulty of the concurrent task were manipulated in a within-subject design so that the benefit of LAS over BAS could be investigated under varying level of workload. In the alarm task four random numbers (with two decimals) were presented to participants. These numbers were randomly taken either from a normal distribution having a mean of 3.00 (noise) or a mean of 4.00 (critical event). Based on their own judgment and on the output of the alarm system, participants had to decide from which distribution these numbers were taken from. This is an experimental paradigm in which participants have direct readily access to raw data (i.e., the numbers). The alarm system was either binary or likelihood. The LAS had four stages (white/green/yellow/magenta) and was designed like in Figure 6d. A criterion was added in the non-alert stage of the BAS to separate the white and green stage and another criterion was added in the alarm stage of the BAS to separate the yellow and the magenta stage. No differences emerged between LAS and BAS in conditions where the difficulty of the concurrent task was low. However, in high workload conditions the performance in the alarm task was significantly better with LAS than BAS. With respect to concurrent task performance no differences emerged in either condition and all effects seemed to be independent of the modality of the alarm signals (auditory vs. visually).

Fifteen years later Wickens and Colcombe (2007) conducted a very similar study. Like Sorkin et al (1988), they used a dual-task paradigm to compare BAS and LAS in which participants could use both their own judgment and the diagnosis emitted by an alarm system to take a decision. Unlike Sorkin et al (1988) they assessed participants' responding time to alarms in order to have a direct measure of the cry-wolf effect. Participants performed either a tracking task or a computational task as concurrent task. They also performed an air traffic conflict detection task that was supported by an alarm system. In this alarm task they had to monitor a screen for potential collision threats between airplanes. If a threat was detected, the participant had to click on the intruding plane to indicate that it has to change its direction. Like in Sorkin et al.'s experiment (1988) the workload and the modality of the alarm (auditory vs. visually) were manipulated. Another variable of interest in this study was the placement of the criterion

in the alarm system (liberal vs. conservative). Wickens and Colcombe (2007) found no clear evidence that the use of LAS improves performance in the alarm task and in the concurrent task. Their explanation for this absence of any effect was the simplicity of the decision in the alarm task. Participants were given the possibility to inspect raw data and this might have reduced the benefit of LAS on performance. They assumed that in real environments the decision task supported by the alarm would be much more complex and it would therefore be reasonable to expect an improvement of performances due to LAS. No effects of the workload, the modality of the alarm, and the placement of the criterion were found. Finally they found that participants needed more time to respond to alarms emitted by LAS than BAS and interpreted this difference as LAS having a greater information content that participants have to process.

Schurtleff (1991), Swanson (2010), and Mc Carley (2009) also conducted studies comparing BAS with LAS using an experimental paradigm in which participants could use both their own judgment about the raw data as well as the diagnosis of the alarm system to take a decision. However they differ from Sorkin (1988) and Wickens & Colcombe (2007) by using a single task paradigm. Like Wickens and Colcombe (2007), Swanson (2010) failed to find any benefit of LAS over BAS and any moderating effect of the placement of the criterion. McCarley (2009) found an effect of the placement of the criterion. A benefit of LAS having an alarm-PPV of .96 (liberal criterion) over BAS having a PPV of .84 was found. However, a LAS having an alarm-PPV of .99 (conservative criterion) had no beneficial effect over this same BAS. Unfortunately, these studies did not report sufficient information about the design of the alarm system and the PPV of each stage so that any comparison between the results of Wickens and Colcombe (2007), Swanson (2010), and McCarley (2009) is difficult. Shurtleff (1991) found a benefit of LAS over BAS but only when the difficulty of decision in the alarm task was increased because, according to Shurtleff, participants could rely less on their own judgment.

A common characteristic of all the studies presented so far is the direct access to task raw data offered to participants. The experimental paradigms used correspond to real-world situations like a radiologist being assisted by an alarm system in order to decide, based on a CT-scan, if a patient has a tumor or not. Another example is an airport security operator being assisted by an alarm system in order to decide if a weapon is present or not in a luggage. In these

paradigms, participants rely on both, their own interpretation of the raw data as well as on the diagnosis emitted by the alarm system to take a decision. As one can see in the results of these studies only small effects emerged. One of the explanations proposed by the authors is that the presence of this additional information might make the decision task too easy so that participants performed by themselves close to the perfect level (i.e., ceiling effect) and no difference can emerge between BAS and LAS. This raises the question of what happens if an extra action is needed from participants to access additional information.

Bustamante (2008), Bustamante and Bliss (2005), Clark and Bustamante (2008), and Clark, Ingebritsen and Bustamante, (2010) have conducted studies using a paradigm in which this access to additional information represents a cost in time (an extra action was needed). In other words, participants have access to Alarm Validity Information (AVI), as it has been described previously. This experimental paradigm corresponds to real-world situations like an airplane 'glass cockpit' for example. Due to display integration no information about the critical event is readily available to pilots when an alarm is emitted. When pilots notice an alarm, they can decide to acknowledge the diagnosis of the alarm by navigating through different layers of a display in order get more information about the real presence of a critical event. In comparison to the experimental paradigms offering a direct readily access to raw data, like the ones presented previously, an additional action is required from the operator to get additional information on the diagnosis of the alarm system.

The experimental paradigm used in these studies is the Multi-Attribute-Task-Battery (MAT-Battery, Comstock & Arnegard, 1992). This experimental paradigm simulates the multitask environment of pilots. Participants complete simulated round-trip flights that are composed of two concurrent tasks and one alarm task. The concurrent tasks are composed of (a) one tracking task simulating the maintain of flight level by keeping with a joystick a circle close to the center of the screen, and (b) one resource-management task simulating the maintaining of a sufficient level of fuel in the tanks. The alarm task simulates the Engine Indication and Crew Alerting System (EICAS). In this task participants have to ensure that they have at least one fully functionally engine at all times. An alarm system (BAS vs. LAS) was assisting them in this task. The LAS told participants that there was either 5% (warning) or 88% (alarm) likelihood of an engine malfunction while the BAS did not provide any differentiated information. After receiving an alarm, participants could either ignore it or cross-check AVI.

If they decided to cross-check AVI, they got access to additional 100% reliable information showing if there is an engine malfunction or not. In the first study conducted by Bustamante and Bliss (2005), a beneficial effect of LAS over BAS has been found. Participants using LAS complied with more true alarms and fewer false alarms than participants using BAS. The workload was also variable of interest in this study. The results show that the benefit of LAS over BAS was even larger under high workload condition.

In one study, Bustamante (2008) provided participants with direct access to raw data, similar to the experimental paradigms presented at the beginning of this review (Sorkin et al., 1988; Wickens & Colcombe, 2007). In a 2x2 design he manipulated the kind of alarm system (BAS vs. LAS) and the access to raw data. The experimental paradigm used was the same as the one presented before (EICAS). Because the experimental paradigm used by Bustamante (2008) provided participants with AVI, he actually compared an access to raw data + AVI vs. access to AVI. The raw data was not 100% reliable. It would only inform participants that there was either a 5% or a 88% likelihood of an engine malfunction. This information was completely redundant with the information provided by the LAS so that LAS users did not actually gain more information than BAS users in the "raw data" condition. Bustamante (2008) expected that BAS users and LAS users would have similar performance when raw data is provided because all participants benefit from likelihood information. However LAS users would show a better performance than BAS users when no raw data is provided. The results confirmed these predictions. The author did not discuss the fact that the raw data and the LAS provided participants with exactly the same likelihood information and that it is therefore not surprising that the performance of LAS users would not benefit from having direct access to raw data. The results might have been different if the information provided by the raw data and the LAS would not have been redundant. Another study having the same design but with the raw data and the likelihood information provided by the LAS not being redundant would be necessary before drawing any conclusions about the interaction effect between the kind of alarm system and the availability of raw data. Unfortunately, no such study has been conducted so far.

Thus far we presented paradigms in which participants have both access to additional information (either through raw data vs. through AVI) and the diagnosis of an alarm system to take a decision. However, it is not always the case that participants have access to additional information. Some studies comparing BAS to LAS have been conducted using an

experimental paradigm in which no direct access to raw data and no AVI are available to participants (Balaud & Manzey, 2013; Wiczorek & Manzey, 2014, Wiczorek et al., 2014). In this case, the diagnosis emitted by the alarm system is the only cue available to decide if a critical event is present or not. This kind of paradigm correspond to real-world situations like an in-flight icing alert in airplanes or police men having to decide if they send agents to a house or not, based on the diagnosis of the burglar alarm system only.

Wiczorek and Manzey (2014) investigated the beneficial effect of LAS over BAS while manipulating the presence of AVI. In one condition, participants had the possibility to cross-check the diagnosis emitted by the alarm system to access 100% reliable information (experimental condition with AVI, similar to the experimental paradigms used by Bustamante), whereas the diagnosis of the alarm system was the only available information in the other condition. The experimental paradigm used was M-TOPS and is described in Part 6.1.2. Wiczorek and Manzey (2014) found a benefit of LAS over BAS on the number of correct decisions made by participants in the condition without AVI only. Interestingly a performance benefit of LAS over BAS was found in the concurrent tasks, independently of the presence of AVI.

Thus far we have presented a few experimental paradigms that have been used to investigate the effect of LAS vs. BAS on participants' performance. These paradigms differ from each other by the access to additional information that they offer or not to participants. We first presented studies offering participants a direct access to raw data. These studies either found no benefit of LAS over BAS or found a beneficial effect of LAS over BAS that is limited to certain conditions only. The latter was the case in high workload conditions (Sorkin et al., 1988) or when the alarm task is difficult (Shurtleff, 1991). We then presented two studies conducted by Bustamante and Bliss (2005) and Bustamante (2008) using an experimental paradigm in which additional information is not directly available to participants but they can ask for it when an alarm system emits its diagnosis (AVI). In this context, a benefit of LAS over BAS was found but this effect disappeared when task raw data was made available to participants (Bustamante, 2008). Finally, we presented one study, which manipulated the presence of AVI vs. no additional information. A benefit of LAS over BAS was found in the condition without AVI. The access to additional information seems to be an important factor moderating the results obtained and will be further discussed in the section 3.3.4.

Studies that compared BAS and LAS also investigated the effect of various factors as potential moderator on the beneficial effect of LAS over BAS. One of these factors, which has been already investigated in numerous studies, is the PPV of the alarm system – see Part 3.2.1 for more details. The PPV has been operationally manipulated either by changing the placement of the criterion or by changing the base rate of critical events. The results about the moderating influence of the PPV are however mitigated. Sorkin et al. (1988) and Wickens and Colcombe (2007) did not find any moderating effect of the PPV whereas other studies did find one (Clark & Bustamante, 2008; Clark, Ingebritsen & Bustamante, 2010; Clark, Peyton and Bustamante, 2009). They found that LAS improve performance over low-PPV BAS (false alarm-prone systems) but not over high-PPV BAS (miss-prone systems). For example Clark, Ingebritsen and Bustamante (2010) showed that that a BAS having a high PPV of 1 and a low NPV of .56 (miss-prone system, conservative criterion) does not benefit from an additional criterion in order to create a LAS. However, a BAS having a PPV of .56 and a NPV of 1 (false alarm-prone system, liberal criterion) benefits from an additional criterion. These results are however difficult to discuss because no information is provided about the PPV- and NPV-characteristics of the LAS in this study. Clark, Peyton and Bustamante (2009) obtained similar results using another experimental paradigm: the weapon-deployment task. Clark and Bustamante (2008) also manipulated the PPV of the alarm systems by manipulating the base rate of critical events instead of the placement of the criterion to create a false alarm-prone vs. miss-prone alarm system. They obtained similar results than Clark et al. (2010). Finally, Zirk (2013) confirmed these results using the experimental paradigm M-TOPS. Even though this was not the purpose of her study, she could demonstrated a benefit of a high-PPV BAS (PPV=.73) over a LAS. Such result was unexpected and has not been previously found in any study.

4.3.2. Investigation of the characteristics of LAS

Most of the studies published so far compared BAS to LAS. Only four studies investigated the effect of LAS characteristics on performance and responding behavior.

Wiczorek (2012) compared three LAS that differed in the placement of their second criterion using the experimental paradigm M-TOPS without check-option. The goal was to investigate

what is the effect of the placement of the criterion on responding behavior towards alerts and on performance. It was expected that the medium-criterion LAS would lead to the best performance. More precisely it was expected that participants would produce less misses with the medium-criterion LAS than the low-criterion LAS. Indeed Wiczorek (2012) expected that participants would tend to adopt an extreme positive response strategy toward the alarm stage of the medium-criterion LAS (PPV = .88) and probability matching toward the alarm stage of the low-criterion LAS (PPV = .63). This would lead participants in the low-criterion LAS condition to inevitably miss more critical events than participants in the medium-criterion LAS. On the other hand it was expected that participants would produce fewer false alarms with the medium-criterion LAS than the high-criterion LAS. Indeed participants would tend to adopt an extreme negative response strategy toward the warning stage of the medium-criterion LAS (PPV = .29) and exhibit probability matching towards the warning stage of the high-criterion LAS (PPV = .35). Consequently, participants using the high-criterion LAS would inevitably produce more false alarms than participants using the medium-criterion LAS.

Results obtained by Wiczorek (2012) do not support these hypotheses. Firstly, participants produced more misses with the medium-criterion LAS than with the low-criterion LAS and, secondly, no difference was found between the medium-criterion LAS and the high-criterion LAS on the amount of false alarms produced by participants. The first result occurred because of the unexpected participants' compliance rate toward alarms emitted by the low-criterion LAS. The second results occurred because of the unexpected higher compliance rate toward warnings emitted by the medium-criterion LAS than by the high-criterion LAS. Wiczorek (2012) replicated this study using the paradigm M-TOPS with access to AVI. Her results show an effect of the placement of the criterion on the amount of false alarms produced by participants only. They produced more false alarms while using the low-criterion LAS than the two other LAS. A look at participants' extreme response strategies explains these results. Because more false alarms are produced by the alarm stage of the low-criterion LAS than by the alarm stage of the medium- and high-criterion LAS and that participants adopted a positive extreme response strategy toward all these alarm stages, participants using the low-criterion LAS produced inevitably more false alarms than participants using the medium- and

high-criterion LAS. The results from these two studies suggest that the placement of the criterion in the LAS has an effect on responding behavior and consequently on performance.

Shurtleff (1991) and Wiczorek et al. (2014) investigated the optimal number of stages in LAS. Shurtleff compared a BAS, a 4-stage LAS, a 6-stage LAS, a 8-stage LAS, and a control condition in which participants did not get any advices from any alarm system. He also manipulated the difficulty to interpret the raw data in the alarm task. The results show an effect of the number of stages on participants' performance in the alarm task only when the task was difficult. Participants had better performances while using the 4-stage LAS and the 8-stage LAS than BAS or no alarm. Wiczorek et al. (2014) compared a BAS with a 3-stage LAS and a 4-stage LAS. She found that participants' performance (i.e., number of misses and false alarms) was better when they used the 4-stage LAS, followed by the 3-stage LAS and the BAS.

Finally, Andre and Cutler (1998) investigated how varying forms of displaying uncertainty in LAS influenced participants' performance. Participants conducted a flight task and had to avoid collision with objects flying in their direction while minimizing the deviation made to avoid them. Participants were informed that the position of the flying object displayed was not 100% accurate so that an alarm system would inform them about how likely it is that the displayed position is actually the real position. This 3-stage LAS would display the degree of uncertainty differently depending on the condition: text, color, or graphical (e.g., a circle having three different possible sizes would appear around the flying object). A baseline was included, in which participants had no additional likelihood information about the accuracy of the position of the flying object. The results show that the graphical display of uncertainty led to the best performance.

4.3.3. Effect of LAS on trust

A few studies assessed participants' trust toward the alarm system. Their authors expected that participants would trust LAS more than BAS because of its greater transparency (Wiczorek, 2011; Wiczorek et al., 2014) and perceived reliability (Ragsdale, 2012; Wiczorek, 2012; Wiczorek et al., 2014). And indeed, all these studies found that participants trusted LAS

more than BAS, even though the overall reliability of BAS and LAS was exactly the same in all experiments. According to these authors the greater transparency of the LAS had an effect on participants' perception of alarm system's errors. Whereas LAS users interpreted only the false alarms, and not the false warnings, as being errors, BAS users interpreted all false alarms as being errors since the BAS did not make any differences between them. Consequently participants perceived LAS as being more reliable than BAS. With perceived reliability and transparency being two components of trust (Lee & Moray, 1992; Wiczorek, 2011), this explains why participants trusted LAS more than BAS.

Wiczorek et al. (2014) also showed that participants trusted a 4-stage LAS more than a BAS. However no differences in trust rating were found between the 3-stage LAS and the 4-stage LAS.

4.3.4. Critical discussion on the literature review

From this review, it emerges that some moderators of the beneficial effect of LAS over BAS might exist and could explain the differences between these results. Some of the results obtained make us believe that the access to additional information might be a relevant moderator. Indeed most of studies providing a direct access to raw data did not find any benefit of LAS over BAS. This is in comparison to studies using an experimental paradigm with AVI or without access to additional information at all. One reason to explain this absence of effect might be that the decision becomes too simple when a direct access to raw data is provided. For instance, Sorkin et al. (1988) used a diagnosis task that participants perform under low workload and without the assistance of an alarm system, close to the maximum level. This might explain why a benefit of LAS over BAS was found under high workload only: only in this condition participants probably did not have enough cognitive resources to well perform the diagnosis task. Wickens and Colcombe (2007) also highlighted the simplicity of the alarm task as a possible explanation of the absence of benefit of LAS over BAS in their study.

The chosen PPV of the alarm systems also seems to be another important factor explaining the results obtained so far. It seems that a benefit of LAS occur over low-PPV BAS only. This

could be another explanation to the absence of benefit of LAS over BAS reported by Sorkin et al. (1988) and Wickens and Colcombe (2007). Indeed in both these studies the BAS used had a PPV above .7. With such a high value of PPV, BAS users obviously tend to respond to all alarms emitted by the alarm system (Manzey et al., 2014). The absence of effect between BAS and LAS on performance in the alarm task could be therefore explained by a ceiling effect. Both BAS users and LAS users adopt optimal behavior and have thus performance close to a perfect level in the alarm task. In the studies conducted by Bustamante and Bliss (2005), Bustamante (2008), and Wiczorek and Manzey (2014) the BAS have a low PPV, .18 and .43 respectively. As a consequence, participants adopted different response strategies toward the alarm stage of the BAS and of the LAS, which in turn, had a direct effect on participants' performance.

It is however difficult to make a comparative analysis and interpretation of these studies as they used very different characteristics in the design of the alarm systems (e.g., placement of the criterion, colors and wording), in the design of the tasks used (e.g., single vs. multitasking, presence of raw data vs. of AVI), and in the design of the experiment (e.g., procedure, pay-off, within-subjects vs. between-subjects design). These characteristics might have influenced participants' behavior and, thus, participants' performance. Some authors like Bustamante and Bliss (2005), Clark and Bustamante (2008), and Sorkin et al. (1988) used a within-subjects design to compare the effect of BAS and LAS on performance. One can speculate that using up to 4 alarm systems with different reliabilities in a single same experiment might have confused participants and made it difficult for them to keep a mental model over time of the reliability of each alarm systems' stages.

The choice of colors and wording in some experiments is also questionable. In the experiment of Sorkin et al. (1988) a 4-stage LAS was used with the following combination of wording and colors: 'silent' - white, 'possible signal' - green, 'likely signal' - yellow, 'urgent signal' - magenta. This choice of colors is not coherent with research investigating the effects of colors on operators' perceptions of the hazardousness and urgency of alarms. For example, it has been shown that yellow is perceived as more urgent or is more susceptible to announce a hazard than magenta (Wogalter, Frederick, Magurno, & Herrera, 1997). In the experiment conducted by Bustamante and Bliss (2005) and Clark and Bustamante (2008), the color yellow and the wording 'warning' were used for alarms emitted by the BAS whereas the color

red and the wording ‘danger’ were used for alarms emitted the LAS. If a benefit of LAS over BAS is found, it makes it thus impossible to know if this effect was caused by the difference in color and wording or by the difference of PPV between these two stages. Unfortunately, participants’ responding behavior toward alarms has not been reported in these articles.

The cost of consulting AVI might also be an important factor influencing the results of these experiments. A possible effect of this cost can be observed while comparing the studies of Wiczorek and Manzey (2014) with the studies of Bustamante (2008), Bustamante and Bliss (2005), and Bustamante et al. (2009). These studies offer participants the possibility to consult AVI when an alarm goes off. However they differ from each other in the cost of consulting AVI (i.e., the payoff and time required). In Wiczorek and Manzey’s experiment, participants did not lose any points each time they consulted AVI whereas in the experiments conducted by Bustamante (2008), Bustamante and Bliss (2005), and Bustamante et al. (2009), participants lost one point each time they consulted AVI while the alarm erroneously went off (i.e., false alarm). Moreover, it required less time for participants to consult AVI in the experiment conducted by Wiczorek and Manzey (2014) than in the experiments conducted by Bustamante (2008), Bustamante and Bliss (2005), and Bustamante et al. (2009). Interestingly Wiczorek and Manzey (2014) did not find any benefit of LAS over BAS whereas the Bustamante’s studies found a benefit of LAS over BAS when AVI was available. Given these results, one can speculate that the difference in cost to consult AVI could have had an effect on the frequency of cross-checking outputs and thus on performance. Unfortunately data about participants’ cross-checking behavior was not reported in Bustamante (2008), and Bustamante et al. (2009). One can only speculates that participants in these studies cross-checked less often the outputs of the alarm system in comparison to participants in Wiczorek and Manzey’s experiment (2014) and that a beneficial effect of LAS over BAS could occur.

Finally the variety of dependent variables used to measure participants’ behavior and performances make any comparison between studies investigating LAS difficult. Most of authors used various forms of the performance indicator *d’* from the signal detection theory (Mc Carley, 2009; Ragsdale, Lew, Dyre & Boring, 2012; Shurtleff, 1991; Sorkin et al., 1988; Swanson, 2010; Wickens & Colcombe, 2007) to report participants’ performance in the alarm task. Bustamante (2008), Bustamante and Clark (2010), Bustamante et al. (2009), Bustamante et al. (2010), and Clark and Bustamante (2008) on the other hand used a performance

indicator called ‘decision-making accuracy’ (Bustamante, 2005). Wiczorek and Manzey (2014) and Wiczorek et al. (2014) used the sum of wrong decisions participants made in interaction with the alarm system, and Wiczorek (2012) used the number of points that participants lost in the alarm task. Finally, Bustamante and Bliss (2005) and Shurtleff (1991) also reported participants’ true alarms response rates and false alarms response rates. Regarding participants’ behavior toward the alarm system, Wickens and Colcombe (2007) and Shurtleff (1991) used the response time to alarms. Bustamante (2008), Bustamante and Clark (2010), Bustamante et al. (2009), Bustamante et al. (2010), and Clark and Bustamante (2008) used the decision-making bias, an indicator developed by Bustamante (2005) taking into account the percentage of hits and false alarms made by participants. Finally, Bustamante and Bliss (2005), Ragsdale (2012), Swanson (2010), Wiczorek and Manzey (2014) and Wiczorek et al. (2014) looked directly at the how often participant did respond either to the overall alarm system or to each stage of the alarm system.

The review has shown that further knowledge regarding the influence of alarm system’s characteristics on participants’ performance is needed. First, some moderators exist and need to be more precisely investigated. Second, likelihood alarm systems can have very different design characteristics (e.g., its number of stages) so it is important to have a clear idea of the effects of these characteristics on performance.

5. Goal of this PhD Dissertation

The main goal of this series of studies is to gain a better understanding of the conditions in which LAS should be used. More precisely, it aims (1) to identify in which situational contexts or task contexts LAS shows a benefit over BAS, (2) to investigate what LAS characteristics lead to optimal performance, and (3) to have a better understanding of the differences between LAS and BAS regarding participants' responding behavior and participants' subjective experience (workload and trust). A practical goal of this PhD Dissertation is to produce a guideline about the use of LAS.

To this purpose three experimental studies were conducted. The first and the second study directly compared BAS to LAS while investigating the influence of the base rate of critical event (Experiment 1) and the influence of the cost of consulting AVI (Experiment 2). In the third study, we examined the effect of the number of stages in LAS.

The studies were conducted using a highly controlled laboratory multitask paradigm. One of the tasks represents an alarm task. Here the participants have to react to the outputs of one alarm system. The two others tasks insured that participants had a level of workload close to real-world environments. The use of a multitask environment allows us a comprehensive understanding of how the characteristics of alarm systems impact participants' attentional allocation and performance in concurrent tasks.

In all these experiments we used the most direct measure of performance: participants' percentage of hits, misses, false alarms, and correct rejections. This serves to process a more precise analysis of participants' performance than any other kind of calculation would allow. The behavioral dependent variables used are the same as in the experiments of Wiczorek and Manzey (2014), Wiczorek (2012), and Wiczorek et al. (2014), so that a direct comparison with these studies will be possible.

6. Experiment 1: The influence of the base rate of critical events

The goal of Experiment 1 is to compare the benefits of BAS vs. LAS under different base rates of critical events (low vs. high). The base rate of critical events affects directly the PPV of alarm systems (see Part 3.2.2 for more detailed explanations). A same alarm system would have a low PPV if it is implemented in a low base rate environment and a high PPV if it is implemented in a high base rate environment. In other words, by manipulating the base rate of critical events, one actually manipulates the PPV of alarm systems.

The PPV of BAS affects participants' responding behavior. If the PPV of BAS is low, participants tend to ignore its emitted alerts and therefore miss some critical events. If the PPV of BAS is high (above .7), participants tend to comply with its emitted alerts and therefore detect most, if not all, critical events. This suggest that the level of PPV of BAS moderates whether or not advantages can be gained by the use of LAS regarding participants' performance in the alarm task. If the PPV of the BAS is low, the implementation of LAS should improve operators´ performance. In such case, LAS is an effective countermeasure against the cry-wolf effect in BAS because the cry-wolf effect will be shifted to alerts having a low likelihood that a critical event is present. However, if the PPV of the BAS is high, the implementation of LAS should not further improve performance since participants already adopt an optimal responding behavior towards BAS.

Results from previous studies confirm such statements. A few studies have reported that LAS improves performance over low-PPV BAS but not over high-PPV BAS (Bustamante & Clark, 2010; Clark & Bustamante, 2008). One study even found a benefit of a high-PPV BAS over LAS (Zirk, 2013). The results from Zirk (2013) suggest that LAS lead to worse performance than high-PPV BAS because it contains a stage (the warning stage) that participants partly ignore. As a consequence, LAS users miss more critical events than high-PPV BAS users, who comply with all alarms. However, this study has not been designed and conducted to compare LAS with BAS having varying level of PPV. The presence of a high-PPV BAS in this study was exploratory and no hypotheses have been made *a priori* about eventual benefits of the high-PPV BAS over LAS.

The current study aims to replicate the results obtained by Zirk (2013) using an experiment designed for this purpose. However, the PPV will not be manipulated by changing the placement of the criterion but by manipulating the base rate of critical events, like it has been done already by Clark and Bustamante (2008). Experiment 1 differs from the experiment conducted by Clark and Bustamante (2008) in three ways. First, participants in our experiment do not have access to any additional information contrary to participants in Clark & Bustamante (2008) experiment. Second, the difference between the two base rate values chosen in Experiment 1 is not as extreme as in the Clark and Bustamante's (2008) experiment and is more similar to real-world environments. Finally, the dependent variables used in Experiment 1 are direct measures of performance and of responding behaviors. The dependent variables used by Clark and Bustamante (2008) were not ideal to gain a thorough understanding of participants´ performances and responding behaviors. Participants´ performance was measured using the weighted sum of the percentage of right answers (hits and correct rejections):

Performance = .5*p(hit) + .5*p(CR).

This dependent variable used by Clark and Bustamante (2008) does not describe exactly how many hits, misses, false alarms, and correct rejections the participants did. It is a pity because one could expect that the benefit of BAS over LAS in a high base-rate context would be in terms of hits and misses but not in terms of false alarms and correct rejections; such statements will be more precisely explained in the hypotheses.

Moreover, Clark and Bustamante (2008) measured participants´ responding behavior using the weighted sum of the percentage of alerts participant decided to comply with:

Bias = .5*p(hit) + .5*p(FA)

Such measure does not precisely show the participants' responding behavior for each stage of the alarm system (warning stage vs. alarm stage). Knowing to what extent participants respond to each stage of the alarm system allows a deeper understanding of the performances. In Experiment 1, a more direct measure of participants´ performance and behavior was used. Participants´ performance is assessed using the percentage of hits, misses, false alarms, and

correct rejections. Moreover, participants´ response rate for each stage of the alarm system is recorded.

Experiment 1 addresses the interaction effect of the type of alarm system (BAS vs. LAS) and the PPV of alarm system (.35 vs. .65) on responding behaviors and performances. The hypotheses are presented in the section 6.2.

6.1. Method

6.1.1. Participants

Fifty-one participants (25 women, 26 men) participated in this study. The average age was 25.12 years (*SD* = 4.82). Among them, 74% were students and 26% were employed. They voluntarily registered to the experiment after having been recruited on an online portal for experiments (Prometei) of the Technische Universität Berlin. None of them was suffering from any distortion of color vision which might interfere with the experiment (i.e., red-green color blindness).

Participants were paid 5€ for their participation and they could get up to 4€ depending on their performance during the experiment. Each of them was randomly assigned to one of the four experimental conditions. An equal repartition of men and women in each group was observed.

6.1.2. Task

Experimental paradigm M-TOPS

A PC-based laboratory task, M-TOPS (Multi-Task Operator Performance Simulation), was used for the three experiments presented in this PhD Dissertation. M-TOPS was already used in various studies investigating decision-making in interaction with alarm systems (Manzey et al., 2014; Wiczorek & Manzey, 2014; Wiczorek et al., 2014). M-TOPS is a Java computer-based program that aims to simulate, in a simplified way, typical multitask demands of

operators in a control room of a chemical plant. It includes three tasks, which are typically required from operators working in this context: a *Resource Ordering Task*, a *Coolant Exchange Task*, and an *Alarm Task.* The *Alarm Task* is the most important for our studies since we manipulated its characteristics for our research purpose. The *Resource Ordering Task* and the *Coolant Exchange Task* are defined as 'concurrent tasks'. The interface M-TOPS is presented in Figure 7. The screen is divided into four parts of equal size while the lower left part is left empty. In following is a description of each task.

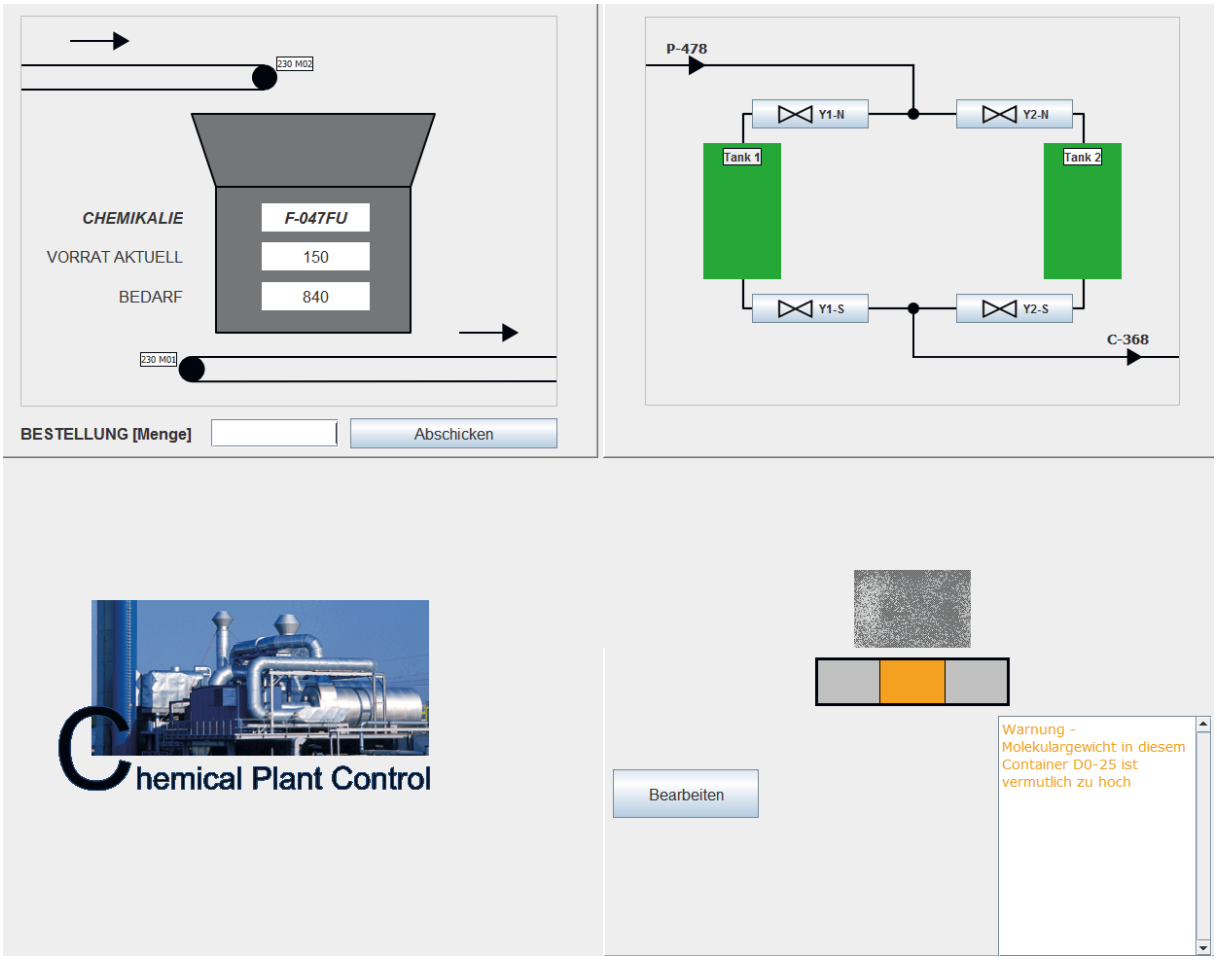


Figure 7. User interface of M-TOPS

Resource Ordering Task (ROT): This task is displayed on the upper left quadrant of the interface in Figure 7 and represents a sort of mental arithmetic task. Participants were told that they have to make sure that there is always the required amount of a specific chemical product for the chemical reaction to keep running. For this purpose, the current value and the required

value of a chemical are shown. Quantities range between 0 and 999. Participants' task is to calculate the difference between these two quantities, type the result into the ordering field and initiate the order by clicking a button. After three seconds, a new order appears on the screen. Participants have 15 seconds to complete this task. If they do not manage to submit the order within 15 seconds, the task is registered as not accomplished and the next order appears automatically. Participants were informed about these temporal conditions in the instructions. As performance measure, the number of correctly sent orders and the percentage of appeared orders participants responded to (i.e., the percentage of orders registered as accomplished, whatever they were correct or not) were sampled.

Coolant Exchange Task (CET): This task is displayed on the upper right quadrant of the interface of Figure 7. Participants were instructed that the containers in the plant have to be cooled down to insure the security of the production. To do so, the coolant in the containers has to be regularly replaced by fresh coolant to prevent over-heating. Participants' task is to initiate a few actions in a given sequence. Two containers are presented to participants with the color green meaning that the coolant has to be changed. Participants have to click on the lower valve to release the used coolant, close this valve, open the upper valve to re-fill the container with fresh coolant and close the upper valve once the container is full. The color of the container informs participants if the coolant has to be changed (green color), if the container is empty and has to be re-fill with fresh coolant (grey color), and if the container is filled with fresh coolant (blue color). A complete exchange-cycle takes 40 seconds plus the time needed by participants to enter the commands (i.e., clicks on the valves). A new task appears after the containers have been successfully filled with fresh coolant and that all valves are closed. As a performance measure the amount of refilling cycles successfully completed was registered.

Alarm Task (AT): This task, displayed in the lower right quadrant of the interface in Figure 7, represents a quality control station in which all containers have to go through before leaving the plant and being delivered to clients. More precisely, a grey square (representing a container) arrives from the left of the quadrant and takes six seconds to reach the control station in the middle of the square where the alarm system is present. One container at a time passes the control station. In the control station the alarm system performs the quality assessment of the chemical end-product of each container and delivers its output to

participants. The alarm system emits its output to participants with the use of colors (in a colored square displayed under the container) and wording (a written message displayed in a quadrant called 'alarm state monitor'). The colored and written information is redundant. The alarm system can be of two kinds: binary (BAS) or likelihood (LAS). If it is BAS, the alarm system emits either a red light combined with the written message 'The molecular weight in this container X is too high', or a green light combined with the written message 'The molecular weight in this container X is okay'. In the case of LAS, a third output option is presented to participants using the color orange and the written message 'The molecular weight in this container X is eventually too high'. The container stays five seconds in the control station during which the output of the alarm system is emitted. Based on this output participants have to decide between sending the container back to the plant (by clicking on the repair – bearbeiten – button) or letting it go to clients (by doing nothing). If the participant chooses to do nothing, the container disappears after having spent five seconds in the control station. After the container left the control station, the next container appears six seconds later. The alarm diagnosis is not 100% reliable so that participants' decision is actually a decision under uncertainty, which is comparable to decisions in real-world settings. The reliability of the alarm outputs can be manipulated. No raw data and no Alarm Validity Information (AVI) were available for participants in Experiment 1.

Alarm system

All alarm systems had the same sensitivity (*d'* =1.8). They had a first criterion set at -1.05 and the LAS had a second criterion set at 0.7.

Pay-off system

Participants got 1.5 points for each correct order in the ordering task, 7.5 points for each tank task fully completed and they lost 2 points for each wrong decision in the alarm task. This pay-off was initially chosen by Wiczorek (2012) with the aim to motivate participants to give their best in all tasks without prioritizing or ignoring a task. In order to be able to compare previous findings from studies conducted so far with M-TOPS with the results of the studies presented in this PhD Dissertation, the same pay-off system was used. A different pay-off system might have influenced decision-making and behavioral responses towards the alarm

system and make the results impossible to compare. This pay-off system was chosen by Wiczorek (2012) for the following reasons: First, this pay-off system creates a concurrence between the different tasks that is similar to the one in real-world settings. In real-world settings, when an alarm goes off, operators have to move their attention to the alarm system without completely abandoning the concurrent tasks (Wickens & Horrey, 2008). To make sure that participants do not completely abandon the concurrent tasks when an alarm goes off, participants can win points in the concurrent tasks only. Second, the pay-off system takes into account the time needed for the realization of the tasks. Participants can therefore win more points in the *Coolant Exchange Task* than in the *Resource Ordering Task* because it takes more time to complete this task. This analysis of time was conducted by Gérard (2012). At the end of the experiment participants get one cent for each accumulated point. This bonus was in addition to the five euros offered for their participation.

6.1.3. Design

A 2 x 2 between-groups design was used for this study. The type of alarm system (BAS vs. LAS) and the PPV of the alarm systems were manipulated (.35 vs. .65). The manipulation of the PPV corresponds to a manipulation of the base rate of critical events (i.e., the ratio of improper containers among the containers). Table 3 resumes the characteristics of the alarm systems used.

Since the NPV of the alarm systems is very high in all the conditions, no effect on participants´ responding behavior and performances in the non-alert stage was expected. The non-alert stage was not taken into account in our hypotheses.

Table 3. Experimental conditions (between-subjects) of Experiment 1

			Characteristics of alarm system					
Condition	Kind of alarm system	Base rate of critical events	*d'*	*c*	Global PPV	PPV alarm stage	PPV warning stage	NPV
BAS35	BAS	.24	1.8	-1.05	.35	.35	-	.97
LAS35	LAS	.24	1.8	-1.05 and 0.7	.35	.78	.21	.97
BAS65	BAS	.52	1.8	-1.05	.65	.65	-	.96
LAS65	LAS	.52	1.8	-1.05 and 0.7	.65	.91	.47	.96

6.1.4. Dependent variables

Participants´ responding behavior for the overall alarm system and for each stage of the alarm system was recorded. Participants´ performance in the alarm task and in the concurrent tasks were also recorded. Finally the trust in the alarm system was also a variable of interest. The operationalization of the dependent variables is detailed here below:

Behavioral data: participants' response rates

These dependent variables correspond to the proportions of alerts participants complied with vs. ignored. Compliance is defined by clicking on the repair button while ignoring is defined by an absence of response. Participants´ response rate is reported in percentage. We recorded participants´ response rates to alerts (non-alerts excluded) and to the different outputs emitted by the alarm systems. These included:

- Compliance rate with alerts;
- Ignoring rate of alerts;

- Compliance rate with alarms;
- Ignoring rate of alarms;

- Compliance rate with warnings;
- Ignoring rate of warnings;

- Compliance rate with non-alerts;
- Ignoring rate of non-alerts.

Behavioral data: participants´ response strategies

Looking at participants´ response strategies allows gaining a more thorough understanding of participants´ response rates. For example, if participants´ response rate to alarms is about 50% one interpretation could be that all participants adopted a sort of probability matching behavior and complied with alarms only half of the time. It could however also be that half of participants exhibited positive extreme responding while the other half exhibited negative extreme responding.

Participants´ response strategies were assessed using three categorical dependent variables:

- Response strategies in response to alarms;
- Response strategies in response to warnings;
- Response strategies in response to non-alerts.

Bliss (1993) initially identified these response strategies (as cited in Bliss, 2003b). They can be classified into three categories:

- Positive extreme responding: participants responded to 90% or more of alerts (alarms and warnings) emitted by the alarm system;
- Negative extreme responding: participants responded to 10% or less of alerts (alarms and warnings) emitted by the alarm system;

- Probability matching: participants responded to 10% to 90% of alerts (alarms and warnings) emitted by the alarm system.

Looking at participants' responding behaviors allows reaching a better understanding of participants´ performances in the alarm task but also in the concurrent tasks. For example, a high compliance rate with alerts can explain a decrease of performance in the concurrent tasks. The less often participants comply with alerts, the more time they can invest in the concurrent tasks.

Performance data

Performance in the alarm task was assessed using four dependent variables:

- Percentage of hits produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of misses produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of false alarms produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of correct rejections produced by participants when the alarm system emitted an alarm or a warning.

Three dependent variables were used to assess participants´ performance in the concurrent tasks:

- Number of correctly sent orders in the Resource Ordering Task (ROT);
- Proportion of orders participants responded to in the Resource Ordering Task (ROT);
- Number of refilling cycles successfully completed in the Coolant Exchange Task (CET).

Subjective data: Trust in the alarm system

Participants completed the FMV questionnaire (Wiczorek, 2011) to indicate their trust towards the entire alarm system. The questionnaire is composed of 16 items and measures four different sub-dimensions of trust: transparency, intention, reliability, and use. Participants responded on a four-point likert scale, including (1) do not agree, (2) rather do not agree, (3) rather agree, and (4) do agree. Participants´ responses at the questionnaire were collected with an online portal for survey named Limesurvey.

6.1.5. Procedure

The experiment took place in the human performance laboratory of the department Arbeits-, Ingenieur- und Organisationspsychologie of the Technische Universität Berlin. Up to four participants participated in the experiment at the same time. Each participant was randomly assigned to a 17-cm screen PC. Each PC had headphones.

Participants first completed an informed consent and a demographic questionnaire and were then invited to start reading the instructions (see Annex D) on the computer screen. Participants were told that the experiment was a simulation of a control room of a chemical plant and that their task was to perform three tasks in parallel in order to assure the good run of the chemical process and to control the quality of the end-product. Each task was conscientiously explained and participants could practice each of them separately during two-minute training sessions. They could then practice all tasks together. At the end of the training session participants were ask to fill in a short right/wrong questionnaire, which was then directly proofed by the instructors. This was done in order to check that the instructions have been well understood. In case of a wrong answer, the statement in question would be explained orally. Participants were then explained that the alarm system was not 100% reliable and that it could sometime provide a wrong diagnosis. This was followed by a 50-trial familiarization session (about 8 minutes) in which participants performed the alarm task only and got acoustic feedbacks via headphones after each decision they made: a sound effect representing coins for a right decision and a ‘buzzer’ for a wrong decision. They were told to use this acoustic information to get an idea of the reliability of the alarm system and it was emphasized that they would work during the experimental session with exactly the same

alarm system. After this familiarization session ended, participants´ perception of the alarm system reliability was asked. More precisely, participants had to report how many times out of 50 containers each output (alarm, warning, and non-alert) were released and for each kind of outputs how many time was the container really improper (see Annex E). The given values were transformed into participants' perceived PPV and NPV using the formula presented in Part 3.2.2. for a manipulation check. After completion of the PPV questionnaire the instructor left the room in order to avoid any instructor effect during the data collection. Participants were reminded of the pay-off of the experiment and it was emphasized that that no acoustic feedbacks would be provided anymore even though they continued working with exactly the same alarm system. The experimental session lasted about 16 minutes in which 100 containers were presented. Participants had to complete the three tasks of M-TOPS at the same time. Finally participants had to complete the FMV questionnaire (Wiczorek, 2011; see Annex F)). Once the experiment was over participants were thanked once more for their participation and the monetary compensation was offered. Table 4 presents the procedure and time-structure of the experiment in a more concise way.

Table 4. Procedure Experiment 1

	Minutes
Welcome	2
Demographic questionnaire	1
Training phase	20
Familiarization session (50 trials)	8
PPV questionnaire	3
Experimental session (100 trials)	15
FMV questionnaire	3
Debriefing	3
Payment	3
Good bye	2
Total	60

6.2. Hypotheses

Participants´ responding behavior

Hypothesis 1:
A main effect of PPV and of ALARM SYSTEM on participants' response rates to alerts (warnings and alarms together) is expected. On one side, participants will comply more with alerts emitted by high-PPV alarm systems than by low-PPV alarm systems. On the other side, participants will comply more with BAS than LAS.

This prediction about the main effect of PPV is based on previous studies showing that the alarm system´s PPV is positively correlated with participants´ compliance rates (Getty et al., 1995; Manzey et al., 2014).

The main effect of ALARM SYSTEM is based on findings from previous studies showing that operators tend to comply more with alerts emitted by BAS than LAS (Bustamante & Bliss, 2005; Wiczorek & Manzey, 2014; Wiczorek et al., 2014).

Hypothesis 2:
In the low-PPV condition, participants will respond to most LAS-alarms and will ignore most LAS-warnings and BAS-alarms.

Regarding the response strategies, participants will exhibit mostly positive extreme responding in response to LAS-alarms and mostly negative extreme responding in response to LAS-warnings. A mixed of negative extreme responding and probability matching is expected in response to BAS-alarms.

Hypothesis 3:
In the high-PPV condition participants will respond to most of LAS-alarms and BAS-alarms, and will ignore most of LAS-warnings (i.e., cry-wolf effect).

Regarding participants´ response strategies, participants will exhibit mostly positive extreme responding in response to LAS-alarms and BAS-alarms and mostly probability matching in response to LAS-warnings.

These hypotheses are based on previous studies showing that participants adapt their response rate to the PPV of alarm systems (Getty et al., 1995; Manzey et al., 2014) and that participants exhibit negative extreme responding in response to alerts having a very low PPV and positive extreme responding in response to alerts having a PPV above .7 (Bliss, 2003b; Manzey et al., 2014). Results of Wiczorek (2012) even suggest that in case of LAS a tendency for positive extreme responding already might be observed for a PPV above .6

Participants´ performance

Hypotheses about participants' performance are the direct consequence of Hypotheses 1, 2, and 3 about participants´ responding behavior.

Hypothesis 4:
An interaction effect between ALARM SYSTEM and PPV on participants´ percentage of hits and misses is expected. In the low-PPV condition LAS users will have better performance than BAS users. The opposite pattern is expected in the high-PPV condition: BAS users will have better performance than LAS users.

In the low-PPV condition it is expected that participants will ignore most the alarms emitted by the BAS and therefore produce more misses and less hits than LAS users who only ignore warnings. In the high-PPV condition it is however expected that BAS users will comply with almost all alarms and therefore produce a very low percentage of misses. However, LAS users will produce some misses by ignoring some of the emitted warnings. Consequently the overall percentage of misses will be bigger for LAS users than BAS users in the high-PPV condition.

Hypothesis 5:

A main effect of ALARM SYSTEM and of PPV on participants' percentage of false alarms and correct rejections is expected. On the one hand, LAS users will have better performance than BAS users. On the other hand, participants interacting with low-PPV alarm systems will have better performance than participants interacting with high-PPV alarm systems.

It is expected that LAS users will have a very accurate responding behavior thanks to the likelihood cues provided by the alarm system. As a consequence, LAS users will show a lower percentage of false alarms and a higher percentage of correct rejections than BAS users. Moreover, participants in the low-PPV condition will tend to ignore more alerts than in the high-PPV condition and thus inevitably produce fewer false alarms and more correct rejections.

Hypothesis 6:

A main effect of PPV and ALARM SYSTEM on participants' performance in the concurrent tasks is expected. Participants in the low-PPV condition will have higher performance in the concurrent tasks than participants in the high-PPV condition. Moreover, LAS users will have better performance than BAS users.

Hypothesis 6 is a direct consequence of Hypothesis 1. The less often participants comply with alerts, the more time they have to work on the concurrent tasks.

Participants´ trust

Hypothesis 7:

A main effect of ALARM SYSTEM and PPV on participants´ trust ratings of the alarm system is expected. Participants will trust LAS more than BAS and will trust high-PPV alarm systems more than low-PPV alarm systems.

*This hypothesis is based on previous studies showing that the reliability of an alarm system has an effect on trust (Lee & Moray, 1992; Lee & See, 2004; Wiczorek, 2011) and that participants trust LAS more than BAS (*Ragsdale, 2012; Wiczorek & Manzey, 2014).

6.3. Results

All participants fully completed the experiment. An alpha level of .05 was used for all statistical tests. The mean (*M*) and standard deviations (*SD*) of all analyses are reported in Annex A.

6.3.1. Manipulation check

Real and estimated PPV and NPV values are displayed in Table 5.

Table 5. Real PPV and estimated PPV of the alarm, warning, and non-alert stage of the BAS and LAS depending on the PPV (low vs. high)

Alarm system	Real PPV	Estimated PPV	Real NPV	Estimated NPV
BAS 35				
alarm stage	.35	.46		
non-alert stage			.97	.90
BAS 65				
alarm stage	.65	.60		
non-alert stage			.96	.87
LAS 35				
alarm stage	.78	.64		
warning stage	.21	.34		
non-alert stage			.97	.90
LAS 65				
alarm stage	.91	.70		
warning stage	.47	.46		
non-alert stage			.96	.92

Six t-tests were conducted to compare the real PPV value of each stage of the alarm systems with the PPV value estimated by participants. These analyses allow for a better understanding of results that might differ from our predictions.

Results show that participants overestimated the PPV of the low-PPV BAS, $t(12) = -3.31$, $p = .01$. They reported a PPV of .46 instead of .35. Participants also underestimated the PPV of the high-PPV LAS, $t(12) = 2.17$, $p = .05$. They reported a PPV of .70 instead of .91.

Such deviations with the real values are actually not problematic because participants still perceived the high-PPV BAS as more reliable than the low-PPV BAS, $t(24) = 2,686$, $p = .01$, and that they also perceived the alarm stage of LAS more reliable than the warning stage of LAS, both in the low-PPV condition and in the high-PPV condition.

Four t-tests were conducted to compare the real NPV value to the NPV value reported by participants in each condition. The results show that participants significantly underestimated the NPV of all alarm systems ($ps > .05$).

6.3.2. Alarm Task: Participants´ response rates

We measured participants´ response rate to alerts (alarms and warnings together), alarms only, and warnings only.

Mean response rates to alerts (alarm stage and warning stage together)

We used a two-way ANOVA with the factors ALARM SYSTEM and PPV for the analysis of the participants´ response rates to alerts. The means of participants´ response rates to alerts (alarm and warning stage together) are displayed in Figure 8.

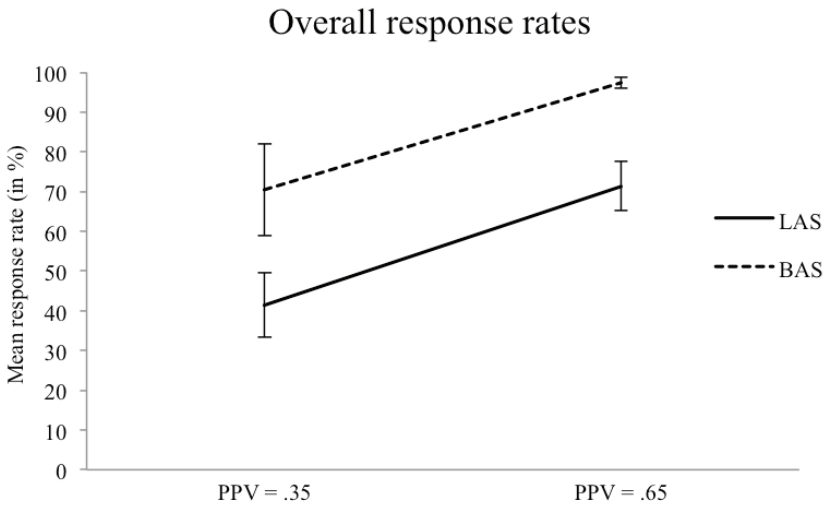


Figure 8. Means and mean standard deviations of participants´ compliance rates with alerts depending on the type of alarm system and the PPV

As it was expected Figure 8 shows that participants complied more with alerts emitted by the high-PPV alarm systems *(M* = 84.32%) than by the low-PPV alarm system *(M* = 55.91%). The main effect of PPV is significant, $F(1,49) = 13.48$, $p = .001$. The main effect of ALARM SYSTEM is also significant, $F(1,49) = 12.64$, $p = .001$. As expected participants complied more with BAS alerts *(M* = 83.87%) than LAS alerts *(M* = 56.36%). This suggests that the use of LAS lead to a greater cry-wolf effect than the use of BAS. These results are going to be discussed more deeply in the Discussion section. No significant interaction effect ALARM SYSTEM x PPV was found, $F(1,49)=0.03$, $p = .85$.

Mean response rates to alarms

A one-way ANOVA with CONDITION (BAS35, BAS65, LAS35, LAS65) as between factor was used for the analysis of the participants´ response rate to alarms. Participants' mean response rates are displayed in Figure 9. The results were as expected. Participants complied with alarms at almost all times, except in the condition BAS35, in which they complied at a significantly lower rate. Here one can already notice that participants complied at a surprisingly high rate with alarms emitted by the BAS35 (*M* = 70.40%). The one-way ANOVA conducted confirms this pattern. Three orthogonal contrasts C1 (3, -1, -1, -1), C2 (0,

2, -1, -1), and C3 (0, 0, 1, -1) were used. The first contrast compares the condition BAS35 with the other conditions and confirms that participants complied less with alarms emitted by the BAS35 than with alarms emitted by the other alarm systems, $F(1,49) = 13.76$, $p =$. 001. The second contrast compares the BAS65 condition to the LAS35 and LAS65 condition while the third contrast compares the LAS35 condition to the LAS65 condition. This shows that participants´ compliance rates with alarms of BAS65, LAS35 and LAS65 do not differ from each other, C2: $F(1,49) = 0.02$, $p = .88$ and C3: $F(1,49) = 0.02$, $p = .89$.

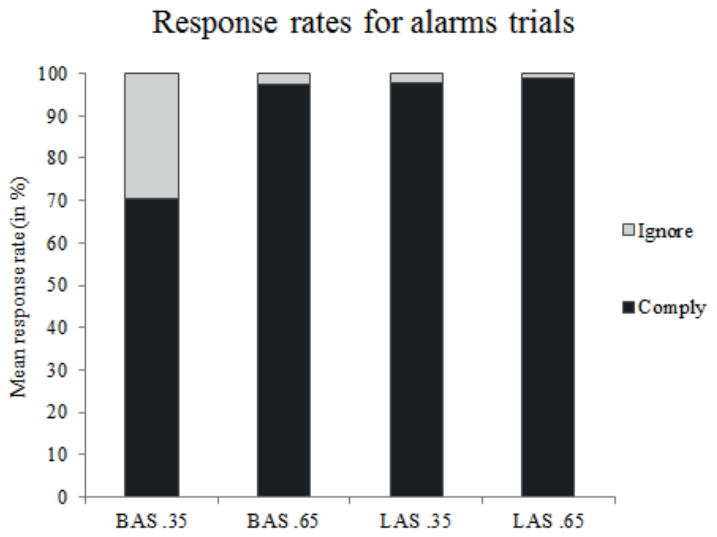


Figure 9. Means of participants´ compliance rates and ignoring rates of alarms depending on the experimental conditions

Mean response rates to warnings

A t-test with the factor PPV was used for the analysis of participants´ response rate to warnings. As one can see on Figure 10, mean response rates to warnings of both conditions are low as expected (i.e., evidence of some cry-wolf-effect). The main effect of PPV is significant, $F(1,23) = 4.00$, $p = .05$. Participants responded less to warnings in the low-PPV condition ($M = 20.31\%$) than in the high-PPV condition ($M = 51.11\%$).

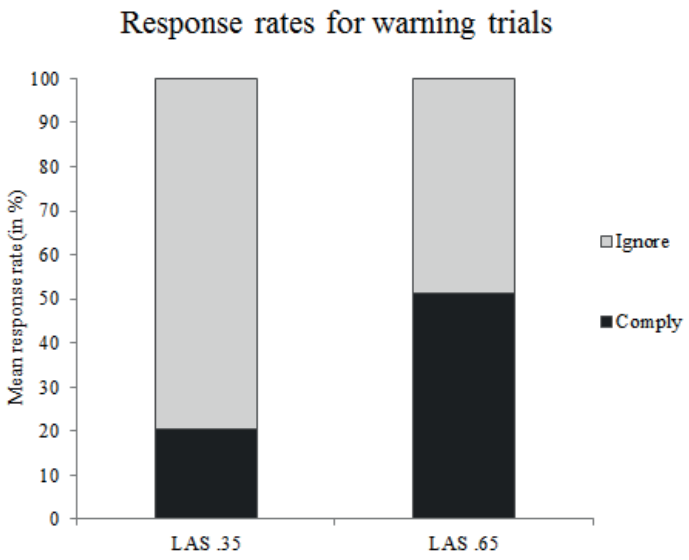


Figure 10. Means of participants´ compliance rates and ignoring rates of warnings depending on the PPV

Mean response rates to non-alerts

Participants ignored almost all non-alerts (M = 98.48%, all conditions together). There is no main effect of PPV, of ALARM SYSTEM, and no interaction effect ALARM SYSTEM x PPV on participants´ response rates to non-alerts, $F(3,46) = 0.64$, $p = .5$.

Participants' behavioral differentiation

Participants complied more with alarms emitted by LAS35 than BAS35, $F(1,22) = 5.05$, $p = .03$ while, as expected, no difference was observed between BAS65 and LAS65 (see above: Mean response rates to alarms).

Participants complied more with alarms than with warnings emitted by the LAS35, $F(1,10) = 51.14$, $p = .00$. They also complied more with alarms than with warnings emitted by the LAS65, $F(1,11) = 20.35$, $p = .00$. This suggests that participants adapted their behavior to the PPV of each stage of the LAS.

6.3.3. Participants´ response strategies

Response strategies for alarms

The different response strategies exhibited by participants in response to alarms are displayed in Figure 11.

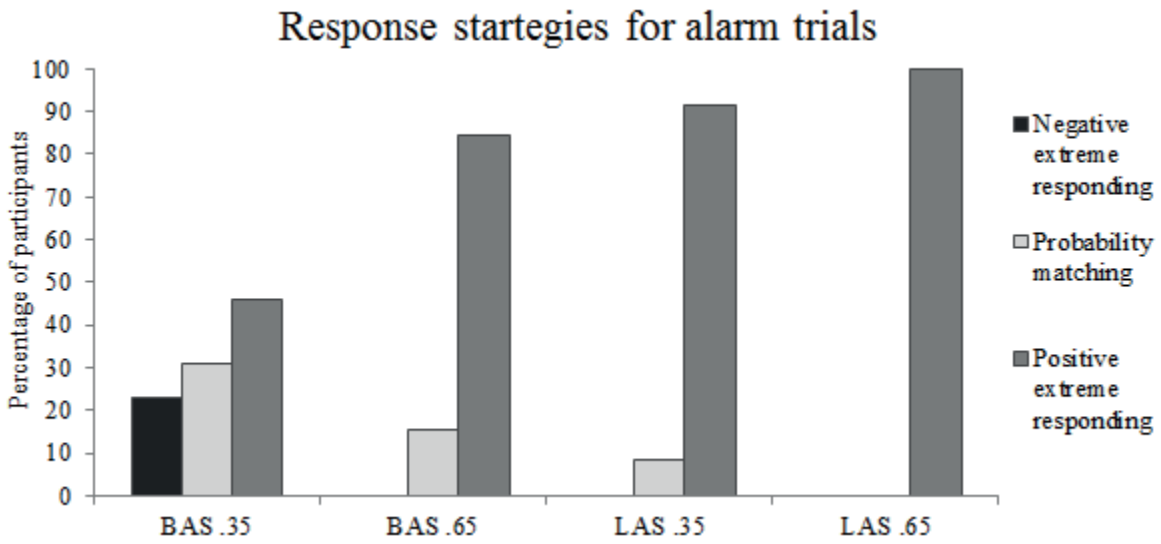


Figure 11. Percentage of participants exhibiting negative extreme responding, positive extreme responding, and mixed strategies in response to alarms as defined in Part 3.3.2.

In order to make comparisons between the four conditions easier, these results are reported as percentages in the Figure 11. The appropriate statistical analysis (Chi2) could not be used to analyze these results because the sample size assumption was not met. Since $n < 15$ in each experimental group, the expected frequencies for each cell is lower than 5. The analysis of these results is descriptive.

Response strategies adopted by participants differ depending on the condition. The dominant strategy adopted by participants was clearly positive extreme responding for alarms emitted by BAS65 (83.33% of participants), LAS35 (91.67%), and LAS65 (100%).

Regarding the BAS35 condition, a mix of different response strategies was adopted by participants. Only a minority of participants exhibited negative extreme responding (21%).

Half of them (50%) exhibited positive extreme responding and 29% showed some evidence of probability matching.

Response strategies for warnings

The different response strategies exhibited by participants in response to alarms are displayed in Figure 12.

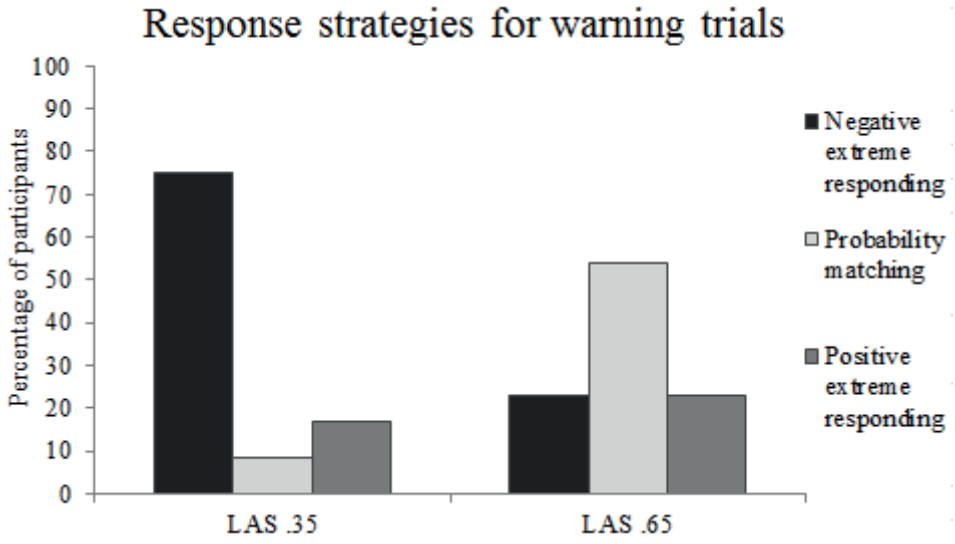


Figure 12. Percentage of participants exhibiting negative extreme responding, positive extreme responding and probability matching in response to warnings as defined in Part 3.3.2.

The dominant strategy adopted by participants in response to LAS35-warnings was negative extreme responding (75%). Concerning participants' behavior towards warnings emitted by LAS65, most participants adopted a sort of probability matching behavior (53.85%), 23.08% of participants adopted a negative extreme responding strategy, and 23.08% adopted a positive extreme responding strategy.

Response strategies for non-alerts

All participants except two (one in the BAS35 condition and one in the BAS65 condition) adopted a negative extreme response strategy for the non-alert stage.

6.3.4. Participants´ performance in the alarm task

All analyses about participants´ performance in the alarm task were performed using 2(ALARM SYSTEM)x2(PPV) between-groups ANOVAs.

Participants' performance in the alarm task was assessed through four dependent variables: The percentage of hits, of misses, of false alarms, and of correct rejections that participants made when the alarm system produced either a warning or an alarm. Participants' performances for the non-alert stage were not taken into account. Participants' performance are displayed in Figure 13 (percentage of hits and misses) and in Figure 14 (percentage of false alarms and correct rejections).

Participants' percentage of hits

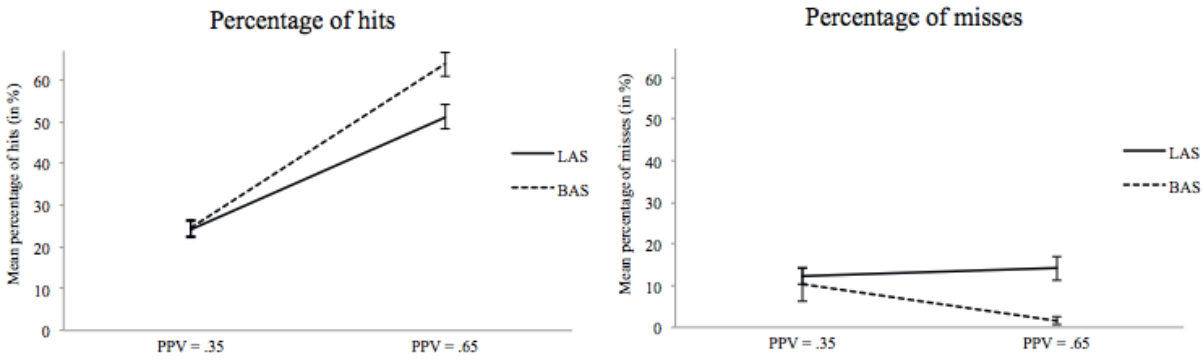


Figure 13. Mean and mean standard deviations of participants' percentage of hits (left panel) and misses (right panel) in the alarm task depending on the type of alarm system and the PPV

Regarding participants' percentage of hits a significant main effect of PPV was found as expected, $F(1,49) = 149.08$, $p = .00$, showing that participants´ percentage of hits was higher in the high-PPV conditions ($M = 57.50\%$) than in the low-PPV conditions ($M = 24.36\%$).

Surprisingly a significant main effect of ALARM SYSTEM was found, $F(1,49) = 5.82$, $p = .02$, showing that BAS users produced a higher percentage of hits ($M = 44.20\%$) than LAS users ($M = 38.19\%$). Actually, this benefit of BAS over LAS concerned participants in the high-PPV condition only. Participants in the BAS65 condition ($M = 63.81\%$) had better performance than participants in the LAS65 condition ($M = 51.18\%$), $F(1,49) = 11.04$, $p = .00$, but no difference was found between LAS35 and BAS35, $F(1,49) = 0.02$, $p = 90$. The interaction effect ALARM SYSTEM x PPV is significant, $F(1,49) = 5.01$, $p = .03$.

Participants' percentage of misses

Regarding participants' percentage of misses no main effect of PPV was found $F(1,49) = 1.54$, $p = .22$. A main effect of ALARM SYSTEM was found, $F(1,49) = 7.24$, $p = .01$. LAS users produced a higher percentage of misses ($M = 13.26\%$) than BAS users ($M = 5.92\%$). The expected interaction effect ALARM SYSTEM x PPV failed the conventional level of significance, $F(1,49) = 3.84$, $p = .056$. However, the analysis of the simple effects shows a statistically significant effect of ALARM SYSTEM in the high-PPV condition. BAS users ($M = 1.58\%$) had better performance (i.e., lower percentage of misses) than LAS users ($M = 14.20\%$) in the high-PPV condition, $F(1,49) = 5.35$, $p = .03$. However, no difference was found between LAS and BAS in the low-PPV condition, $F(1,49) = 0.26$, $p = .61$ whereas a benefit of LAS over BAS was expected.

Participants' percentage of false alarms

Figure 14 shows participants' percentage of false alarms.

Regarding participants' percentage of false alarms, as it was expected, BAS users produced a higher percentage of false alarms ($M = 39.67\%$) than LAS users ($M = 18.76\%$). The main effect of ALARM SYSTEM is significant, $F(1,49) = 16.45$, $p = .00$. The main effect of PPV is not significant, $F(1,49) = 0.83$, $p = .37$. Participants in the high-PPV condition did not produce a higher percentage of false alarms than participants in the low-PPV condition. Finally, the interaction effect ALARM SYSTEM x PPV on participants' percentage of false alarms is also not significant, $F(1,49) = 2.13$, $p = .15$.

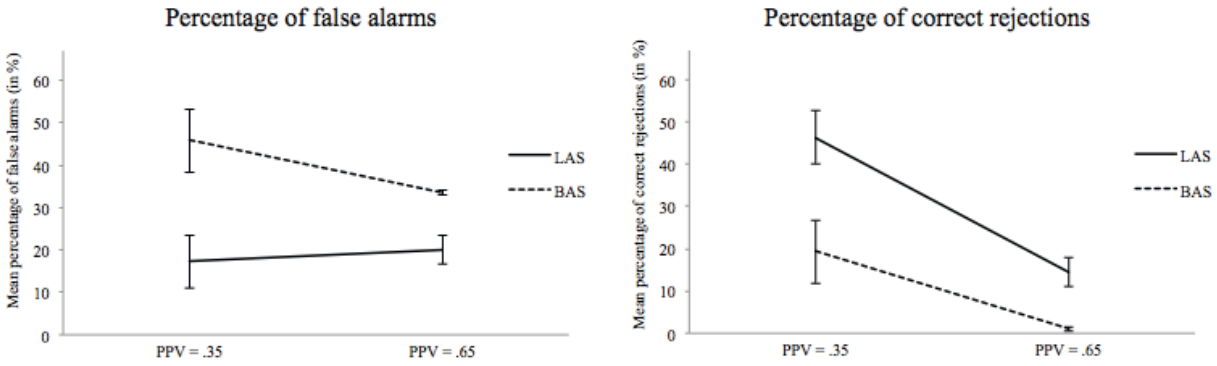


Figure 14. Mean and mean standard deviations of participants' percentage of false alarms (left panel) and correct rejections (right panel) in the alarm task depending on the type of alarm system and the PPV

Participants' percentage of correct rejections

Figure 14 shows participants' percentage of correct rejections. Regarding participants' percentage of correct rejections, as expected, LAS users *(M* = 29.77%) had better performance than BAS users *(M* = 10.21%). This main effect of ALARM SYSTEM is significant, $F(1,49) = 15.29$, $p = .00$. A main effect of PPV was also found, $F(1,49) = 23.51$, $p = .00$. Participants interacting with a low-PPV alarm system had better performance *(M* = 32.30%) than participants working with a high-PPV alarm system *(M* = 7.79%). There is no significant interaction effect ALARM SYSTEM x PPV on participants' percentage of correct rejections, $F(1,49) = 1.73$, $p = .20$.

6.3.5. Participants´ performance in concurrent tasks

All analyses about the participants´ performance in the concurrent tasks were performed using 2(ALARM SYSTEM)x2(PPV) between-groups ANOVAs.

Performance in the Coolant Exchange Task (CET)

As one can see on Figure 15, participants using low-PPV alarm systems completed more refilling cycles ($M = 14.96$) than participants using high-PPV alarm systems ($M = 13.50$). This main effect of PPV is significant, $F(1,49) = 6.40$, $p = .02$. The main effect of ALARM

SYSTEM and the interaction effect ALARM SYSTEM x PPV were not significant, $F(1,49) = 2.20$, $p = .15$, and $F(1,49) = 0.69$, $p = .41$ respectively.

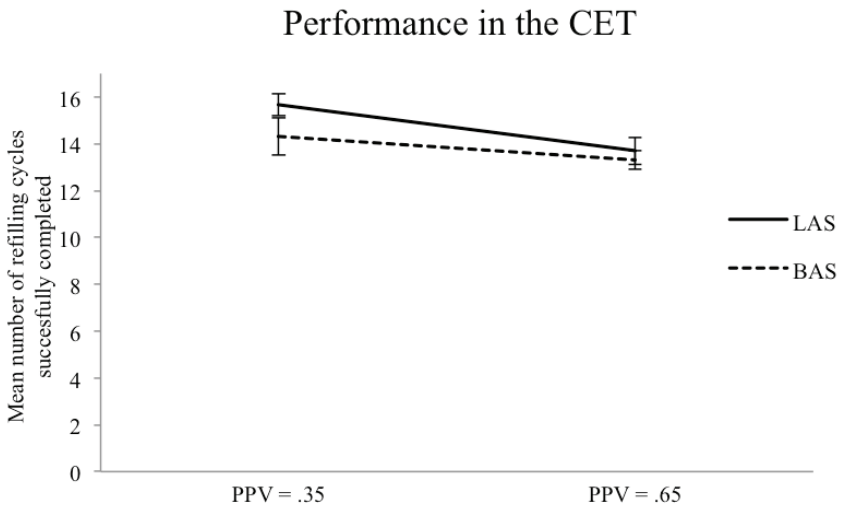


Figure 15. Means and mean standard deviations of the number of refilling cycles successfully completed depending on the type of alarm system and the PPV

Performances in the Resource Ordering Task (ROT)

Regarding the number of correctly sent orders by participants in the ROT results did not show any main effect of ALARM SYSTEM, $F(1,49) = 0.01$, $p = .93$ and of PPV, $F(1,49) = 1.03$, $p = .316$. Also, no interaction effect ALARM SYSTEM x PPV was found, $F(1,49) = 0.01$, $p = .92$.

Regarding the percentage of appeared orders participants responded to in the ROT, results did not show any main effect of ALARM SYSTEM, $F(1,49) = 1.07$, $p = .31$, and of PPV, $F(1,49) = 2.21$, $p = .14$. Furthermore, no interaction effect ALARM SYSTEM x PPV was found, $F(1,49) = 0.84$, $p = .36$.

6.3.6. Participants´ subjective ratings

Trust in the alarm system: FMV trust scale

A 2(ALARM SYSTEM)x2(PPV) between-groups ANOVA was performed on participants´ trust ratings. No significant main effect of ALARM SYSTEM, $F(1,49) = 0.03$, $p = .16$, of PPV, $F(1,49) = 0.29$, $p = .54$, and no significant interaction effect (ALARM SYSTEM x PPV), $F(1,49) = 0.75$, $p = .87$, were found.

A more precise analysis of the subscales of the FMV trust scale shows a significant main effect of ALARM SYSTEM on the subscale use, $F(1,49) = 3.94$, $p = .05$. Participants reported that the LAS ($M = 2.96$) was more useful than the BAS ($M = 2.60$).

6.4. Discussion

The main goal of Experiment 1 was to investigate if the base rate of critical events moderates the beneficial effect that LAS is meant to have over BAS according to literature. Because the manipulation of the base rate has a direct effect on the PPV of alarm systems, we defined our experimental conditions directly by their PPV. A benefit of LAS over BAS was expected in the low-PPV condition (i.e., low base rate) whereas a benefit of BAS over LAS was expected in the high-PPV condition (i.e., high base rate) in regard to the number of critical events correctly detected.

There were four experimental conditions in this experiment: low-PPV BAS (BAS35), high-PPV BAS (BAS65), low-PPV LAS (LAS35), and high-PPV LAS (LAS65). The research hypotheses reflect the various research questions in which we have a particular interest: how operators respond to alerts (H1, H2, and H3), how this responding behavior affect participants' performance both in the alarm task (H4, H5) and in the concurrent tasks (H6), and what happens at the subjective level by measuring trust (H7).

Before being able to analyze the results it was necessary to make sure that our manipulation of the PPV was successful. To this purpose we analyzed to what extent participants' perceived PPV differed from the real PPV values. The results show that the perceived and real PPV

values differ in the BAS35 condition (overestimation) and in the LAS65 condition (underestimation). A possible explanation for this misguided evaluation of BAS35 could be the human bias of overestimating the probability of rare events (Kahneman & Tvesky, 1984). This is not problematic regarding the experimental manipulation because participants still perceived BAS35-alarms as less reliable than BAS65-alarms (estimated PPV= .60), showing that the manipulation of the PPV was successful. Moreover participants perceived LAS-alarms as more reliable than LAS-warnings. This is mandatory since the hypotheses are based on the assumptions that LAS users will differentiate their behavior in a different way than BAS users.

Interestingly, participants underestimated the NPV of alarm systems. Again this might equally be explained by participants' bias to overestimate low probabilities (Kahneman & Tvesky, 1984). As such participants might have overestimated the occurrence of critical events in the non-alert stage. However, this had no effect on participants' responding behavior since 96% of them adopted an extreme negative responding strategy toward the non-alert stages.

The current experiment aimed to test nine hypotheses. The results concerning each hypothesis are presented and discussed in following.

In Hypothesis 1, a main effect of PPV on participants´ overall response rates was expected. It was expected that participants would comply more with alerts in the high-PPV than in the low-PPV condition. The results confirm this. This finding support results from previous studies showing that the PPV is a great predictor of participants´ response rate (Getty et al., 1995; Manzey, et al., 2014).

Hypothesis 1 also predicted that participants would comply more with BAS than LAS (main effect of ALARM SYSTEM). The findings confirm this hypothesis and therefore replicate results from Bustamante & Bliss (2005) and Wiczorek & Manzey (2014). At a first look such results seem counterintuitive. Since LAS is implemented in order to reduce the cry-wolf effect, one might expect that participants would comply more with LAS than BAS. Nevertheless, and as expected, the results show the opposite trend. Two reasons explain this finding. First, the cry-wolf effect does not completely disappear when interacting with LAS but is shifted to the warning stage. Second, LAS emits more warnings than alarms. In the low-PPV condition LAS emits 48 warnings and 18 alarms. In the high-PPV condition LAS emits

45 warnings and 33 alarms. Since participants ignore most of warnings and that the biggest proportion of alerts are warnings, a reduction of the overall compliance rate towards alerts can be observed. These results are particularly interesting in multitask environments like M-TOPS because the overall compliance rate has a direct effect on the amount of attentional resources and free time that participants can allocate to concurrent tasks.

Hypothesis 2 focused on participants´ responding behavior in the low-PPV condition only. It was expected that BAS users would ignore most alarms emitted by BAS while LAS users would differentiate their behavior: ignore warnings but comply with alarms. The results partly confirm these expectations. Participants responded as expected toward LAS. However participants´ compliance rate with BAS was surprisingly high. Participants complied with 70.4% of BAS-alarms. A greater cry-wolf effect was expected in this stage. Complying with a BAS having a PPV of .35 is not a rational strategy because participants will produce a lot of false alarms and in this experiment, false alarms have the same cost as misses (-2 points). Two reasons might explain the high compliance rates with BAS-alarms. On the one hand, participants might have interpreted misses (sending an improper container to clients) as being more consequential than false alarms (sending a container, which is okay, back to the chemical plant), because of the cover story. On the other hand, participants overestimated the PPV of BAS-alarms (.46 instead of the real value of .35).

A look at participants' response strategies confirms these surprising results. Half of the participants in the BAS35 condition exhibited positive extreme responding. Yet, a middle value of PPV such as .35 should, in theory, mostly lead to the adoption of a probability matching (Bliss, Gilson, et al., 1995). However, only 29% of participants exhibited probability matching toward BAS35.

Participants´ response strategies toward LAS confirm our expectations. A majority of participants exhibited positive extreme responding towards alarms and negative extreme responding towards warnings.

Hypothesis 3 focused on participants responding behavior in the high-PPV condition. Contrary to the .35 PPV condition, no cry-wolf effect was expected for BAS and BAS users were expected to mostly comply with alarms. Regarding LAS users, they were expected to comply with most alarms while some evidence of cry-wolf effect would be observed in the

warning stage; however, to a lesser extent than in the low-PPV condition. The experiment's results confirm Hypothesis 3. It is interesting to note here that BAS65 users complied with 97.34% of emitted alarms.

Hypothesis 4 and 5 were about participants' performance in the alarm task. An interaction effect ALARM SYSTEM x PPV was expected to affect participants' percentage of hits and misses (H4). More precisely, a benefit of LAS over BAS was expected in the low-PPV condition whereas a benefit of BAS over LAS was expected in the high-PPV condition. The findings confirm these predictions only partially. A benefit of BAS over LAS emerged in the high-PPV condition, however, and contrary to our hypothesis, no difference between LAS and BAS emerged in the low-PPV condition. The latter result can be explained by participants' high compliance rates towards BAS. By complying with most of alarms, BAS35 users detected a high numbers of critical events; indeed as many as LAS35 users that benefited from the likelihood information.

The absence of benefit of LAS35 over BAS35 contrasts with results from previous studies (Bustamante & Bliss, 2005; Bustamante, 2008; Clark & Bustamante, 2008; Clark, Ingebritsen & Bustamante, 2010; Wiczorek & Manzey, 2014; Wiczorek et al., 2014), who reported a benefit of LAS over a low-PPV BAS in the alarm task. A direct comparison between our results and the results reported by Wiczorek & Manzey (2014) and Wiczorek et al. (2014) is particularly interesting since the same task environment M-TOPS was used. However, one difference between the studies was how performance was operationalized. Wiczorek & Manzey (2014) and Wiczorek et al. (2014) defined participants' performance as the sum of wrong decisions participants made in interaction with the alarm system (false alarms and misses together). If the same definition is applied to the results of Experiment 1 one arrives at the same conclusions as Wiczorek & Manzey (2014) and Wiczorek et al. (2014): participants have better performance with LAS than with low-PPV BAS. Indeed, a two-way ANOVA with the factors ALARM SYSTEM and PPV for the analyses of the sum of wrong decisions show a significant interaction effect of ALARM SYSTEM x PPV, $F(1,49) = 17.99$, $p = .00$. A simple effect of ALARM SYSTEM in the low-PPV condition on the sum of wrong decisions emerged, $F(1,49) = 22.79$, $p = .00$. Participants made less wrong decisions with LAS ($M = -41.33$) than BAS ($M = -77.85$). This is the consequence of the lower percentage of false alarms in LAS35 than in BAS35 since there is no difference between LAS35 and BAS35 on

the percentage of misses produced by participants. Interestingly, no simple effect of ALARM SYSTEM in the high-PPV condition emerged, $F(1,49) = 0.48$, $p = .49$.

The reasons why Bustamante was able to demonstrate a benefit of LAS over low-PPV BAS, whereas Experiment 1 failed to demonstrate this benefit, remain unclear, yet might be explained by the differences in the experimental paradigm used.

The unexpected absence of difference between BAS and LAS in the low-PPV condition is probably also responsible for the significant main effect of ALARM SYSTEM found on participants' percentage of hits and misses. In the high-PPV condition and in the low-PPV condition a benefit of BAS over LAS has been found (even though this benefit is only significant in the high-PPV condition).

As mentioned previously, a statistically significant benefit of BAS over LAS emerged in the high-PPV condition and is reflected in the mean percentages of hits and misses. These results are the direct consequence from participants´ responding behavior. Since participants almost systematically initiated reparative actions when an alarm of BAS65 went off, almost no improper container left the plant (1.58% of misses). Yet, LAS users ignored about 50% of the warnings even though 47% of these warnings were correct (PPV = .47). Consequently, LAS users inevitably ignored some true warnings and produced misses. This is not the first time that a benefit of BAS over LAS has been found. Zirk (2012) reported similar results in a study, which was initially not designed to investigate this difference.

Regarding participants' percentage of false alarms and correct rejections (H5), a main effect of ALARM SYSTEM and a main effect of PPV were expected. On one hand, it was expected that participants would demonstrate better performance with LAS than with BAS (i.e., produce less false alarms and more correct rejections). One the other hand, it was expected that participants would have better performance with the low-PPV alarm system than with the high-PPV alarm systems.

As expected, the results show that participants had better performance with LAS than BAS. The reason explaining the benefit of LAS over BAS is that LAS allows a more accurate responding behavior than BAS. By complying with most alarms, even though some of them were false, BAS users produced a lot of false alarms. LAS users ignore most of alerts likely to

be false (warnings) and produce therefore fewer false alarms than BAS users. In conclusion, even though LAS does not always help participants to detect more critical events, it can still help them to have a more accurate behavior and to produce fewer false alarms.

A main effect of PPV has only been found on participants' percentage of correct rejections. Participants have better performance in the low-PPV condition than in the high-PPV condition. Similar results were expected for participants' percentage of false alarms. They, however, did not emerge. Actually, BAS35 users produced a high amount of false alarms because they complied with most of alarms even though they were false. This had an impact on the overall percentage of false alarms in the low-PPV condition and explains why no difference between the low-PPV and the high-PPV condition was found.

In Hypothesis 6, a main effect of PPV and ALARM SYSTEM on participants' performances in concurrent tasks was expected. The findings only confirm this prediction for the main effect of PPV. Participants in the low-PPV condition had better performance than participants in the high-PPV condition. Yet, this effect only emerged in the Coolant Exchange Task. Two reasons can explain these results. First, participants in the low-PPV condition experienced 34 non-alerts while participants in the high-PPV condition experienced only 22 non-alerts. Consequently, and since participants mostly ignore non-alerts, participants in the low-PPV condition had more free time to allocate to the concurrent tasks. Second, participants in the low-PPV condition complied less with alerts than participants in the high-PPV condition. Hence, they again had more free time to work on concurrent tasks. Such results mean that even though high-PPV alarm systems improve performance in the alarm task, they in parallel decrease performance in the concurrent tasks.

Concerning the expected main effect of ALARM SYSTEM on participants' performance in concurrent tasks, no difference between BAS and LAS emerged. This confirms results from Wiczorek et al. (2014) but at the same time contrasts with the results from Wiczorek and Manzey (2014).

In the experiment of Wiczorek et al. (2014), participants´ overall response rate toward alerts was the same whatever the kind of system used (BAS vs. LAS). This is therefore not surprising that no difference between BAS and LAS users has been observed on participants' performance in the concurrent tasks. However, this is surprising that no benefit of LAS over

BAS emerged in Experiment 1 since, similarly to the Wiczorek and Manzey's (2014) experiment, LAS users complied less with alerts than BAS users and had therefore more time to work on concurrent tasks.

One difference between the experimental paradigm used by Wiczorek and Manzey (2014) on one hand, and the experimental paradigm used in Experiment 1 on the other, is the number of concurrent tasks participants had to work on. Participants in the experiment of Wiczorek and Manzey (2014) had only one concurrent task to work compared to two concurrent tasks in our studies. The significant benefit of having more time might emerge only if participants have only one concurrent task to work on because the entire free resources are then concentrated on a same task instead of being divided on two different tasks.

Unfortunately other studies (Bustamante & Bliss, 2005; Bustamante & Clark, 2010; Clark & Bustamante, 2008) did not report participants' performance in the concurrent task.

In Hypothesis 7, a main effect of ALARM SYSTEM and PPV on participants' trust ratings was expected. The results however show that the PPV has no effect on participants´ trust ratings. This is surprising that participants interacting with high-PPV alarm systems did not report more trust in the system than participants working with low-PPV alarm systems. This might be caused by the wrong estimations that participants did of the PPV of the BAS35 and LAS65. This might have affected participants' trust ratings.

Nonetheless, LAS users reported more trust in the alarm system than BAS users. This effect emerged on the subscale use only, which makes sense since LAS provides more detailed information than BAS. It is however surprising that no effect emerged on the subscale *reliability* since this finding has been reported by two earlier studies (Ragsdale, 2012; Wiczorek & Manzey, 2014). These two studies did not use the FMV trust questionnaire to measure trust in the alarm system. Thus, one could wonder if the FMV trust questionnaire was the most appropriate tool to assess trust differences between the conditions in Experiment 1. The four-point scale used in the questionnaire might be responsible for a low discriminative power and a scale with more response categories might have been more appropriate. Preston & Colman (1999) have for example shown that a four-point scale performs relatively poorly in comparison to a five-, six-, or seven-point scale to discriminate participants' responses.

Another measure of trust might have a greater discriminative power and thus allow differences between the experiment's conditions to appear. A visual analog scale (a single line containing only five verbal anchors ranging from "my trust is very strong" to "I barely trust the system" and no gradations), like the one used by Wiczorek & Manzey (2014), might be a better alternative in this regard.

This first experiment was successful in identifying a moderator (the PPV) on the potential benefit of LAS over BAS. The results suggest that high-PPV BAS are better than LAS to support operators in detecting critical events. However, LAS improves performance with regard to false alarms in comparison to low-PPV BAS. Not to forget, these difference in alarm systems' PPV correspond actually to difference in the base rate of critical events. Therefore, these results suggest that LAS does not improve detection performance (hits) in high base rate environments. The findings also support the claim made by Parasuraman et al. (1997) that the same alarm system can affect operators´ responding behavior differently depending on the base rate of critical events. Though having the exact same technical characteristics (i.e., criterion), alarm systems in Experiment 1 still had different effects on participants' responding behavior depending on the base rate of critical events.

These results have important practical applications. There are various environments in which the base rate of critical events can vary over time. For example, the density of traffic can change the base rate of critical events of environments in which warning collision systems are implemented (Parasuraman et al., 1997). Operators' behavior can also change the base rate of critical events (Meyer & Bitan, 2002). A nurse who demonstrates health preventive behaviors towards patients might avoid the occurrence of critical events. As a consequence, the preventive nurse will experience a lower base rate of critical events in comparison to a nurse who does not initiate such preventive behaviors. A practical solution might be the implementation of adaptive systems that modify its characteristics depending on the base rate of critical events. This solution is further detailed in Part 9.3.

7. Experiment 2: The influence of the cost of consulting AVI

The main objective of Experiment 2 is to investigate if the cost of consulting Alarm Validity Information (AVI), in terms of time and amount of effort spent, impacts the potential beneficial effect of LAS over BAS.

Previous studies conducted with the paradigm M-TOPS have shown that the beneficial effect of LAS over low-PPV BAS disappears when alarm validity information is provided to participants (Wiczorek & Manzey, 2014). When AVI is provided to participants a ceiling effect is observed and both BAS and LAS users reach almost optimal performance in the alarm task. Participants produce almost no incorrect responses because (a) BAS users cross-checked almost all alarms and (b) LAS users complied with most alarms and cross-checked most warnings. Wiczorek and Manzey (2014) concluded that if AVI is available implementing LAS does not improve performance in the alarm task over BAS.

These results contrast with results from previous studies showing a benefit of LAS over BAS in settings that provide AVI to participants (Bustamante & Bliss, 2005; Clark & Bustamante, 2008; Bustamante & Clark, 2010). Unfortunately Clark and Bustamante (2008) and Bustamante and Clark (2010) did not report results about participants' behavior. Only Bustamante and Bliss (2005) reported participants' overall response rates. Their results show that participants cross-checked up to 60% of the emitted alerts. This means that LAS users had an advantage over BAS users for at least 40% of alerts. For 40% of alerts, LAS users benefited from likelihood information helping them to determine if the alert was more likely to be true or false in comparison to BAS users that did not benefit from likelihood information. This might explain why a benefit of LAS over BAS has been found by Bustamante and Bliss (2005). It might also explain the similar results reported by Clark and Bustamante (2008) and Bustamante and Clark (2010) since the same alarm systems and the same experimental paradigm was used in these studies.

One difference between all of the above mentioned experiments is the cost of consulting AVI. In the experiment of Wiczorek and Manzey (2014) consulting AVI was made very easy.

Participants just had to click a button to access AVI. This required little time and effort. Moreover, participants did not lose any points by consulting AVI. In contrary, consulting AVI required more effort in the experiments of Bustamante and Bliss (2005), Clark and Bustamante (2008), and Bustamante and Clark (2010). In these experiments the concurrent task and the alarm task were displayed on different screens. In order to consult AVI participants first had to turn by 90° to see the second computer screen and then press the space bar on the keyboard located in front of the computer. In addition, participants lost points each time they cross-checked an alert that was correct. Such pay-off might have particularly discouraged participants to consult low-PPV warnings because it would have made them loose points almost systematically. This might explain why participants in the experiment of Bustamante and Bliss (2005) cross-checked less alerts than in the experiment of Wiczorek and Manzey (2014) as well as the difference observed in performance.

However, this interpretation regarding the cost of consulting AVI remains speculative because so far no study has investigated to what extent the cost of consulting AVI affects the impact of LAS vs. BAS on performance. Only one study conducted with BAS shows that participant decrease their cross-checking frequencies of alarms when the cost of consulting AVI increases (Manzey et al., 2014). This study did not report any effect of the cost of consulting AVI on participants' performance.

Experiment 2 aims to investigate the effect that the cost of consulting AVI has on BAS vs. LAS users' performance. We expected that BAS users and LAS users would reduce their cross-checking behavior when the cost of consulting AVI increases. Contrary to BAS, LAS allows a more precise reduction of cross-checking. For example, LAS users could decide to reduce their use of AVI only in the case that they get an alarm and still cross-check warnings. Thus, they would have little chance to miss a critical event. Consequently, increasing the cost of consulting AVI might only slightly decrease performance of LAS users. BAS users, on the other hand, do not benefit from any additional likelihood information to decide which alarms do not need to be checked. Subsequently, increasing the cost of AVI might strongly decrease performance of BAS users in the alarm task.

The same BAS and LAS than in the low-PPV condition of Experiment 1 were used for Experiment 2. The dependent variables are the same as in Experiment 1 with the exception

that an additional measure of trust and of participants´ workload was added. The latter measure was used to check if the experimental manipulation of the cost of AVI worked.

7.1. Method

7.1.1. Participants

Sixty-one participants (32 men, 29 women) participated in Experiment 2. Participants' age ranged from 18 to 37 with a mean age of 26.08 (*SD* = 3.98). Among them 89.8% were students, 8.5% were employed, and 1.7% had another status. They registered voluntarily to the experiment after having been recruited via an online portal for experiments (Prometei) of the Technische Universität Berlin. None of them was suffering from any distortion of vision, such as red-green color blindness, which might have interfered with the experiment.

Participants were paid €5 for their participation and, in addition, they had the chance to gain up to €4 depending on their performance in the experiment. Each participant was randomly assigned to one of the four experimental conditions. An equal repartition of men and women in each group was observed.

7.1.2. Task

Experimental paradigm M-TOPS

Participants completed the three tasks from the experimental paradigm M-TOPS (described in Part 6.1.2.). However, the main difference with Experiment 1 is the provision of Alarm Validity Information (AVI) to participants. This means that participants have the opportunity to cross-check the validity of the output of the alarm system before responding. In this version, a 'check' button – prüfen – is added above the 'repair' button – bearbeiten –. The interface M-TOPS with AVI is presented in Figure 16.

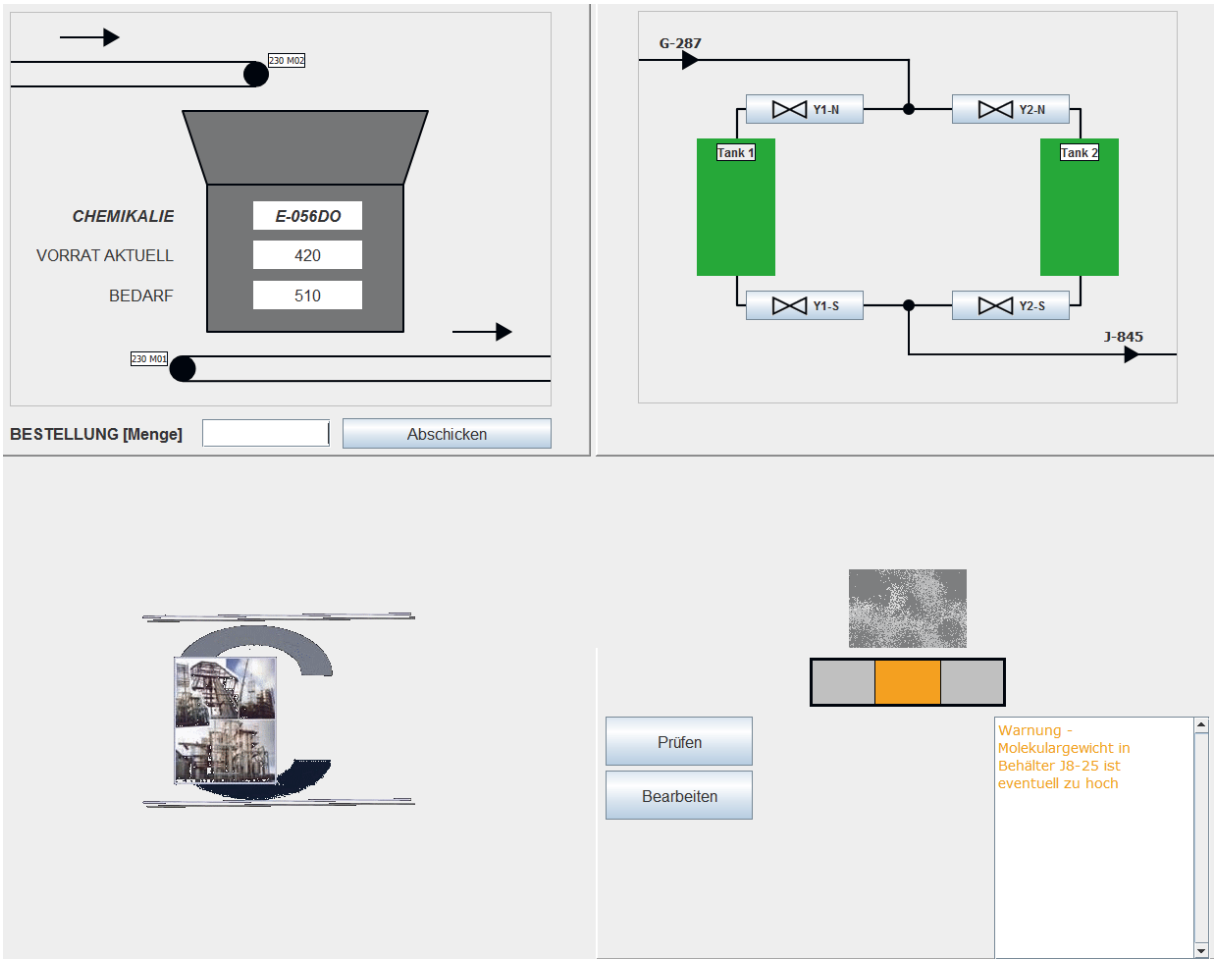


Figure 16. User interface of M-TOPS with access to AVI

In order to consult AVI participants can click on the 'check' button and a picture appears where the grey square is displayed (see Figure 17). This picture displays 40 letters. Participants are instructed that the presence of the letter K means that the chemical content of the container is improper, i.e., that the molecular weight in this container is too high. The AVI is 100% reliable. Participants have as much time as they need to search the letter K on the picture. While consulting AVI any actions on the concurrent tasks are not permitted. This is to avoid that participants click on the repair button merely to get more time to invest in the concurrent tasks. Once participants have finished scanning the picture they can either click on the 'repair' button – bearbeiten – to send the container back to the plant or on the 'continue' button – weiter – to send the container to clients.

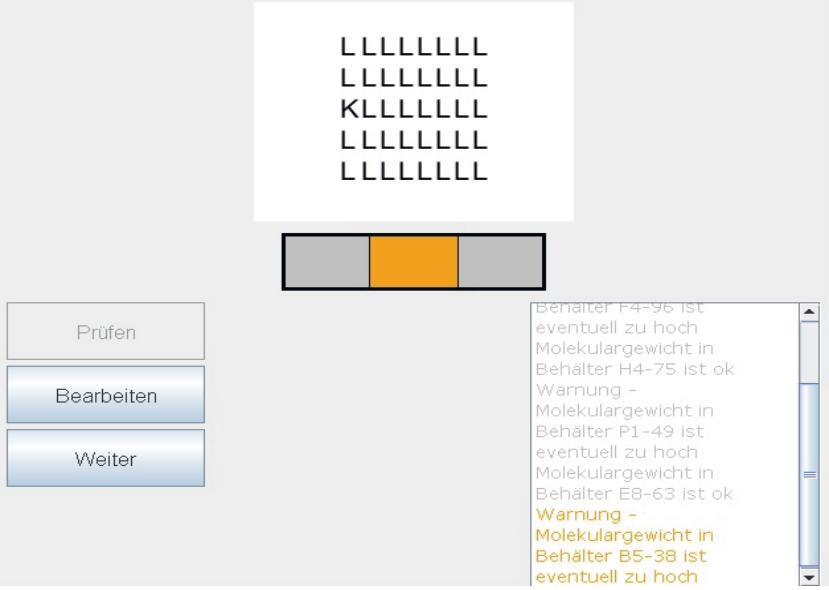


Figure 17. The alarm task after participants clicked the 'check' button

The experimental session stops after 100 trials, which corresponds to approximately 20 minutes, depending on how often participants consulted AVI. However, participants were only informed that the experimental session would stop after exactly 20 minutes. This was for participants to believe that there is a real trade-off between the time spent consulting AVI and the time spent working on the concurrent tasks. This makes our experimental settings closer to real multitask settings in which consulting AVI has a real cost regarding the amount of time participants can invest in the concurrent tasks.

Alarm system

The alarms systems used had the same reliability characteristics than in the .35 PPV condition of Experiment 1. Table 6 illustrates the alarm systems' main characteristics.

Table 6. Characteristics of BAS and LAS in Experiment 2

Type of alarm system	Base rate of critical events	Characteristics of the alarm system					
		d'	*c*	Global PPV	PPV alarm stage	PPV warning stage	NPV
BAS	.24	1.8	-1.05	.35	.35	-	.97
LAS	.24	1.8	-1.05 and 0.7	.35	.78	.21	.97

Pay-off system

The pay-off system was the same as in Experiment 1. It is described in Part 6.1.2.

7.1.3. Design

A 2 x 2 between-groups design was used for this study. The type of alarm system (BAS vs. LAS) and the cost of consulting AVI (low-cost vs. high-cost) were manipulated.

The cost of consulting AVI was manipulated (a) by manipulating the waiting time before participants could initiate an action (repair vs. next), and (b) by manipulating the difficulty of the visual search task required when consulting AVI. Participants in the high-cost condition can only initiate an action five seconds after accessing AVI whereas participants in the low-cost condition do not have any such waiting period. Moreover, the nature of the distractors used in the visual search task is different depending on the condition. Participants in the low-cost condition have to search for the letter K among the letters L while participants in the high-cost condition have to search for the letter K among various differing letters. As a consequence, the letter K is more salient in the low-cost condition (i.e., pop-out effect) making the visual search clearly easier and less demanding than in the high-cost condition. The difference between AVI in the low-cost and in the high-cost condition is presented in Figure 18.

Figure 18. AVI in the low-cost condition (left) vs. high-cost condition (right)

7.1.4. Dependent variables

Behavioral data: participants' response rates

Like in Experiment 1, we recorded participants´ response rates to the different stages and to the overall alarm system (non-alerts excluded). These included:

- Compliance rate with alerts;
- Cross-checking rate of alerts;
- Ignoring rate of alerts;

- Compliance rate with alarms;
- Cross-checking rate of alarms;
- Ignoring rate of alarms;

- Compliance rate with warnings;
- Cross-checking rate of warnings;
- Ignoring rate of warnings;

- Compliance rate with non-alerts;
- Cross-checking rate of non-alerts;
- Ignoring rate of non-alerts.

As in Experiment 1, even though participants' response rate to non-alerts was recorded, no effect on this variable was expected. Since the NPV of all alarm systems was very high and was kept constant across all conditions, it was expected that participants would always ignore non-alerts.

Participants´ response strategies were not analyzed in this experiment. Participants adopt such strategies in order to optimize performance of decisions taken under uncertainty. Because participants have the possibility to cross-check the output of the alarm system in this experiment, it does make little sense to analyze participants´ response strategies.

Performance data

As in Experiment 1, four dependent variables were used to record participants' performance. These are:

- Percentage of hits produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of misses produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of false alarms produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of correct rejections produced by participants when the alarm system emitted an alarm or a warning.

Three dependent variables were used to assess participants´ performance in the concurrent tasks:

- Number of correctly sent orders in the Resource Ordering Task (ROT);
- Proportion of orders participants responded to in the Resource Ordering Task (ROT);
- Number of refilling cycles successfully completed in the Coolant Exchange Task (CET).

Subjective data

Participants' subjective experience was measured using three dependent variables:

- Trust in the alarm system (FMV questionnaire, Wiczorek 2011);
- Trust in the alarm system (item);

Participants indicated how much they trust the entire alarm system on a scale from 0 to 10 (See Annex G). This item was included in order to have an additional measure of participants' trust other that the FMV questionnaire. Indeed, results from Experiment 1 raised some doubts about a potential lack of sensitivity of the FMV questionnaire for these investigations.

- Workload.

The workload is defined as the load of information placed on cognitive processes. Participants' workload was assessed using the NASA Task Load Index (Hart & Staveland, 1988; see Annex H). The mean of the six single scales was considered as overall workload measure.

7.1.5. Procedure

The procedure was the same as in Experiment 1. The trust item and the NASA-TLX questionnaire were administrated directly after the experimental session and before the FMV trust questionnaire. Table 7 resumes the procedure of Experiment 2.

Table 7. Procedure Experiment 2

	Minutes
Welcome	2
Demographic questionnaire	1
Training phase	20
Familiarization session (50 trials)	8
PPV questionnaire	3
Experimental session (100 trials)	15
NASA-TLX + Trust item + FMV questionnaire	6
Debriefing	2
Payment	2
Good bye	1
Total	60

7.2. Hypotheses

Participants´ responding behavior

Hypothesis 1:

A main effect of COST on participants' cross-checking rate of alerts is expected. Participants will consult AVI less often when the cost of it increases.

Hypothesis 2:

A main effect of COST on participants' compliance rate and cross-checking rate of alarms is expected. Participants will reduce their cross-checking rate of alarms in favor of compliance when the cost of AVI increases.

Hypothesis 3:

A differentiation in behavior is expected. On the one hand, participants will comply more with and cross-check less LAS-alarms than LAS-warnings (H3a). On the other hand,

participants will cross-check more BAS-alarms than LAS-alarms in the low-cost condition (H3b).

Hypothesis 3a is based on findings showing that participants tend to comply more with and cross-check less alarms having a high PPV than a low PPV (Manzey et al., 2014). Hypothesis 3b is based on results of Wiczorek and Manzey (2014) showing that participants crossed-check almost all BAS-alarms while they cross-check only a minority of LAS-alarms. This hypothesis concerns the low-PPV condition only. Based on results from previous studies, it is not possible to formulate any hypothesis about participants' cross-checking behavior in the high-cost condition. A difference between BAS users and LAS users might emerge, or not, in the high-cost condition, depending on the extent to which BAS users and LAS users reduce their cross-checking behavior.

Participants' performance

Hypothesis 4:

A main effect of COST on participants´ percentage of hits, misses, correct rejections, and false alarms is expected. Participants will show a higher percentage of hits and correct rejections and a lower percentage of misses and false alarms in the low-cost condition than in the high-cost condition.

Participants in the low-cost condition will have better performance in the alarm task than participants in the high-cost condition because they will cross-check more often the outputs of the alarm systems.

Hypothesis 5:

An interaction effect ALARM SYSTEM x COST on participants´ percentage of hits and misses is expected. More precisely, no difference between LAS and BAS is expected in the low-cost condition whereas a benefit of LAS over BAS is expected in the high-cost condition.

A ceiling effect is expected in the low-cost condition. Both LAS users and BAS users will have performance close to perfect. This is the direct consequence of (a) BAS users cross-checking the majority of alarms, and (b) LAS users cross-checking the majority of LAS-warnings and

either cross-checking or complying with LAS-alarms. However, it is expected that participants will reduce their cross-checking behavior in the high-cost condition. Contrary to BAS, LAS allows a more precise reduction of cross-checking. LAS users might cross-check less alarms and choose to rather comply with them. This would only slightly decrease their performance because LAS-alarms are very reliable. BAS users, on the other hand, do not benefit from any additional likelihood information to decide which alarms do not need to be cross-checked. Increasing the cost of AVI might strongly decrease performance of BAS users in the alarm task. As a consequence, a benefit of LAS over BAS will emerge in the high-cost condition.

Hypothesis 6:

An interaction effect ALARM SYSTEM x COST on participants´ percentage of false alarms and correct rejections is expected. A benefit of BAS over LAS is expected in the low-cost condition and a benefit of LAS over BAS is expected in the high-cost condition.

It is expected that BAS users will have performance close to optimal in the low-cost condition. Nevertheless, LAS users will produce some false alarms by complying with the few false alarms produced by the LAS-alarms. Overall, LAS users will produce a higher percentage of false alarms and a lower percentage of correct rejections than BAS users in the low-cost condition. This opposite pattern is expected in the high-cost condition. Based on results from Experiment 1 showing that participants comply with most of alarms emitted by a BAS having a PPV of .35, it is expected that BAS users will reduce their cross-checking behavior in favor of complying with BAS-alarms. Because the PPV of this alarm stage is low (.35) BAS users will necessarily produce some false alarms. As a consequence, BAS users will produce more false alarms and less correct rejections than LAS users, who benefit from access to likelihood information. A benefit of LAS over BAS is expected in the high-cost condition.

Hypothesis 7:

A main effect of COST on participants´ performances in the concurrent tasks is expected. Participants in the low-cost condition will have higher performance than participants in the high-cost condition.

Participants need more cognitive resources to analyze AVI in the high-cost condition than in the low-cost condition. It is therefore expected that participants in the high-cost condition will have less cognitive resources available to invest in the concurrent tasks in comparison to participants in the low-cost condition.

Participants' workload

Hypothesis 8:
A main effect of COST on participants' workload is expected. Participants in the high-cost condition will report more subjective workload than participants in the low-cost condition.

Participants need more cognitive resources to analyze AVI in the high-cost condition than in the low-cost condition.

Participants´ trust

Hypothesis 9:
A main effect of ALARM SYSTEM on participants´ trust ratings is expected. Participants will report having more trust in LAS than in BAS.

7.3. Results

Two participants failed to complete the experiment. They did not work on the Coolant Exchange Task during the experimental session so we excluded them from the analyses.

An alpha level of .05 was used for all statistical tests. The mean (*M*) and standard deviations (*SD*) of all analyses are reported in Annex B.

7.3.1. Manipulation check

Table 8 shows real and estimated PPV and NPV values.

Table 8. Real PPV and estimated PPV of the alarm, warning, and non-alert stage of the BAS and LAS depending on the cost of consulting AVI

Alarm system	Real PPV	Estimated PPV	Real NPV	Estimated NPV
BAS Low-cost				
alarm stage	.35	.59		
non-alert stage			.97	.74
BAS High-cost				
alarm stage	.35	.40		
non-alert stage			.97	.83
LAS Low-cost				
alarm stage	.78	.79		
warning stage	.21	.51		
non-alert stage			.97	.79
LAS High-cost				
alarm stage	.78	.75		
warning stage	.21	.38		
non-alert stage			.97	.86

Three t-tests were conducted to compare the real PPV value of each stage to the mean value of the PPV estimated by participants.

Results show that participants overestimated the PPV of the BAS in the low-cost condition only, $t(13) = 6.56$, $p = .00$. They reported a PPV of .59 instead of .35. However, they reported a correct estimation of the PPV of the BAS in the high-cost condition (estimated PPV = .40).

Results also show that participants overestimated the PPV of the warning stage of LAS (real PPV = .21), $t(30) = 7.45$, $p = .00$, both in the low-cost condition (estimated PPV = .51) and in the high-cost condition (estimated PPV = .38).

More importantly, participants perceived the PPV of LAS-alarms as more reliable than the PPV of BAS-alarms, $t(58) = 6.82$, $p = .00$, and LAS-warnings, $t(30) = 7.18$, $p = .00$.

Finally, four t-tests were conducted to compare the real NPV value to the NPV reported by participants in each condition. Participants underestimated the NPV of all alarm systems, ps < .05.

7.3.2. Alarm Task: Participants´ response rates

We measured participants' response rate to the overall LAS (alarm and warning stages together), to alarms, to warnings signals, and to non-alerts. Participants' response rates correspond to three dependent variables: participants' compliance rates, participants' cross-checking rates, and participants' ignoring rates. We used a two-way ANOVA with the factors ALARM SYSTEM and COST on each of these variables. In total nine two-way ANOVA were conducted. In addition, three t-tests were conducted for the analysis of participants' response rate to warnings.

Mean response rates to alerts (alarm stage and warning stage together)

The means of participants' compliance rate, cross-checking rate, and ignoring rate of alerts are displayed in Figure 19.

Participants' compliance rate with alerts

As illustrated in Figure 19, participants' compliance rate with alerts is higher in the high-cost condition *(M* = 16.46%) than in the low-cost condition *(M* = 5.25%). This main effect of COST is significant, $F(1,57) = 10.39$, $p = .00$. Moreover, participants complied more with alerts emitted by LAS *(M* = 14.42%) than by BAS *(M* = 6.71%). This main effect of ALARM SYSTEM is significant, $F(1,57) = 5.06$, $p = .03$. The interaction effect ALARM SYSTEM x COST is not significant, $F(1,57) = 00.00$, $p = .94$.

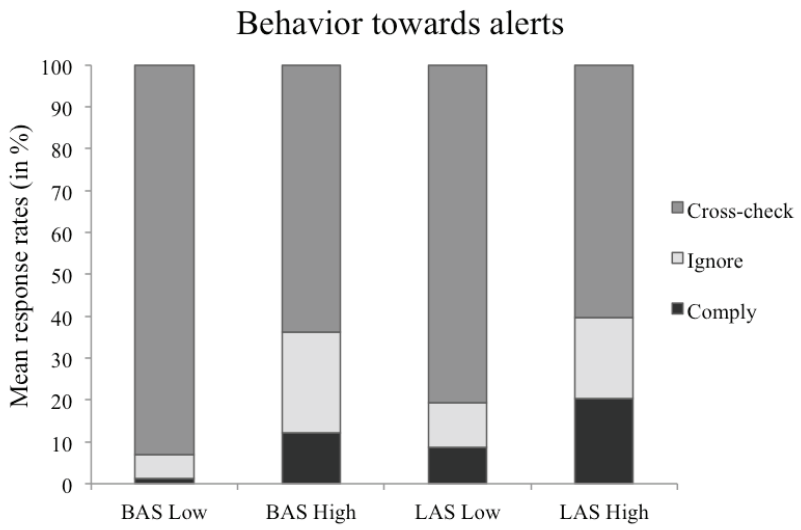


Figure 19. Means of the proportions of different kinds of responses to alerts depending on the experimental conditions

Participants' cross-checking rate of alerts

Participants in the low-cost condition cross-checked alerts more often *(M* = 86.41%) than participants in the high-cost condition *(M* = 61.96%), which suggests that our experimental manipulation did work. This expected main effect of COST is significant, *F(*1,57) = 31.82, *p* = .00. Besides, participants cross-checked more often alerts emitted by BAS *(M* = 78.46%) than LAS *(M* = 70.72%). The main effect of ALARM SYSTEM is significant, *F(*1,57) = 9.16, *p* = .00. The interaction effect ALARM SYSTEM x COST is not significant, *F(*1,57) = 0.04, *p* = .85.

Participants' ignoring rate of alerts

The main effect of ALARM SYSTEM, of COST and the interaction effect ALARM SYSTEM x COST are not significant, *p*s > .05.

Mean response rates to alarms

The means of participants' compliance rate, cross-checking rate, and ignoring rate of alarms are displayed in Figure 20.

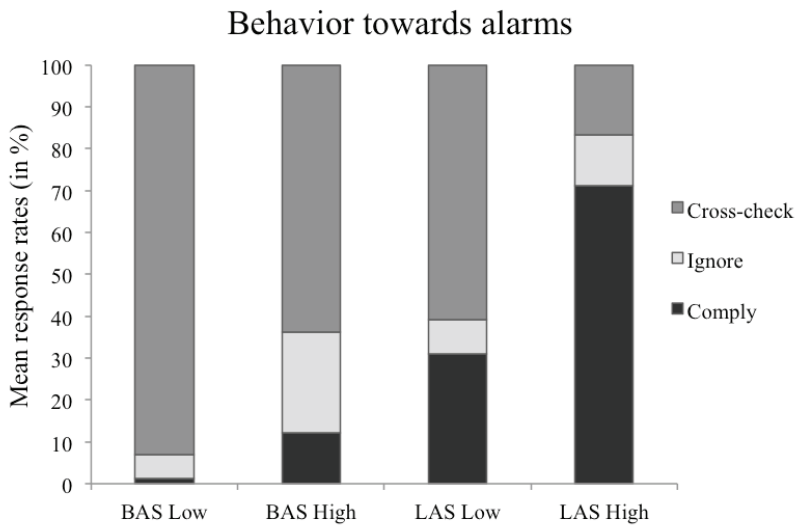


Figure 20. Means of the proportions of different kinds of responses to alarms depending on the experimental conditions

Participants' compliance rate with alarms

As one can see in Figure 20 and as expected in Hypothesis 2, participants' direct compliance rate is higher in the high-cost condition *(M* = 42.69%) than in the low-cost condition *(M* = 17.04%). The main effect of COST is significant, $F(1,57) = 10.06$, $p = .00$. Figure 20 further shows that LAS users complied more with alarms *(M* = 50.36%) than BAS users *(M* = 6.71%). The main effect of ALARM SYSTEM is significant, $F(1,57) = 30.06$, $p = .00$. The interaction effect ALARM SYSTEM x COST is not significant, $F(1,57) = 10.06$, $p = .08$.

Participants' cross-checking rate of alarms

Participants' cross-checking rates tend to show the opposite pattern than participants' compliance rate. Participants in the low-cost condition cross-checked more often the diagnosis of alarms *(M* = 75.84%) than participants in the high-cost condition *(M* = 39.45%) as expected in Hypothesis 2. The main effect of COST is significant, $F(1,57) = 18.32$, $p = .00$. Moreover participants cross-checked alarms more often when emitted by BAS *(M* = 78.46%) than by LAS *(M* = 39.43%). This effect was expected in the low-cost condition only (Hypothesis 3b). The results show that this effect also exists in the high-cost condition. This suggests that the placement of an additional criterion in BAS that transforms it into LAS impact participants' cross-checking frequencies of alarms. The main effect of ALARM

SYSTEM is significant, $F(1,57) = 21.55$, $p = .00$. The interaction effect ALARM SYSTEM x COST is not significant, $F(1,57) = 0.76$, $p = .39$.

Participants' ignoring rate of alarms

The main effect of COST failed the conventional level of significance, $F(1,57) = 3.37$, $p = .07$. The main effect of ALARM SYSTEM, and the interaction effect ALARM SYSTEM x COST are not significant, $ps > .05$.

Mean response rates to warnings

The means of participants' compliance rate, cross-checking rate, and ignoring rate of warnings are displayed in Figure 21.

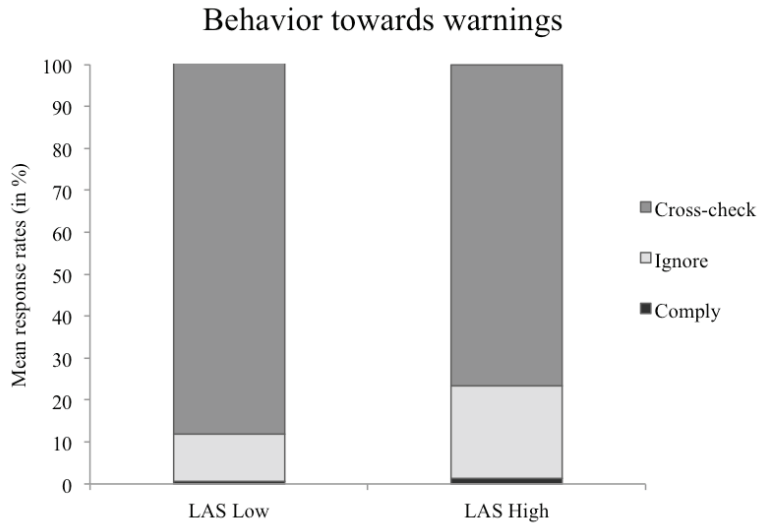


Figure 21. Means of the proportions of different kinds of responses to warnings depending on the experimental conditions

Participants' compliance rate with warnings

As one can see in Figure 21, almost none of the participants complied with warnings. No difference emerged between the low- and high-cost conditions, $F(1,29) = 1.38$, $p = .25$.

Participants' cross-checking rate of warnings

One can observe in Figure 21, that participants in the low-cost condition cross-checked slightly more warnings than participants in the high-cost condition. This main effect of COST is however not significant, $F(1,29) = 2.10$, $p = .15$.

Participants' ignoring rate of warnings

The main effect of COST on participants' ignoring rate of warnings is not significant, $F(1, 29) = 1.83$, $p = .19$.

Mean response rates to non-alerts

The means of participants' compliance rate, cross-checking rate, and ignoring rate of non-alerts are displayed in Figure 22.

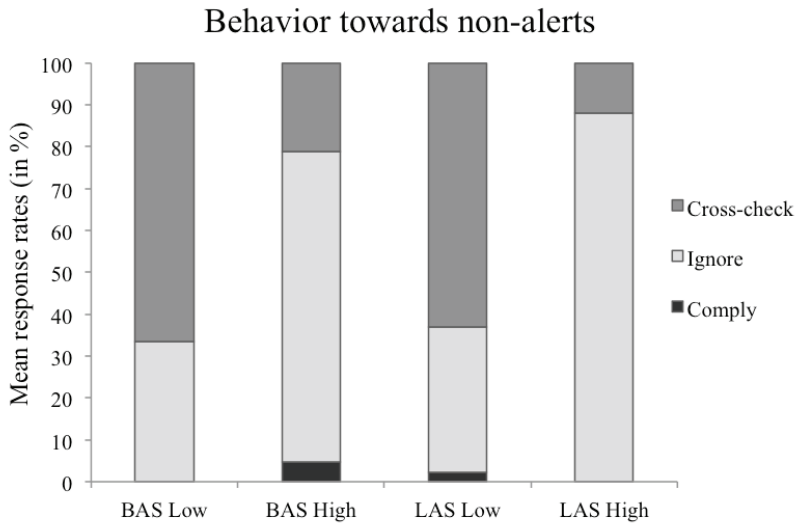


Figure 22. Means of the proportions of different kinds of responses to non-alerts depending on the experimental conditions

Participants' compliance rate with non-alerts

As one can see in Figure 22, almost none of the participants complied directly with non-alerts emitted by the alarm systems (i.e., clicked on the repair button). The main effect of ALARM SYSTEM, COST and the interaction effect ALARM SYSTEM x COST are not significant, *p*s $< .05$.

Participants' cross-checking rate of non-alerts

Surprisingly, participants in the low-cost condition cross-checked non-alerts even though the NPV of the alarm systems is very high in this experiment. Participants in the low-cost condition cross-checked a majority of non-alerts *(M* = 64.61%) whereas participants in the high-cost condition cross-checked only a few non-alerts *(M* = 16.33%). The main effect of COST is significant, *F(*1,57) = 25.95, *p* = .00. No main effect of ALARM SYSTEM and no interaction effect ALARM SYSTEM x COST revealed significance, *p*s > .05.

Participants' ignoring rate of non-alerts

Participants ignored non-alerts in the low-cost condition less often *(M* = 34.22%) than in the high-cost condition *(M* = 81.34%). This main effect of COST is significant, *F(*1,57) = 23.58, *p* = .00. No main effect of ALARM SYSTEM and no interaction effect ALARM SYSTEM x COST revealed significance, *p*s > .05.

Investigation of behavioral differentiation of participants induced by the use of LAS

Comparison between participants' responding behavior towards alarms vs. warnings emitted by the LAS

A t-test was used for this analysis. Participants complied more with alarms emitted by the LAS *(M* = 48.70%) than with warnings *(M* = 0.97%), *F(*1,29) = 39.31, *p* = .00. Moreover participants cross-checked more warnings *(M* = 81.94%) than alarms emitted by the LAS *(M* = 40.74%), *F(*1,29) = 36, *p* = .00. These results confirm Hypothesis 3a and suggest that participants differentiate their responding behavior between alarms and warnings emitted by LAS.

7.3.3. Participants´ performance in the alarm task

All analyses about participants´ performance in the alarm task were performed using 2(ALARM SYSTEM)x2(COST) between-groups ANOVAs.

Four participants were excluded from the analyses of the percentage of hits and misses based on their outlying Standard Deviation of Residuals (SDR) and Cook values. Moreover, the

presence of these participants had clearly a problematic impact on the homogeneity of the variances, which would have make the use of an ANOVA for these analyses impossible.

Participants' percentage of hits

Figure 23 displays the results of participants' percentage of hits.

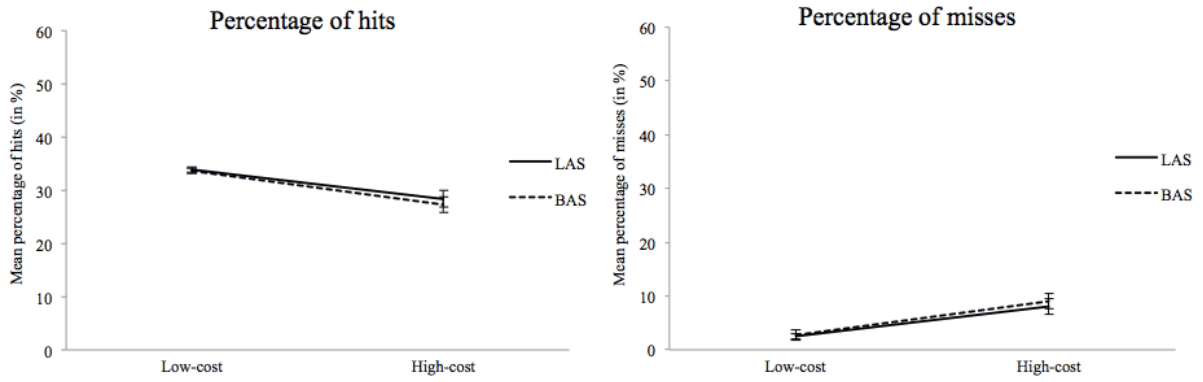


Figure 23. Mean and mean standard deviations of participants' percentage of hits (left panel) and misses (right panel) in the alarm task depending on the type of alarm system and the cost of consulting AVI

As expected in Hypothesis 4, participants in the low-cost condition produced significantly more hits *(M* = 33.75%) than participants in the high-cost condition *(M* = 27.86%). The main effect of COST is significant, *F(*1,53) = 26.81, *p* = .00. BAS users and LAS users do not differ with regards to the percentage of hits they produced. The main effect of ALARM SYSTEM is not significant, *F(*1,53) = 0.30, *p* = .58. Finally, the expected interaction effect ALARM SYSTEM x COST does not appear at a descriptive level and is not significant, *F(*1,53) = 0.15, *p* = .70.

Participants' percentage of misses

Figure 23 displays the results of participants' percentage of misses. As expected, participants in the low-cost condition *(M* = 2.61%) produced significantly less misses than participants in the high-cost condition *(M* = 8.50%). The main effect of COST is significant, *F(*1,53) = 26.81, *p* = .00. The use of BAS vs. LAS did not lead to any differences on the percentage of

misses. The main effect of ALARM SYSTEM is not significant, $F(1,53) = 0.30$, $p = .58$. Finally, and contrary to our expectations, the interaction effect ALARM SYSTEM x COST is not significant, $F(1,53) = 0.15$, $p = .70$.

Participants' percentage of false alarms

Results about participants' percentage of false alarms are displayed in Figure 24.

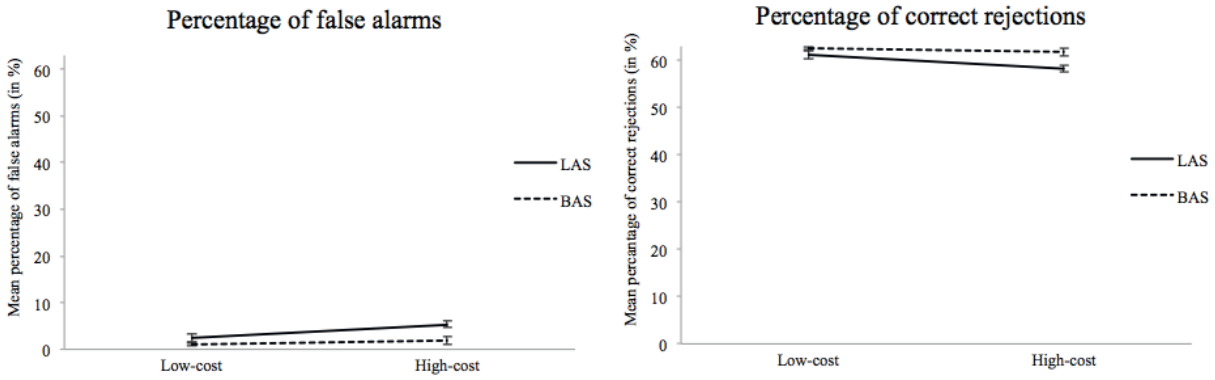


Figure 24. Mean and mean standard deviations of participants' percentage of false alarms (left panel) and correct rejections (right panel) in the alarm task depending on the type of alarm system and the cost of consulting AVI

Three outliers have been deleted from this analysis based on their SDR and Cook values. Participants in the low-cost condition produced as expected significantly fewer false alarms *(M* = 1.76%) than participants in the high-cost condition *(M* = 3.84%). The main effect of COST is significant, $F(1,54) = 11.00$, $p = .00$. The results also show that participants using the BAS produced significantly fewer false alarms *(M* = 1.33%) than participants using the LAS *(M* = 3.86%). This effect was expected for the low-cost condition only. The main effect of ALARM SYSTEM is significant, $F(1,54) = 5.92$, $p = .02$. Finally, and contrary to our expectations, there is no interaction effect ALARM SYSTEM x COST, $F(1,54) = 1.86$, $p = .18$.

Participants' percentage of correct rejections

Figure 24 displays the results about participants' percentage of correct rejections. The same three outliers deleted from the above analysis were also deleted from this analysis based on their outlying values. As expected, participants in the low-cost condition produced as expected significantly more correct rejections *(M* = 61.87%) than participants in the high-cost condition *(M* = 59.79%). The main effect of COST is significant, *F(*1,54) = 11.00, *p* = .00. The results further show that BAS users produced significantly more correct rejections *(M* = 62.30%) than LAS users *(M* = 59.76%). Such effect was expected for the low-cost condition only. The main effect of ALARM SYSTEM is significant, *F(*1,54) = 5.92, *p* = .02. Finally, the expected interaction effect ALARM SYSTEM x COST is not significant, *F(*1,54) = 1.86, *p* = .18.

7.3.4. Participants´ performance in concurrent tasks

All analyses about the participants´ performance in concurrent tasks were performed using 2(ALARM SYSTEM)x2(COST) between-groups ANOVAs.

Performance in the Coolant Exchange Task (CET)

Participants' performance are displayed in Figure 25.

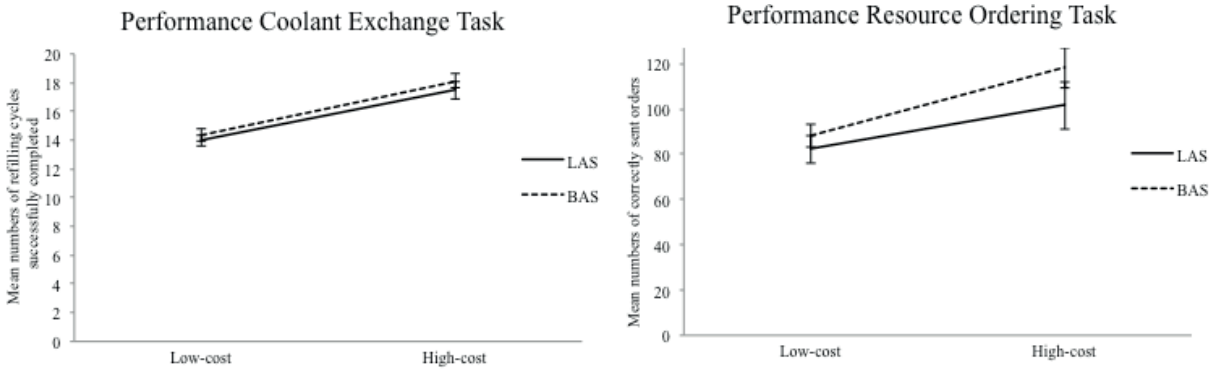


Figure 25. Means and mean standard deviations of the number of refilling cycles successfully completed in the CET (left panel) and of the number of correctly sent orders in the ROT (right panel) depending on the type of alarm system and the cost of consulting AVI

Contrary to our expectations, participants in the high-cost condition successfully completed more refilling cycles *(M* = 17.79) than participants in the low-cost condition *(M* = 14.17). This main effect of COST is significant, *F(*1,58) = 55.41, *p* = .00. The main effect of ALARM SYSTEM and the interaction effect ALARM SYSTEM x COST were not significant, *F(*1,57) = 0.87, *p* = .36, and *F(*1,57) = 0.00, *p* = .96 respectively.

Performances in the Resource Ordering Task (ROT)

Participants' performance are displayed in Figure 25 (see above). One outlier was deleted from the two analyses of participants´ performance in ROT because of its outlying values on SDR and Cook. Participants' results concerning the amount of correctly sent orders in the ROT are displayed in Figure 25. Contrary to our expectations, participants in the high-cost condition sent more correct orders *(M* = 109.86) than participants in the low-cost condition *(M* = 85.07). This main effect of COST is significant, *F(*1,56) = 9.57, *p* = .00. The main effect of ALARM SYSTEM and the interaction effect ALARM SYSTEM x COST were not significant, *F(*1,56) = 1.98, *p* = .17, and *F(*1,56) = 0.27, *p* = .61 respectively.

Regarding the percentage of appeared orders participants responded to in the ROT, results did not show any significant effects, *p*s > .05.

7.3.5. Participants´ subjective ratings

A 2(ALARM SYSTEM)x2(COST) between-groups ANOVA was performed on participants' workload and trust ratings.

Workload

Results did not show any significant effect on participants' workload, *p*s > .05.

A more precise analysis of the subscales of the NASA-TLX scales shows a significant main effect of COST on the subscale temporal demand, *F(*1,57) = 5.21, *p* = .03. Contrary to our expectations, participants in the low-cost condition reported more temporal demand *(M* = 8.02) than participants in the high-cost condition *(M* = 7.01).

Trust in the alarm system: item

As one can see in the left panel of Figure 26, participants reported more trust towards LAS *(M* = 5.57) than towards BAS *(M* = 3.90). This main effect of ALARM SYSTEM is significant, *F(*1,57) = 8.83, *p* = .00. The main effect of COST and the interaction effect ALARM SYSTEM x COST were not significant, *F(*1,57) = 0.06, *p* = .80, and *F(*1,57) = 2.18, *p* = .15 respectively.

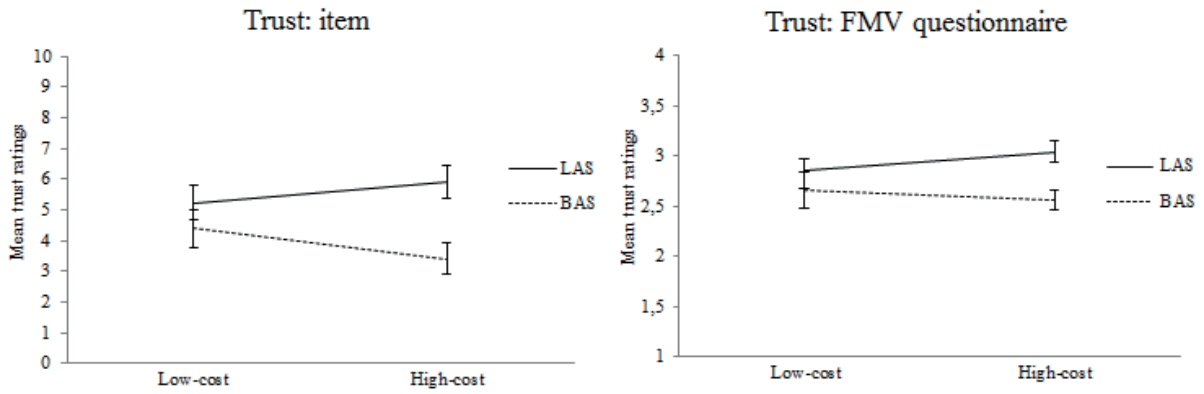


Figure 26. Participants' mean trust ratings at the item (left panel) and at the FMV questionnaire (right panel) depending on the type of alarm system and the cost of consulting AVI

Trust in the alarm system: FMV trust scale

One participant was not included in this analysis as he had left the room before responding to the FMV questionnaire. As shown in the right panel of Figure 26, participants reported more trust towards LAS *(M* = 2.94) than towards BAS *(M* = 2.61). This main effect of ALARM SYSTEM is significant, *F(*1,56) = 6.37, *p* = .02. The main effect of COST and the interaction effect ALARM SYSTEM x COST were not significant, *F(*1,56) = 0.09, *p* = .76, and *F(*1,56) = 1.02, *p* = .32 respectively.

A more precise analysis of the subscales of the FMV trust scale shows a significant main effect of ALARM SYSTEM on the subscale reliability, *F(*1,56) = 14.83, *p* = .00 as well as on the subscale use, *F(*1,56) = 11.51, *p* = .00. Participants reported that the LAS was more

reliable *(M* = 2.49) than the BAS *(M* = 1.87) and that the LAS was more useful *(M* = 2.62) than the BAS *(M* = 2.53).

7.4. Discussion

Experiment 2 was conducted to investigate the influence of the cost of consulting AVI on participants' performance with BAS and LAS. The cost of consulting AVI was operationalized by manipulating the time and amount of effort spent by participants. No benefit of LAS over BAS was expected when the cost of consulting AVI is low. This prediction was based on previous results conducted with the paradigm M-TOPS that showed that LAS do not bring any benefit over BAS with regard to alarm task performance when AVI is provided to participants (Wiczorek & Manzey, 2014). However a benefit of LAS over BAS was expected in the high-cost condition. More precisely, it was expected that all participants would cross-check alerts less often when the cost of consulting AVI increases. However this would affect BAS users' performance in the alarm task stronger than LAS users' performance because LAS allows for a more precise reduction of cross-checking than BAS.

We tested four experimental conditions in this experiment, namely low-cost BAS, high-cost BAS, low-cost LAS, and high-cost LAS. The dependent variables were the same as in Experiment 1: how participants respond to alarms (H1, H2, H3), and what are their performance in the alarm task (H4, H5, H6) as well as in the concurrent tasks (H7). In comparison to Experiment 1, we added a measure of participants' subjective workload (H8). Trust was again a variable of interest (H9).

The manipulation of the alert-PPV was successful. Participants perceived LAS-alarms as more reliable than BAS-alarms so that a difference is likely to occur on participants´ response rates between these two conditions. Participants also correctly perceived LAS-alarms as more reliable than LAS-warnings so that they are likely to differentiate their behavior towards LAS. This is mandatory because if participants do not perceive LAS-alarms as being different from LAS-warnings and BAS-alarms, one cannot expect LAS users to differentiate their behaviors. This would be problematic because our hypotheses are based on the assumption that LAS users differentiate their behavior.

Participants overestimated the PPV of BAS-alarms in the low-cost condition and the PPV of warnings in both the low- and high-cost condition. As explained in Experiment 1, this is probably due to the human bias of overestimating low probabilities (Kahneman & Tvesky, 1984). This overestimation might affect participants´ response rate, especially in the high-cost condition. In the low-cost condition, one expects participants to cross-check the majority of alerts anyway. Participants underestimated the NPV of the system in all conditions; and, most importantly, these NPV values do not differ from each other.

As per our first hypothesis (H1) a main effect of COST on participants' cross-checking rate of alerts was expected. The results confirm this hypothesis. Participants cross-checked alerts in the high-cost condition less often than in the low-cost condition. This suggests than the experimental manipulation of the cost of consulting AVI was successful and had a direct impact on participants' responding behavior.

Interestingly this main effect is caused by participants´ behavior towards alarms only. As expected in Hypothesis 2, participants cross-checked more alarms in the low-cost condition than in the high-cost condition. Participants reacted the same way to warnings whatever the cost was. A slight reduction of cross-checking frequencies of warnings emerged in the descriptive data, but did prove to be significant.

One could wonder why participants reduced their cross-checking behavior of BAS-alarms and not of LAS-warnings even though the perceived PPV of both stages was almost the same (.50 and .45 respectively). Contrary to BAS users, LAS users had the option to decide what type of alerts they thought would be most useful to cross-check. Because alarms are more reliable than warnings, LAS users probably decided to reduce their cross-checking of alarms only.

It is also interesting to note that, when increasing the cost of consulting AVI, BAS users reduced their cross-checking behavior and started to ignore alarms rather than to comply with them. In this Experiment, 'ignoring alarms' was the second most chosen reaction to BAS-alarms after 'cross-checking'. This choice makes sense with regards to the PPV of alarms (.35) but stands in contradiction with findings from Experiment 1. In Experiment 1, BAS users mostly complied with alarms even though the PPV was exactly the same than in Experiment 2. One possible explanation of this difference in results between the two

experiments could be the impression participants had of the amount of time they could invest in concurrent tasks. Participants might have thought that they are limited in time. In 20 minutes they should achieve the best performance both in the alarm task and in the concurrent tasks. In Experiment 1, participants invested very little time in the alarm task contrary to participants in the high-cost condition of Experiment 2 who cross-checked the majority of BAS-alarms. Because participants of Experiment 2 had already invested a lot of time cross-checking alarms, they might have chosen to ignore rather than comply with the rest of the BAS-alarms in order to liberate more time for the concurrent tasks. Complying with alarms definitely requires more time than ignoring them. Participants have to click on a button while no action is required to ignore alarms. Participants from Experiment 1 did not face such trade-off and might have preferred to comply with alarms.

Interestingly, an effect of COST has also been found on participants' cross-checking rate as well as on their ignoring rate of non-alerts. However, participants´ response rates towards non-alerts was not a focus in this study. Since the non-alert stage was designed very reliable (NPV= .97), it was expected that participants would always ignore non-alerts. As such we did not expect that any of our experimental manipulations would have an effect on participants´ behavior towards non-alerts. Nevertheless, participants cross-checked 64.61% of the non-alerts in the low-cost condition in comparison to 16.33% in the high-cost condition. Similar effects have been report by Wiczorek and Manzey (2014), to a smaller extent. What could explain this effect?

If one considers the time allocated to each task, this behavior is not rational because cross-checking non-alerts having such a high NPV is clearly a waste of time. A possible explanation for this effect is that participants underestimated the NPV of non-alerts (.81 instead of .97) and might have behaved according to this wrong NPV evaluation.

Another explanation of this effect could be the cost associated to successively selecting different responses/actions (Kornblum, 1973; Rumelhart & Norman, 1982). In the low-cost condition, BAS and LAS users cross-checked most of the alerts emitted by the alarm systems. Since participants have the routine to cross-check alerts, it represents a judicious strategy to expand this routine to the non-alerts and automatically select the response 'cross-checking' when a non-alert is emitted. This decreases response time to keep the same routine rather than

systematically analyzing the output of the alarm system and successively selecting different response (cross-check vs. ignore) depending on what output is emitted. Such an explanation is consistent with the fact that participants did not cross-check non-alerts in the high-cost condition because participants had no systematic cross-checking routine of alerts.

Hypothesis 3 aimed to test the behavioral differentiation of participants induced by the use of LAS. Our results confirm that that a behavioral differentiation occurred. This suggests that a behavioral differentiation, as already observed in Experiment 1, also occurs when AVI is available. Participants complied more with LAS-alarms than LAS-warnings and cross-checked less LAS-alarms than LAS-warnings (H3a). This difference is more pronounced in the high-cost condition. Concerning Hypothesis 3b, participants complied more with LAS-alarms than with BAS-alarms and cross checked less LAS-alarms than BAS-alarms in the low-cost condition. This effect is even more unequivocal in the high-cost condition, which suggests that in situations in which resources are limited, participants particularly use the cues offered by LAS in order to optimize their performance.

In Hypothesis 4 a main effect of COST on participants' performance (percentage of hits, misses, false alarms, and correct rejections) was expected. The results confirmed this hypothesis. Participants had better performance in the low-cost condition than in the high-cost condition. These results are not surprising given participants cross-checking rates of alerts. Because participants cross-checked less alerts in the high-cost condition, they produced more errors.

Nonetheless, some misses and false alarms have been produced even though participants consulted AVI, suggesting that some participants' errors occurred during the visual search task that was required to analyze the AVI. In the low-cost condition participants did not make any errors while consulting AVI whereas participants in the high-cost condition produced in average 2.93 errors: 0.21 false alarms and 2.72 misses. This difference between false alarms and misses while consulting AVI is significant, $F(1,57) = 32.13$, $p = .00$. The errors made by participants contribute to the main effect of COST on participants´ percentage of false alarms and misses reported above. It is important to note that the errors made by participants in the visual search task are not exclusively responsible for the main effect of COST. If the trials in which false alarms and misses have been made after cross-checking AVI are deleted, the

effect of COST on participants' numbers of misses and false alarms remains significant, $F(1,57) = 3.98$, $p = .05$ and $F(1,57) = 7.95$, $p = .01$ respectively. In addition, most of the errors made by participants occurred in trials, in which participants did not use AVI (9.10 errors), and only a minority of errors were made after consulting AVI (2.93 errors).

It is relevant to point out that the errors participants produced while consulting AVI are almost exclusively misses. This is not surprising since errors produced in visual search tasks tend to be misses (Wickens et al., 2013). This might be explained by an expectancy effect (Wolfe, Horowitz & Kenner, 2005; Wolfe, Horowitz et al., 2007). Because the base rate of critical events was low (i.e., only 24% of containers were improper), participants might have rather expected to not find the visual target (i.e., the letter K) and might have stopped their search earlier than participants who would not have such low expectations. Such results are very interesting because it means that when consulting AVI consists in performing a visual search task and the base rate of a critical event is low, there is a risk that operators miss critical events even though they consulted AVI. One countermeasure to this could be to sometimes introduce a mock target (Wilkinson, 1964). For example, providing radiologist CT scans displaying a tumor or screening luggage at the airport security having a prohibited luggage item inside. Such methods are already implemented by aviation security agencies (Wickens et al., 2013).

In line with Hypothesis 5, an interaction effect ALARM SYSTEM x COST on participants´ percentage of hits and misses was expected. No difference between LAS and BAS was expected in the low-cost condition while a benefit of LAS over BAS was expected in the high-cost condition. More precisely, it was expected that the performance of BAS users would strongly decrease whereas the performance of LAS users would decrease only slightly from the low-cost condition to the high-cost condition. As a consequence, performance would be better with LAS than BAS.

The results do not confirm this hypothesis. The results show that LAS users and BAS users have the same performance. In the low-cost condition a ceiling effect emerged as expected. LAS users and BAS users exhibited performance close to perfect. However, results show that performance of LAS users and BAS users both deceased to exactly the same extent in the high-cost condition. This is surprising because at a behavioral level, participants behaved

exactly as expected. On the one hand, BAS users significantly reduced their cross-checking behavior so that they inevitably produced more errors. On the other hand, LAS users differentiated their responding behavior depending on the likelihood information provided by the LAS; they benefited from an advantage in comparison to BAS users. Two reasons can explain these results. First, performance of BAS users did not decrease very strongly because they still cross-checked a lot of alarms. Second, LAS users reduced their cross-checking behavior towards warnings when the cost of cross-checking increased and started to ignore them. Even though this effect was descriptive only and not statistically significant, this had an effect on performance of LAS users. Ignoring warnings having a PPV of .21 definitely increases participants' misses, especially that most of LAS-alerts are warnings. A statistical analysis on participants´ percentage of hits and misses on warnings trials only shows that LAS users had worse performance in the high-cost condition (6.07 hits and 3.93 misses) than in the low-cost condition (9.07 hits and 0.93 misses). This main effect of COST on participants´ amount of hits and misses in warning trials is significant, $F(1,26) = 19.96$, $p = .00$. These two reasons explain why LAS users experienced a decrease of performance similar to BAS users when the cost of AVI was increased.

Hypothesis 6 focused on participants' percentage of false alarms and correct rejections. An interaction effect ALARM SYSTEM x COST was expected with a benefit of BAS over LAS in the low-cost condition and a benefit of LAS over BAS in the high-cost condition. The results only partially confirm this hypothesis because an unexpected benefit of BAS over LAS was found not only in the low-cost condition but also in the high-cost condition. In the low-cost condition, LAS users produced more false alarms and less correct rejections than BAS users because they complied with about 30% of LAS-alarms having a PPV of .78. In the high-cost condition however, the benefit of BAS over LAS can be mainly explained by the unexpected responding behavior of BAS users mentioned above. BAS users reduced their cross-checking behavior of only 29.22% in favor of ignoring the alarm system. As a consequence, BAS users rather produced misses than false alarms. One can actually see in the descriptive data that the percentage of false alarms produced by BAS users is almost the same in the low- and high-cost condition, 0.97% and 1.79% respectively. Whereas BAS users produced significantly more misses in the high-cost condition than in the low-cost condition, 9.09% and 2.71% respectively. Such results were surprising as it was expected that BAS users

would reduce their cross-checking behavior in favor of complying with instead of ignoring alarms which would have led them to rather produce false alarms than misses.

Hypothesis 7 concerned participants' performance in the concurrent tasks. A main effect of COST on the performance was expected. As such we expected participants in the low-cost condition to have better performance than participants in the high-cost condition. This is because the visual search task requires fewer resources in the low-cost condition than in the high-cost condition. The results show a main effect of COST but in the opposite direction as the one expected: participants had a better performance in the high-cost condition than in the low-cost condition. From a first look, one could think that this pattern of performance is caused by participants' cross-checking behavior. Participants cross-checked less alerts in the high-cost condition than in the low-cost condition. Hence, one could think that participants had more time for the concurrent tasks. This explanation is not plausible because the experimental sessions did not last a fixed time but instead stopped after 100 containers went through the control station. This means that cross-checking would not make participants loose time that they could have invested in the concurrent tasks. Another more likely explanation is that participants in the low-cost condition experienced more switching-cost than participants in the high-cost condition and this affected participants' performance. Such an explanation becomes obvious after looking at participants´ estimation of their workload. It was expected that participants would experience more workload in the high-cost condition than in the low-cost condition (Hypothesis 8) because of the extra visual search task required from participants when consulting AVI (i.e., no pop-out effect). Nevertheless, the results from NASA-TLX show a main effect of COST on the subscale 'temporal demand' in the opposite direction as expected: participants reported more workload on the temporal demand scale in the low-cost condition than the high-cost condition. Participants were presented the temporal demand in the following way: 'How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?' One possible explanation for this unexpected effect could be that the frequency at which participants switched attention between the alarm task and the concurrent tasks had an effect on their response to the temporal demand item of NASA-TLX. In the low-cost condition participants could decide to repair vs. ignore the container directly after cross-checking AVI. In the high-cost condition participants had to wait 5 seconds after clicking on

the repair button to give their response, regardless if they had already finished the visual task or not. As a consequence, the frequency at which participants had to switch attention between tasks was higher in the low-cost condition than in the high-cost condition. This requirement to switch tasks imposed a higher cognitive demand on participants in the low-cost condition and might have lowered their performance in the concurrent tasks in comparison to participants from the high-cost condition (Rogers & Monsell, 1995). This main effect of COST observed on participants' evaluation of their temporal demand is likely to explain the main effect of COST found on participants' performance in the concurrent tasks. These results could be pretty worrying regarding the successful impact of our experimental manipulation of the cost. The goal of our experimental manipulation was to impose more workload to participants in the high-cost condition than to participants in the low-cost condition, not the opposite. However, because participants' compliance rate is lower in the high-cost condition than the low-cost condition this confirms that participants perceived the AVI as more costly in the high-cost condition than the low-cost condition. Still, this is clearly a limitation of the experiment. Another experimental design in which the low-cost condition does not represent a higher switching cost for participants would have been of help in this regard

The last hypothesis (H9) focused on participants' trust. A main effect of ALARM SYSTEM on trust was expected. The results confirmed our hypothesis. Participants reported more trust towards LAS than BAS, in the trust item as well as in the trust questionnaire. A more precise analysis of the subscale of the trust questionnaire showed that this main effect of ALARM SYSTEM occurs on the subscales of reliability and use. These results support findings previously reported by Ragsdale (2012) and Wiczorek and Manzey (2014).

Experiment 2 shows that LAS does not present any benefit over BAS regarding participants´ performance if AVI is provided to participants, whatever the cost of consulting AVI. Interestingly, even a benefit of BAS over LAS has even been found regarding participants´ percentage of false alarms and correct rejections. As a conclusion, there is no interest in implementing a LAS in settings in which AVI is available. A low-cost AVI can even increases operators´ workload because of the switching costs caused by the constant switches of attention focus. This can result in a decrease of performance in the concurrent tasks. Such findings about participants' performance in the concurrent tasks are interesting but can only be applied to real-world settings in which switching frequencies are very high. In most real-

world settings operators are rarely faced with alarms and have therefore less time pressure while consulting AVI in comparison to the participants in our experiment. More important, this study emphasizes the necessity of considering switching costs in alarm system research because it can help interpreting non-rational behaviors (e.g., participants´ responding behavior to non-alerts in this experiment). The importance of considering switching costs in research on alarms will be further discussed in the main discussion.

8. Experiment 3: The optimal number of stages in likelihood alarm systems

The main objective of Experiment 3 is to investigate the optimal number of stages in LAS. [1]

Contrary to BAS, LAS emits different types of alert signals, each of them corresponding to a different likelihood that a critical event is present. This provides operators with additional information that enable to adapt their responding behavior and hence increase their chances to correctly comply with true alarms and to ignore false alarms.

This raises the question of which degree of specificity of alarm systems (i.e., number of stages of LAS) is optimal for operators. Shurtleff (1991) and Wiczorek et al. (2014) have already investigated this question. Shurtleff (1991) compared a BAS, a 4-stage LAS, a 6-stage LAS, an 8-stage LAS, and a control condition in which participants did not have access to any alarm system. In this study, the difficulty of the decision in the alarm task was also manipulated. Shurtleff's (1991) results show that the number of stages only has an effect on participant's performance when the decision task supported by the alarm system is difficult. In this case, participants demonstrated a better performance with the 4-stage LAS and 8-stage LAS than with the BAS or with no alarm system. However, the alarm system used in the study conducted by Shurtleff (1991) to investigate the influence of the number of stages did not use the colors and wording characteristics of an alarm system but its design was more similar to a decision assistant system. The diagnosis of the alarm system was indeed presented by changing the length of a horizontal bar: the longer the bar, the greater was the probability that a signal was present. This narrows the generalization of these results to alarm system research. Wiczorek et al. (2014) compared a BAS, a 3-stage LAS, and a 4-stage LAS using the experimental paradigm M-TOPS. The results shows that a 4-stage LAS improve participants' performance (less incorrect decisions) in comparison to a 3-stage LAS.

Experiment 3 aims to replicate the findings of Wiczorek et al. (2014) using different PPV alarm characteristics and also aims to further investigate the question of the optimal number of stages in LAS by comparing a 3-stage, 4-stage, and 5-stage LAS. The same measures as in

[1] Part of these results has been published already in Balaud and Manzey (2015).

Experiment 1 and 2 were used. Participants´ trust was measured using only the FMV questionnaire and, as in Experiment 2, workload was a variable of interest.

8.1. Method

8.1.1. Participants

Forty-eight participants (22 men, 26 women) participated in this study. Participants ranged in age from 18 to 44 with a mean age of 27.02 (*SD* = 5.77). Among them 72.9% were students, 20.8% were employed, and 6.3% had another status. Participants registered voluntarily to the experiment after having been recruited on an online portal for experiments (Prometei) of the Technische Universität Berlin. None of them was suffering from any distortion of vision, such as red-green color blindness, which might have interfered with the experiment.

Participants were paid €5 for their participation and, in addition, they had the chance to gain up to €4 depending on their performance in the experiment. Each participant was randomly assigned to one of the four experimental conditions. An equal repartition of men and women in each group was observed.

8.1.2. Task

Experimental paradigm M-TOPS

Participants completed the three tasks of the experimental paradigm M-TOPS (described in Part 6.1.2.). There was no possibility for participants to consult AVI. The LAS differed in their number of stages, namely three stages, four stages, and five stages. In the 4-stage LAS, a yellow warning stage was added and the following written message was shown: ‘The molecular weight in this container Nr.X might be too high’. In the 5-stages LAS a yellow-orange warning stage was added and the following written message was shown: ‘The molecular weight in this container Nr.X is possibly too high’.

Alarm system

All alarm systems had the same sensitivity (d' = 1.8). The systems differed in the number of stages they had. The first criterion separating the non-alert stage from the other stages was kept constant at all times (c = -1.05). Figure 27 shows the three LAS used.

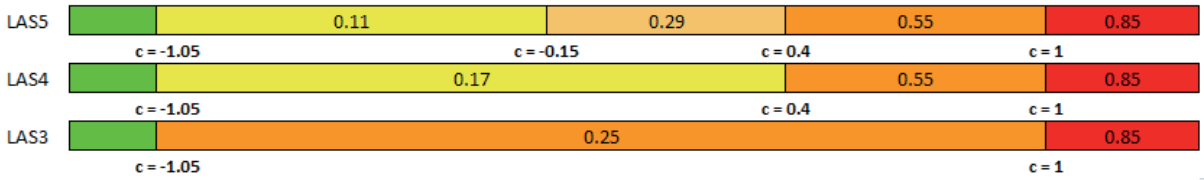


Figure 27. Systems characteristics of the three LAS

The numbers reported in the boxes correspond to the PPV of each stage and the number reported under each separation corresponds to the criterion. The colors are the ones used for the experiment. They were chosen according to findings from previous studies investigating the link between colors and perceived urgency or perceived hazard (Braun & Silver, 1995; Chapanis, 1994; Jacobs & Suess, 1975; Smith-Jackson & Wogalter, 2000; Wogalter, Conzola, & Smith-Jackson, 2002).

Pay-off system

The pay-off system was the same as in Experiment 1. This is described in Part 6.1.2.

8.1.3. Design

The experimental design was composed of a single between-subjects variable, namely the number of stages. This variable had three conditions: 3-stage (LAS3), 4-stage (LAS4), and 5-stage (LAS5).

8.1.4. Dependent variables

Behavioral data: participants' response rates

Like in Experiment 1, we recorded participants´ response rates to the different stages and to the overall system (non-alerts excluded). These included:

- Compliance rate with alerts;
- Ignoring rate of alerts;

- Compliance rate with alarms;
- Ignoring rate of alarms;

- Compliance rate with orange warnings;
- Ignoring rate of orange warnings;

- Compliance rate with yellow-orange warnings;
- Ignoring rate of yellow-orange warnings;

- Compliance rate with yellow warnings;
- Ignoring rate of yellow warnings;

- Compliance rate with non-alerts;
- Ignoring rate of non-alerts.

Behavioral data: participants´ response strategies

Participants´ response strategies for each stage were also recorded:

- Response strategies in response to alarms;
- Response strategies in response to orange warnings;
- Response strategies in response to yellow-orange warnings;
- Response strategies in response to yellow warnings;

- Response strategies in response to non-alerts.

Performance data

As in Experiment 1, four dependent variables were used to record participants' performance. These are:

- Percentage of hits produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of misses produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of false alarms produced by participants when the alarm system emitted an alarm or a warning;
- Percentage of correct rejections produced by participants when the alarm system emitted an alarm or a warning.

As in Experiment 1, three dependent variables were used to assess participants´ performance in the concurrent tasks. These are:

- Number of correctly sent orders in the Resource Ordering Task (ROT);
- Proportion of orders participants responded to in the Resource Ordering Task (ROT);
- Number of refilling cycles successfully completed in the Coolant Exchange Task (CET).

Subjective data

Participants' subjective experience was measured using two dependent variables:

- Trust in the alarm system (FMV questionnaire, Wiczorek 2011);
- Workload.

The NASA-TLX was used to assess participants' workload. See Part 7.1.4 for further information about this measure.

8.1.5. Procedure

The procedure was the same as in Experiment 1 and 2. Table 9 resumes the procedure.

Table 9. Procedure Experiment 3

	Minutes
Welcome	2
Demographic questionnaire	1
Training phase	20
Familiarization session (50 trials)	8
PPV questionnaire	3
Experimental session (100 trials)	15
FMV questionnaire + NASA-TLX	6
Debriefing	2
Payment	2
Good bye	1
Total	60

8.2. Hypotheses

Participants´ responding behavior

Hypothesis 1:

It is expected that participants will adapt their response rates to the PPV of each stage so that participants´ response rate to each stage will be significantly different from the other stages.

Regarding the response strategies, participants will exhibit mostly positive extreme responding in response to alarms and mostly negative extreme responding in response to warnings having a PPV below .5. Finally, it is expected that participants will mostly exhibit probability matching in response to orange warnings.

Hypothesis 2:

It is expected that the addition of a new criterion in a LAS will lead to a behavioral differentiation of participants´ response rates. In particular, it is assumed that the cry-wolf effect will be shifted from the warning stage of LAS3 to the yellow-warning stage LAS4. Moreover, participants will comply more with the orange warnings of LAS4 than with warnings of LAS3. A similar effect is expected between the 4- and 5-stage LAS.

This hypothesis makes sense if one looks more precisely at the placement of the criterion chosen to make these LAS stages. LAS4 was created by adding a criterion in the stage of LAS3 having the lowest PPV (orange-warning stage). The goal of this criterion is to reduce the cry-wolf effect by creating 2 stages: the orange-warning stage with a better PPV to which participants will respond more often (reduction of the cry-wolf effect) as well as the yellow-warning stage with a worse PPV in which the cry-wolf effect will still occur. The same logic applies to create LAS5. A criterion was set in the stage of LAS4 that has the lowest PPV (yellow-warning stage) in order to create the yellow-orange-warning stage with a better PPV (reduction of the cry-wolf effect) as well as the yellow-warning stage with a lower PPV (the cry-wolf effect would still occur here).

Participants´ performance

Hypothesis 3:

A main effect of the variable NUMBER OF STAGES on participants' performance is expected. The more stages exist, the better participants' performance will be in terms of percentage of hits, misses, false alarms, and correct rejections.

> **Hypothesis 3a:** Participants´ percentage of hits and correct rejections will increase with the number of stages.
>
> **Hypothesis 3b:** Participants´ percentage of misses and false alarms will decrease with the number of stages.

These hypotheses are a direct consequence of the Hypotheses 1 and 2. Participants adopt a more differentiated behavior towards LAS4 than LAS3. LAS4 users will respond more often to

orange warnings than to yellow warnings of LAS4 whereas LAS3 users did not benefit from any cues to differentiate their behavior. By behaving in this way, LAS4 users will respond more often to true alarms and less often to false alarms. As a consequence LAS4 leads to a higher percentage of hits and correct rejections and a lower percentage of misses and false alarms than LAS3. The same logic applies when comparing LAS4 to LAS5.

Hypothesis 4:

A main effect of NUMBER OF STAGES on participants' performance in the concurrent tasks is expected. A decrease of performance in concurrent tasks is only expected in the LAS5 condition.

Too much specificity (stages) in the alert display might increase workload and thus have an effect on performance in concurrent tasks. However, a decrease of performance is only expected in the LAS5 condition because, as shown by Wiczorek et al. (2014), LAS4 does not lead to a decrease of performance compared to LAS3.

Participants´ trust

Hypothesis 5:

A main effect of NUMBER OF STAGES on participants' trust towards the alarm system is expected. The more stages the LAS has, the more participants will trust this system.

This hypothesis is based on previous literature showing that the more information is provided by a system, the more trust towards this system users have (Lee & See, 2004).

Participants´ workload

Hypothesis 6:

The more stages the LAS has, the higher participants' subjective workload is.

The more information users have to process, the higher is the workload.

8.3. Results

All participants fully completed the experiment. An alpha level of .05 was used for all statistical tests. The mean (*M*) and standard deviations (*SD*) of all analyses are reported in Annex C.

8.3.1. Manipulation check

Table 10 shows real and estimated PPV and NPV values.

Table 10. Real PPV and estimated PPV of the LAS-alerts

Alarm system	Real PPV	Estimated PPV	Real NPV	Estimated NPV
LAS3				
alarm stage	.85	.87		
warning stage	.25	.30		
non-alert stage			.97	.93
LAS4				
alarm stage	.85	.81		
orange-warning stage	.55	.65		
yellow-warning stage	.17	.26	.97	.89
non-alert stage				
LAS5				
alarm stage	.85	.76		
orange-warning stage	.55	.61		
yellow-orange-warning stage	.29	.41		
yellow-warning stage	.11	.29	.97	.83
non-alert stage				

Nine t-tests were conducted to compare the real PPV value of each stage of LAS with the PPV estimated by participants.

Results show that participants overestimated the PPV of the yellow-warning stage of LAS4, *t(*15) = -2.98, *p* = .01. They reported a PPV of .26 instead of .17. Participants also overestimated the PPV of the yellow-warning stage and of the yellow-orange-warning stage of LAS5, *t(*15) = -3.58, *p* = .00 and *t(*15) = -2.15, *p* = .05 respectively. They reported a PPV of .29 instead of .11 and a PPV of .41 instead of .29 for the yellow warning stage and the yellow-orange warning stage of LAS5 respectively.

These deviations are actually not problematic because participants correctly perceived the difference between each stage in LAS. For example, participants using LAS4 correctly estimated that the PPV of the yellow-warning stage *(M* = .26), is lower than the PPV of the orange-warning stage *(M* = .65), t(15) = 3.33, *p* = .01, and that the PPV of the orange-warning stage is lower than the PPV of the alarm stage PPV *(M* = .81), t(15) = 5.99, *p* = .00. This shows that the instructions and the familiarization task were enough to give participants a decent idea of the PPV of each stage.

Three t-tests were conducted to compare the real NPV value to the NPV reported by participants in each condition. The results show that participants significantly underestimated the NPV of all non-alert stages, *p*s <.05.

8.3.2. Alarm Task: Participants´ response rates

We measured participants' response rate to the overall LAS (alarm and warning stages together), to alarms, to each kind of warnings signals, and to non-alerts.

Mean response rates to alerts (alarm stage and warning stages together)

We used a one-way ANOVA with NUMBER OF STAGES (LAS3, LAS4, LAS5) as between factor. Two orthogonal contrasts C1 (-1, 2, -1) and C2 (-1, 0, 1) were used. Participants' mean response rates are displayed in Figure 28. As the graph shows, participants complied more with LAS4 *(M* = 68.46%) than with LAS5 *(M* = 56.66%) and LAS3 *(M* = 57.53%). The statistical analysis does not confirm the trend observed at a descriptive level. Both contrasts are not significant, *p*s > .05. There is no effect of NUMBER OF STAGES on participants' response rates to the overall LAS.

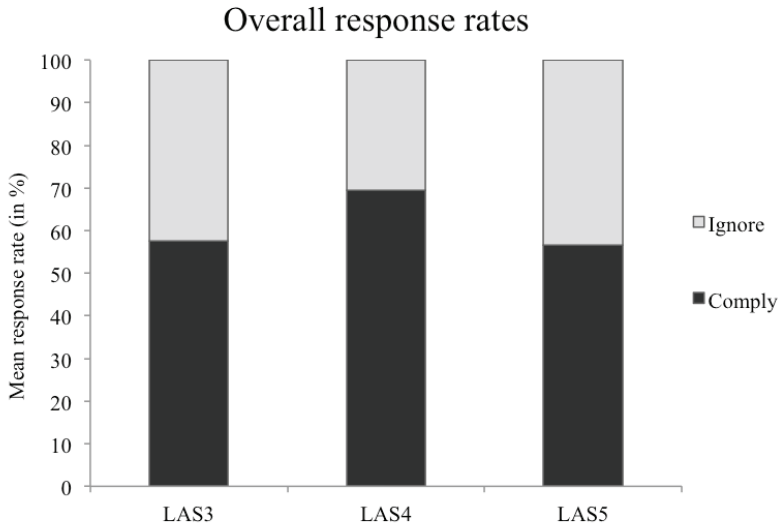


Figure 28. Means of participants´ compliance rates and ignoring rates of LAS-alerts depending on the number of stages

Mean response rates towards LAS3

A t-test with the factor STAGE (alarm, orange-warning) was used for the analysis of participants´ response rates to each stage of LAS3. As one can see in Figure 29, participants complied more with alarms *(M* = 98.56%) than with warnings *(M* = 16.51%) emitted by LAS3. This main effect of STAGE is significant, $F(1,15) = 120.58$, $p = .000$.

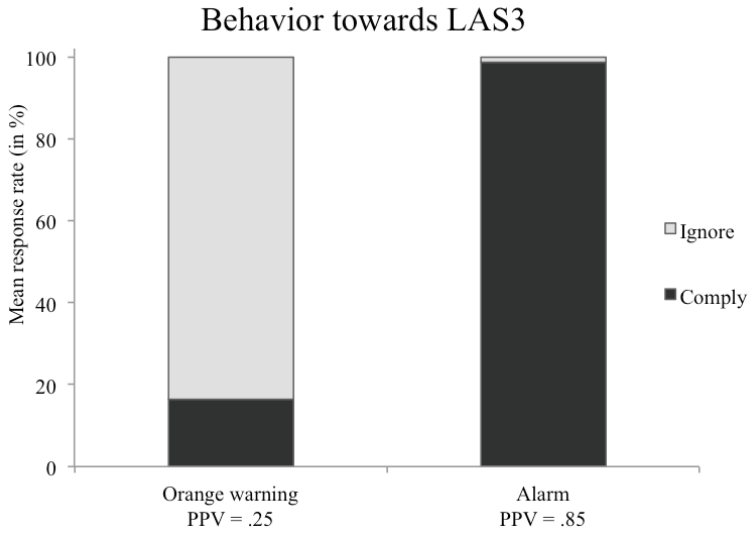


Figure 29. Means of participants' compliance rates and ignoring rates as per LAS stages

Mean response rates towards LAS4

A one-way ANOVA with STAGE (alarm, orange-warning, yellow-warning) as within factor was used for the analysis of the response rate toward LAS4. Participants' mean response rates are displayed in Figure 30. Participants responded to 96.63% of alarms, 97.16% of orange warnings, and 14.58% of yellow warnings.

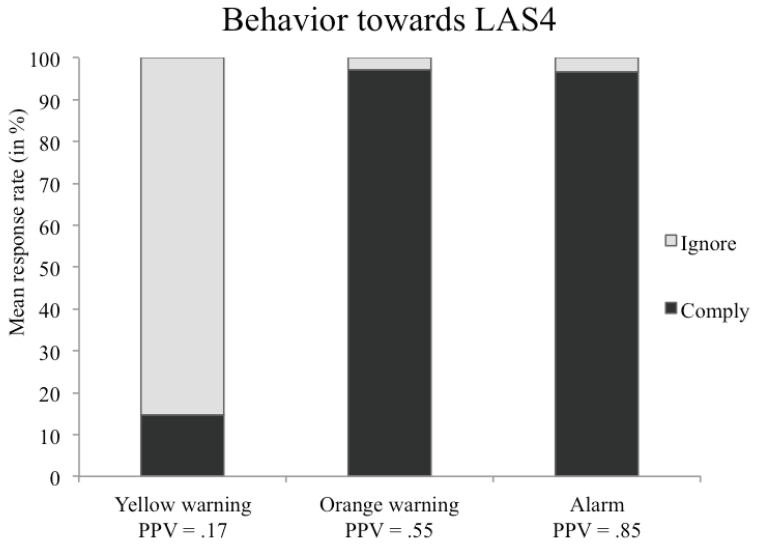


Figure 30. Means of participants' compliance rates and ignoring rates as per LAS stages

According to Hypothesis 1, we used a linear contrast C1(-1, 0, 1) and a quadratic contrast C2 (-1, 2, -1) to test if participants responded depending on the PPV of each stage. The significant linear contrast suggest that participants complied more with the alarm stage than with the yellow-warning stage, $F(1,15) = 111.68$, $p = .00$. The significant quadratic contrast however shows that participants' mean response rate in the orange-warning condition differs from the linear trend C1, $F(1,15) = 111.03$, $p = .00$. This is probably due to the high compliance rate observed in the orange-warning stage.

A more precise analysis of these results actually demonstrates that participants' compliance rates do not differ from each other in the alarm and in the orange-warning condition, $F(1,15) = 0.10$, $p = .76$. Nevertheless, they differ from each other in the yellow-warning and the orange-warning condition, $F(1,15) = 116.64$, $p = .00$.

As expected in Hypothesis 2 participants complied significantly more with the orange-warning stage of LAS4 (97.16%) than with the orange-warning stage of LAS3 (16.51%), $F(1, 30) = 107.91$, $p = .00$. Also, one additional criterion in the orange-warning stage of LAS3 led to a significant reduction of the cry-wolf effect in the orange-warning stage of LAS4.

One can note here that participants' response rate to orange warnings *(M* = 97.16%) is extremely high given that the PPV of this stage is .55. Such high response rate is usually expected for high-PPV stages (i.e., over .7).

Mean response rates towards LAS5

A one-way ANOVA with STAGE (alarm, orange-warning, yellow-orange-warning, yellow-warning) as within factor was used for the analysis of the response rate to LAS5. As expected, the pattern of results is linear. Participants' mean response rates are displayed in Figure 31. Participants responded to 95.19% of alarms, to 86.36% of orange warnings, to 35.71% of yellow-orange warnings, and to 9.37% of yellow warnings.

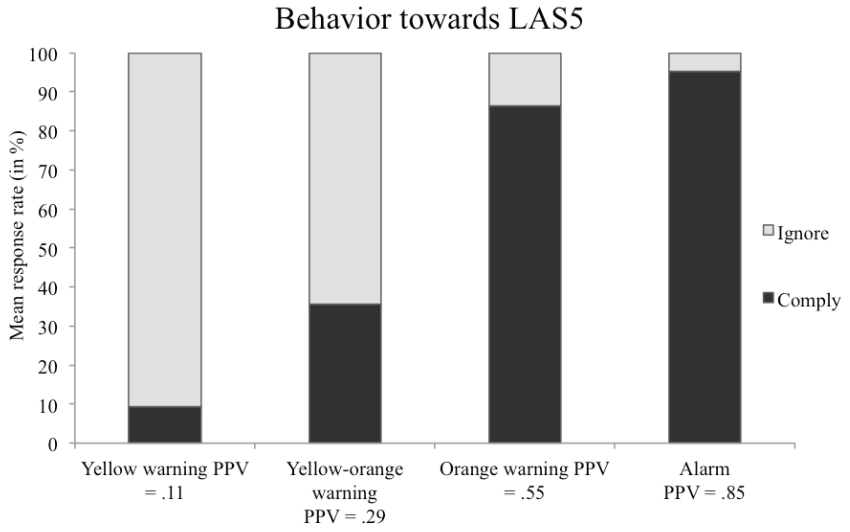


Figure 31. Means of participants' compliance rates and ignoring rates as per LAS stages

In line with Hypothesis 1, we used a linear contrast C1(-3, -1, 1, 3), a quadratic contrast C2 (1, -1, 1, -1), and a cubic contrast (-1, 3, -3, 1) to test if participants responded depending on the PPV of each stage. The linear contrast is significant, $F(1,15) = 120.34$, $p = .00$, as well as

the cubic contrast, $F(1,15) = 5.31$, $p = .04$. This means that the pattern of results is not completely linear, as we expected. The high compliance rate obtained in the orange-warning stage is probably responsible for the significance of the cubic contrast. This is confirmed by the fact that participants' compliance rate does not differ in the orange-warning stage and the alarm stage, $F(1,15) = 2.46$, $p = .14$.

Even though participants complied more with the yellow-orange-warning stage of LAS5 (M = 35.71%) than with the yellow-warning stage of LAS4 (M = 14.58%), this effect is not significant $F(1,30) = 2.65$, $p = .11$. The addition of one more criterion to the orange-warning stage of LAS4 did not lead to any significant reduction of the cry-wolf effect in the yellow-orange-warning stage of LAS5 even though this was expected in Hypothesis 2.

One can note here that the response rate to orange warnings (M = 86.36%) is extremely high given that the PPV of this stage is .55. Such high response rate is usually expected for high-PPV stages (i.e., above .7).

Mean response rates to non-alerts

A one-way ANOVA with NUMBER OF STAGES (LAS3, LAS4, LAS5) as between factor was used for the analysis of participants' response rate to non-alerts. There is no effect of NUMBER OF STAGES on participants' response rate to non-alerts, $F(2, 47) = 0.253$, $p = .78$.

8.3.3. Participants´ response strategies

Response strategies towards LAS3

Figure 32 shows the response strategies adopted by participants in the 3-stage LAS condition.

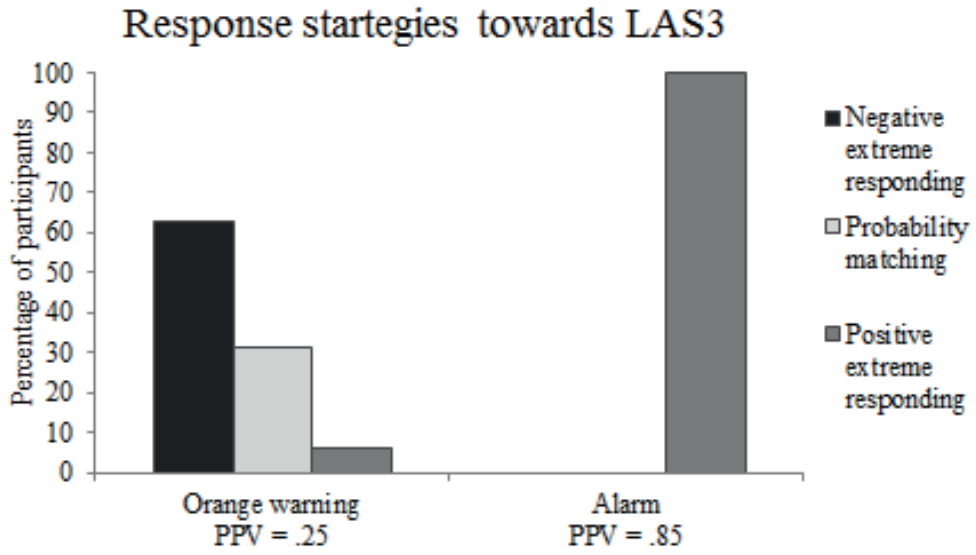


Figure 32. Percentage of participants exhibiting negative extreme responding, positive extreme responding, and probability matching when interacting with the 3-stage LAS (as defined in Part 3.3.2.)

All participants (n = 16) adopted an extreme positive responding strategy in response to alarms. Negative extreme responding turned out to be the dominant strategy adopted by participants for the orange-warning stage of LAS3 (n = 10), $\chi^2(2, n = 16) = 7.63$, $p = .02$. Five participants exhibited a sort of probability matching behavior and one participant adopted a positive extreme responding toward the orange-warning stage of LAS3.

Response strategies towards LAS4

Figure 33 shows the response strategies adopted by participants for the 4-stage LAS.

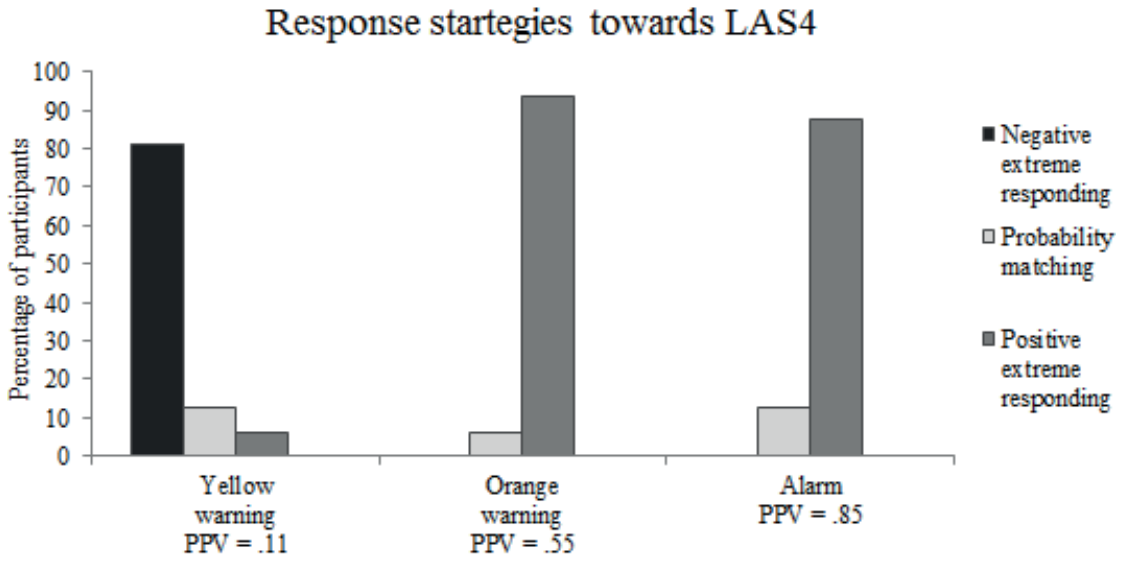


Figure 33. Percentage of participants exhibiting negative extreme responding, positive extreme responding, and probability matching when interacting with the 4-stage LAS (as defined in Part 3.3.2.)

The dominant strategy adopted by participants in response to alarms was positive extreme responding (n = 14), $\chi^2(1, n = 16) = 9.00$, $p = .00$. The same pattern of results is observed for the orange-warning stage where 15 participants adopted a positive extreme responding strategy, $\chi^2(1, n = 16) = 12.25$, $p = .00$. In the yellow-warning stage participants exhibited mostly a negative extreme responding (n = 13), $\chi^2(2, n = 16) = 16.63$, $p = .00$. Two participants exhibited a sort of probability matching and one participant exhibited a positive extreme responding towards the yellow-warning stage of LAS4.

Response strategies towards LAS5

Figure 34 shows participants´ response strategies in the 5-stage LAS condition.

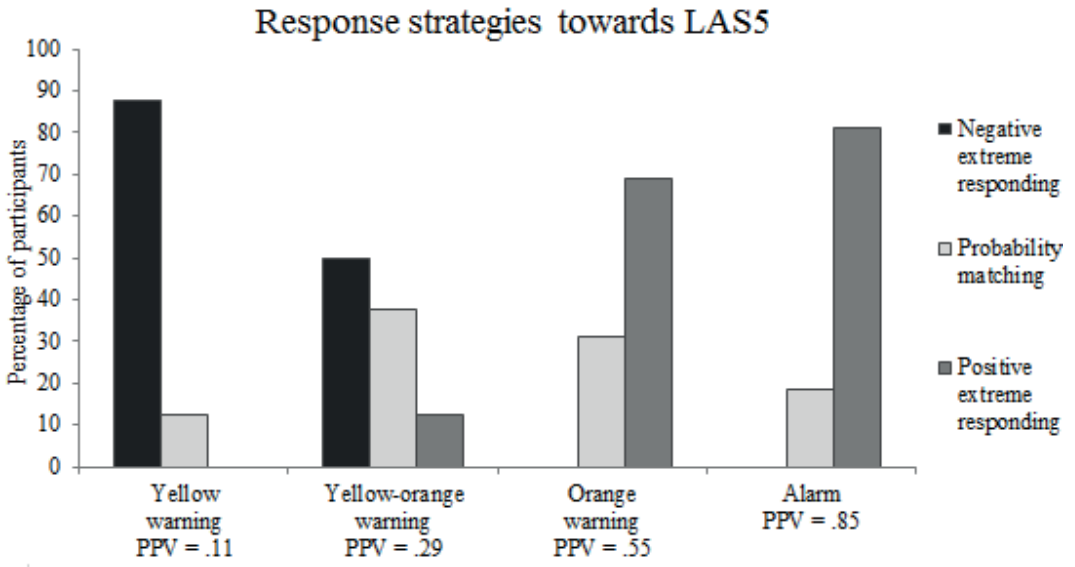


Figure 34. Percentage of participants exhibiting negative extreme responding, positive extreme responding and probability matching when interacting with the 5-stage LAS (as defined in Part 3.3.2.)

The dominant strategy adopted by participants in response to alarms was positive extreme responding (n = 13), $\chi^2(1, n = 16) = 6.25$, $p = .01$. In the orange-warning stage of LAS5, most of participants (n = 11) exhibited positive extreme responding and a minority (n = 5) probability matching. This difference is not significant, $\chi^2(1, n = 16) = 2.25$, $p = .13$. The dominant strategy adopted by participants in the yellow-orange-warning stage of LAS5 was negative extreme responding (n = 8) followed by probability matching (n = 6) and positive extreme responding (n = 2). These differences are not significant, $\chi^2(2, n = 16) = 3.50$, $p = .17$. Finally, the dominant strategy adopted by participants in response to yellow warnings of LAS5 was negative extreme responding (n = 13) while a minority of participants (n = 2) exhibited probability matching. This difference is significant, $\chi^2(1, n = 16) = 9.00$, $p = .00$.

Response strategies for non-alerts

All participants (n = 48) adopted a negative extreme responding strategy in response to non-alerts.

8.3.4. Participants´ performance in the alarm task

All analyses about participants' performance in the alarm task and in the concurrent tasks were performed using a one-way ANOVA with NUMBER OF STAGES (LAS3, LAS4, LAS5) as between factor. The same two orthogonal contrasts were used for all analyses: C1 (2, -1, -1) and C2 (0, -1, 1). The first contrast C1 tests if LAS3 differs from LAS4 and LAS5 while the second contrast C2 tests if LAS4 and LAS5 differ from each other.

Participants' performance in the alarm task was assessed using four dependent variables: the percentage of hits, the percentage of misses, the percentage of false alarms, and the percentage of correct rejections that participants made when the alarm system produced either a warning or an alarm. Participants' performance in the non-alert stage was not taken into account.

Participants' percentage of hits

Results about participants' percentage of hits are displayed in Figure 35.

Two participants were excluded from this analysis based on their outlying SDR and Cook values. These two participants produced very high percentages of hits *(M* = 36.36%) in comparison to other participants *(M* = 23.72%).

Contrary to our expectations, results on participants´ percentage of hits do not exhibit a linear trend. As expected, participants using LAS3 produced significantly less hits *(M* = 17.64) than participants using LAS4 *(M* = 26.33). Nonetheless, participants using LAS5 *(M* = 26.42) did not produce more hits than participants using LAS4. The first contrast C1 confirms that participants produced significantly less hits with LAS3 than with LAS4 and with LAS5,

F(1,44) = 52.91, *p* = .00. The second contrast C2 shows that participants' percentage of hits does not differ between LAS4 and LAS5, *F*(1,44) = 0.01, *p* = .94 (C2).

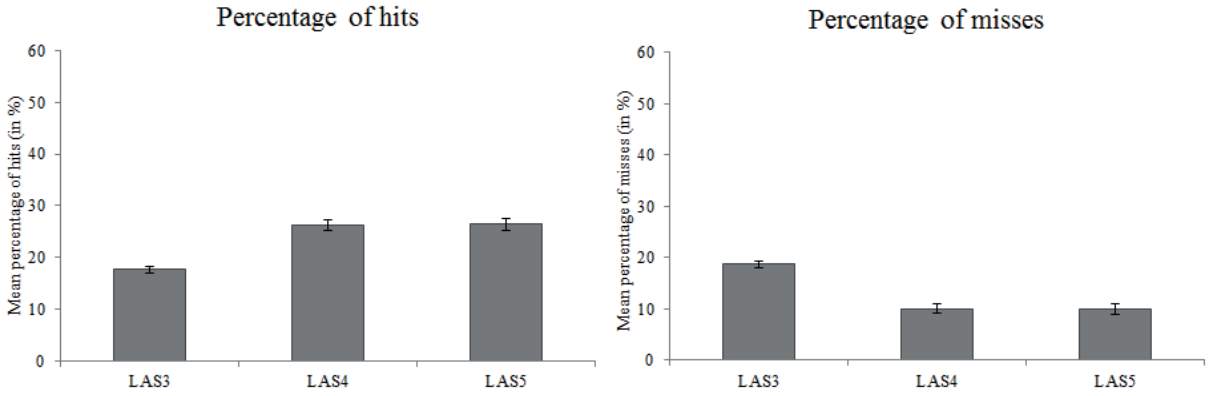


Figure 35. Means and mean standard deviations of participants' percentage of hits (left panel) and misses (right panel) in the alarm task depending on the number of stages

Participants' percentage of misses

Results about participants' percentage of misses are displayed in Figure 35. The two participants excluded from the analysis on percentage of hits were also excluded from the analysis on percentage of misses based on their SDR and Cook values. Both these participants produced no misses *(M* = 0%).

Regarding participants´ percentage of misses, results do not exhibit a linear trend, which is contrary to our expectations. As expected, participants using LAS3 produced significantly more misses *(M* = 18.72) than participants using LAS4 *(M* = 10.04). On the other side participants using LAS5 *(M* = 9.94) did not produce less misses than participants using LAS4. The first contrast C1 confirms that participants produced significantly more misses with LAS3 than LAS4 and LAS5, *F*(1,44) = 52.91, *p* = .00. The second contrast C2 shows that participants' percentage of hits does not differ between LAS4 and LAS5, *F*(1,44) = 0.01, *p* = .94 (C2).

Participants' percentage of false alarms

Results about participants' percentage of false alarms are displayed in Figure 36. One participant was excluded from this analysis based on his SDR and Cook value. This participant produced a very high percentage of false alarms *(M* = 63.63%) in comparison to the other participants *(M* = 15.38%).

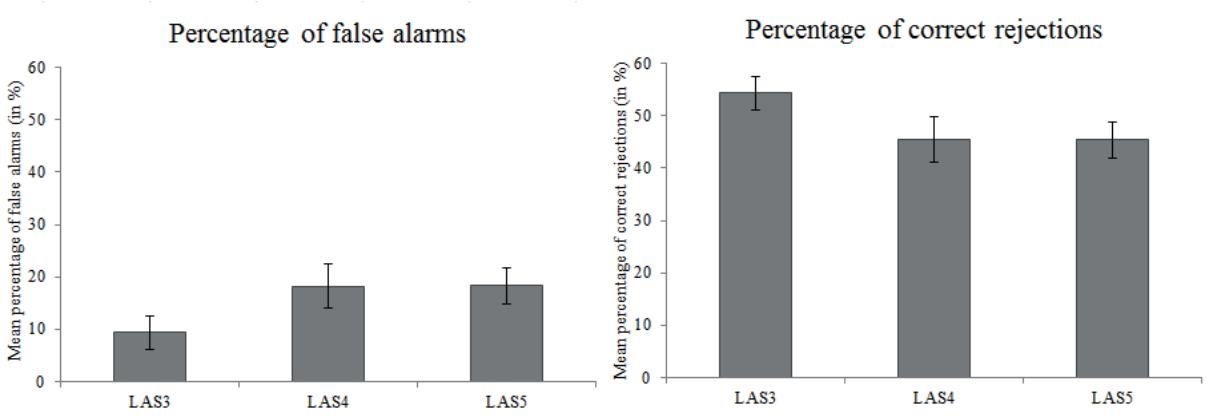


Figure 36. Means and mean standard deviations of participants' percentage of false alarms (left panel) and correct rejections (right panel) in the alarm task depending on the number of stages

Regarding participants´ percentage of false alarms, results do not exhibit a linear trend. Moreover, contrary to our expectations, LAS3 users exhibit the best performance. They produced significantly fewer false alarms *(M* = 9.29) than LAS4 users *(M* = 18.18) and LAS5 users *(M* = 18.28). The first contrast C1 confirms that participants produced significantly fewer false alarms with LAS3 than LAS4 and LAS5, $F(1,45) = 3.84$, $p = .05$. The second contrast C2 shows that participants' percentage of false alarms does not differ between LAS4 and LAS5, $F(1,45) = 0.00$, $p = .99$ (C2).

Participants' percentage of correct rejections

Results about participants' percentage of correct rejections are displayed in Figure 36. The participant excluded from the analysis on percentage of false alarms was also excluded from this analysis based on his SDR and Cook values.

Results do not exhibit a linear trend. Participants using LAS3 produced contrary to our expectations significantly more correct rejections *(M* = 54.34) than participants using LAS4 *(M* = 45.45) and participants using LAS5 *(M* = 45.36). The first contrast C1 confirms that participants produced significantly more correct rejections with LAS3 than LAS4 and LAS5, *F(*1,45) = 52.91, *p* = .00. The second contrast C2 shows that participants' percentage of correct rejections did not differ between LAS4 and LAS5, *F(*1,45) = 0.01, *p* = .94 (C2).

8.3.5. Participants´ performance in concurrent tasks

All analyses about participants' performance in concurrent tasks were performed using a one-way ANOVA with NUMBER OF STAGES (LAS3, LAS4, LAS5) as between factor. The same two orthogonal contrasts were used for all these analyses, namely C1 (1, 1, -2) and C2 (-1, 1, 0).

Performance in the Coolant Exchange Task (CET)

It was expected that participants in the LAS5 condition would exhibit worse performance in the Coolant Exchange Task (CET) than participants in the LAS3 and LAS4 condition (i.e., they would complete less refilling cycles). Our results do not support this hypothesis. The main effect of NUMBER OF STAGES is not significant, *F(*2, 45)= 0.06, *p* =.943.

Performance in the Resource Ordering Task (ROT)

It was expected that participants in the LAS5 condition would have worse performance in the ROT than participants in LAS3 and LAS4 conditions (i.e., they would send less orders and less sent orders would be correct). Based on his SDR and Cook values, one participant was excluded from the analysis on the number of correctly sent orders in ROT. This participant sent extremely more correct orders (285 orders) than the other participants *(M* = 114.64). Participants in the LAS5 condition sent less correct orders *(M* = 105.53) than participants in LAS3 *(M* = 118.38) and LAS5 conditions *(M* = 119.44). However the main effect of NUMBER OF STAGES on the amount of correct orders sent is not statistically significant, *F(*2,44)= 0.41, p=.66.

Three participants were excluded from the analysis on the variable 'percentage of orders participants responded to in the ROT' based on their SDR and Cook values. They all sent fewer orders (less than 75%) than the other participants *(M* = 96.22%). There is no main effect of NUMBER OF STAGES on the percentage of appeared orders participants responded to, *F(*2,42) = 0.475, *p* = .63.

8.3.6. Participants' subjective ratings

A one-way ANOVA with NUMBER OF STAGES (LAS3, LAS4, LAS5) as between factor was used for the analysis of participants' workload and trust ratings.

Workload

The main effect of NUMBER OF STAGES on participants' workload ratings is not significant, *F(*2,45)= 1.05, *p* =.36.

Trust

The main effect of NUMBER OF STAGES on participants' trust ratings is not significant, *F(*2, 45) = 0.44, *p* =.65.

A main effect of NUMBER OF STAGES has been found on the subscale Intention, *F(*2, 45) = 3.180, p=.051. LAS5 users reported more trust *(M* = 3.55) than LAS4 users *(M* = 3.16) and LAS3 users *(M* = 3.27) on this subscale.

8.4. Discussion

Experiment 3 aims to investigate the number of stages of likelihood alarm systems that provide the optimal specificity of information for human performance in interaction with LAS. We specifically investigated the effect of three different LAS on responding behavior, performance, and workload. The LASs differed with respect to the number of stages they consist of.

Participants correctly perceived the relative reliability of each stage. However, participants overestimated the PPVs of all low-PPV stages (PPV < .5) of LAS4 and LAS5 and they underestimated the NPV of alarm systems. Similar estimations also occurred in Experiment 1 and 2.

Participants did not precisely adapt their responding behavior to the PPV of each stage. Our findings only partially confirmed Hypothesis 1, in which a linear pattern between PPV and participants´ response rate was expected. Interestingly, participants responding behavior shows a kind of dichotomization depending if the PPV of the alert stage is below or above .5. Participants tend to comply with almost all alerts emitted by stages having a PPV above .5. For example, participants complied with more than 85% of the orange warnings emitted by LAS4 and LAS5 even though the PPV of this stage was .55. However, Experiment 3 shows that participants tend to differentiate their responding behavior more clearly in stages having a PPV bellow .5. Similar responding behaviors toward many-stage LASs were reported by Wiczorek et al. (2014). These results raise two questions. First, why did participants dichotomize their responding behavior? Second, why did participants dichotomize their behavior in such a way? The next paragraphs aim to answer these questions.

First, such dichotomization is a rational strategy considering that the more stimulus-response alternatives there are, the longer it takes participants to make a choice (Hick, 1952; Hyman, 1953). By adopting the same response strategy towards all stages having a PPV above .5, participants reduced the number of 'stimulus-response' alternatives. At a subjective level, participants probably merged stages with PPVs above .5 into a single stage and automatically comply to alerts from this stage.

The second question that arises from our results concerns the way participants dichotomize their responding behavior. Why participants differentiate their behavior for alerts stages having a PPV < .5 and did not merge these stages in order to facilitate decision? One possible explanation might be regarding the subjective pay-offs that participants attributed to the different performance outcomes. Even though the experiment's instructions stated that a false alarm and a miss represent exactly the same cost, participants might have still perceived the consequences of misses as more serious than those of false alarms. Some participants made such statements during the debriefing session. The experiment's cover story is likely to have

supported this effect as participants acted as operators in a chemical plant. It is very probable that participants actually evaluated that sending an improper container to clients (miss) has more serious consequences than sending a proper container back to the factory. Another possible explanation for this effect refers to an effect called 'commission bias' (Guenzler & Manzey, 2013). In this case, it might have been more tolerable for participants to adopt extreme response strategies towards all stage having a PPV above .5 rather than adopting an extreme response strategy towards all stages having a PPV bellow .5. This is because participants might have preferred committing an error after having initiated an action rather than committing an error after having ignored an alert.

Looking at these results, one can already conclude that there is no need to have a few stages with a PPV above .5. It is better to make one alarm stage and make sure its PPV still leads to positive extreme responding strategies. This reduces the complexity of the decision-making task by reducing the numbers of stimuli. As a consequence it might improve participants' response time and thus performance, particularly in the ongoing concurrent tasks.

As per our second hypothesis, we expected that the addition of a new criterion would systematically lead participants to differentiate their behavior. This hypothesis was confirmed only when comparing LAS3 to LAS4. This effect also emerged when comparing LAS4 to LAS5 but only at a descriptive level. The addition of a new criterion in the yellow-warning stage of LAS4 led to more compliance with warnings emitted by the yellow-orange-warning stage of LAS5. However, this effect is not significant.

The number of stages does not affect participants´ overall response rates to alerts. Wiczorek et al (2014) already reported that participants do not reduce their overall response rates between LAS3 and LAS4 and our results show that this finding is equally true for LAS5.

Regarding participants performance in the alarm task (Hypothesis 3), a linear trend did not emerge as expected. With respect to the percentage of hits and misses participants had better performances with the LAS4 and LAS5 than with the LAS3. Contrary to our expectations no significant differences emerged between LAS4 and LAS5.This is probably due to the fact that even though LAS5 provides more differentiated information, participants actually adopted more rational behavior in response to LAS4 than in response to LAS5. Almost no participant

exhibited probability matching for LAS4 whereas this was quite often exhibited for LAS5; this was especially true in stages having a PPV below .5.

Against our expectations (Hypotheis 3b), with respect to the percentage of false alarms and correct rejections, participants had lower performances with LAS4 and LAS5 than with LAS3. This finding stands in contradiction with results reported by Wiczorek et al. (2014) that demonstrated that participants produce fewer false alarms with LAS4 than with LAS3. The high response rate to orange warnings in the LAS4 and LAS5 might explain these results. By complying with more than 93% of warnings having a PPV of .55, participants produced a great amount of false alarms in comparison to participants in the LAS3 condition who mainly ignored the .25 PPV warnings and produced mostly correct rejections. Nonetheless, the percentage of hits is a better indicator of performance than the percentage of false alarms since most alarms systems are used in environments where misses are more costly than false alarms.

At a first look, it seems that a 4-stage LAS incorporates the maximum degree of complexity above which performance does not improve further. Is this always the case? Can we generalize this finding to all 4-stage LASs? The answer is probably no. It turns out that with the criterion and sensitivity settings chosen in this experiment, LAS4 led operators to adopt the best strategies so as to reach optimal performance. This is partly due to the fact that participants exhibit positive extreme responding in response to the orange-warning stage and alarm stage of LAS. Participants behaved as if the alarm stage started from the middle criterion. One could imagine that if the orange-warning stage and the alarm stage of LAS4 would have been merged into one single alarm stage to make a 3-stage LAS, then this 3-stage LAS might have also lead to optimal performance. One of the things that can be learned from this experiment is that participants interacting with LAS of more than 4-stage seem to respond to a few stages in the same way; probably in order to reduce the complexity of the decision task. In terms of behavior, the results we found for the 4-stage LAS in Experiment 3 are probably similar to what would be found with either a 3-stage LAS, in which the orange and red stage would form only one alarm stage, or with a BAS, in which the yellow-warning stage would be integrated to the non-alert stage and the orange-warning stage would be integrated to the alarm stage. More research is needed to really understand why participants adopt these response strategies. Such insights could guide engineers in the design of LAS.

Participants' performance in concurrent tasks does not decrease with the use of LAS5. Hence, Hypothesis 4 was not confirmed. This is probably due to the fact that participants' workload did not increase with the greater amount of information provided by LAS5. Indeed regarding participants' workload no difference between the three different LAS was observed. Consequently, we could not confirm Hypothesis 6. Moreover, participants´ overall response rate to alerts is the same for the three different LAS. It would however be interesting to know if a higher number of stages than five affect the workload since alarm systems having more than five stages are sometimes used in real-world settings.

It was expected that the more stages LAS has, the more participants trust the system (Hypothesis 5). The results show that LAS5 users reported more trust on the subscale intention than LAS3 and LAS4 users. LAS5 users might have interpreted the biggest amount of likelihood information as an intention from the alarm system to help.

In conclusion this study suggests that 4-stage LAS provide the optimal degree of specificity and that a higher degree of specificity does not further improve performance. However some questions remain. For example, are these results caused by the specificity of information provided by LAS4 in comparison to LAS3, or by the characteristics of the LAS4 criterion design? More research is needed to better understand to what extent 3-stage LAS, using different settings than in Experiment 3, could lead to the same performance as LAS4.

9. General discussion

The objective of the present work was to get a better understanding of the conditions in which likelihood alarm systems should be used, as well as, investigating what LAS characteristics lead to the best possible performance. LAS has been considered as a promising alternative to binary alarm systems in order to reduce the cry-wolf effect (Sorkin et al., 1988). Contrary to BAS, LAS emits different types of alerts depending on the likelihood that a critical event occurs. Two main benefits of LAS over BAS had been suggested in the literature. First, LAS might better support operators to differentiate between true alarms and false alarms and therefore make their decisions more accurate. Second, LAS might allow for a better allocation of operators' attention between the alarm task and other concurrent tasks they are responsible for. As a consequence, it has been suggested that operators' performance should improve with the implementation of LAS.

Despite these assumptions, studies investigating the benefit of LAS over BAS have provided very inconsistent results. From the review of former studies investigating LAS (Part 4.3) it was concluded that some moderators exist, which might affect the potential benefit that LAS has over BAS. More specifically, it was proposed that the base rate of critical events and the cost and ease of consulting to Alarm Validity Information (AVI) might be among the most important moderators.

One goal of this PhD Dissertation has been to investigate these two potential moderators. For this purpose, Experiment 1 and Experiment 2 were conducted to investigate the effect of the base rate of critical events and the cost of consulting AVI on BAS and LAS users. A second goal has been to identify which LAS design characteristics best support the operators' decision-making. To this end, the effect of the number of stages in LAS (Experiment 3) has been examined. These studies were conducted using a laboratory multitask paradigm, which simulates in a simplified way typical tasks of operators in a control room of a chemical plant. The results from these three studies will be discussed in the following section and should provide important theoretical and practical implications for researchers and practitioners.

9.1. Summary and discussion of the results

A first conclusion from our results is that participants use the likelihood information provided by LAS. They take the different stages of the LAS into account and adopt their behavior accordingly. When AVI is not available to participants (as in Experiment 1 and 3), participants complied more with the alarms than with the warnings emitted by the 3-stage LAS. However, when LAS is composed of more than three stages, participants differentiate their behaviors only towards warnings that have a PPV below .5 and not for higher stages. In the latter case they systematically exhibit extreme responding strategies. Participants also differentiate their responding behavior when AVI is available. In Experiment 2 participants clearly cross-checked more and complied less with LAS-warnings than with LAS-alarms, whatever the cost of consulting AVI.

These results suggest that adding a new stage in an alarm prompts the user to take it into account and adapt his behavior to it. This also confirm results on BAS from Getty et al. (1995), and Manzey et al. (2014), as well as results on LAS by Ragsdale (2012), Wiczorek (2012) and Wiczorek et al. (2014) that show that participants adapt their behavior to the PPV of LAS.

Interestingly, no benefit of LAS over BAS could be shown with regards to the number of critical events correctly detected (hits). This is despite clear evidence that participants used the information provided by the LAS to adapt their responding behavior accordingly. The results even show a benefit of high-PPV BAS over LAS in Experiment 1.

Results from Experiment 1 clearly highlight one issue related to the use of LAS: By providing more stages LAS induce a probability matching behavior, which in some cases can decrease performance in comparison to BAS. This is one of the most important contributions of this PhD Dissertation. Results from Experiment 1 illustrate this. BAS users complied with almost all high-PPV alarms and therefore did almost not miss any critical event. However, LAS users only complied with LAS-alarms and ignored a significant number of LAS-warnings. As a consequence, LAS users missed some critical events announced by the warnings.

This should make researcher reconsider the way they describe the benefit of LAS over BAS. The benefit of LAS over BAS does not reside exclusively in the fact that LAS provide more information to the user. LAS only present a benefit over BAS if the extra likelihood information induces a probability matching behavior, which is desirable (e.g., when the BAS is responsible for a cry-wolf effect). This might explain why previous studies failed to find a benefit of LAS over high-PPV BAS (Swanson, 2010; Wickens & Colcombe, 2007). It is very important to inform practitioners about the reasons why LAS lead to better performance than BAS so that LAS is not systematically chosen over BAS.

Another important result of this PhD Dissertation that emerged out of Experiment 2 is that likelihood information becomes useless in a setting in which AVI is provided to participants. BAS users were inclined to cross-check all alarms, which led to a close to perfect performance and the missing of almost none critical events. This shows that in such a case the implementation of LAS does not further improve performance. This is true regardless of the cost of consulting AVI. Nonetheless, the generalizability of such a conclusion is limited since a few authors have shown a benefit of LAS over BAS using an experimental paradigm with access to AVI (Bustamante & Bliss, 2005; Clark & Bustamante, 2008; Bustamante & Clark, 2010). Clearly then, future research is required to understand more precisely in which conditions LAS lead to better performance than BAS, in settings where AVI is provided to participants.

The results from this PhD Dissertation only show a benefit of LAS over BAS regarding the number of false alarms produced by participants and only in settings without AVI (Experiment 1). Nonetheless, LAS led to worse performance than BAS in terms of number of false alarms produced by participants when AVI was available. These results raise the question why LAS users demonstrate a better performance than BAS users when AVI is not available (Experiment 1) whereas BAS users demonstrate a better performance than LAS users when AVI is available (Experiment 2).

When AVI is not available (Experiment 1), whereas the adoption of a sort of probability matching toward warnings leads participants to miss critical events, it at the same time helps them avoiding false alarms. LAS presents a benefit over low-PPV BAS in this regard. When AVI is available (Experiment 2), the stages of LAS provoke participants to reduce their cross-

checking of LAS-alarms in favor of complying. Since alarms were not 100% reliable, LAS users ended-up producing more false alarms than BAS users.

The results from Experiments 1 and 2 do not provide evidence that LAS improve concurrent task performance over BAS. This is surprising because results from Experiment 1 and 2 show clear evidence that LAS users allocate less attention to the alarm task than BAS users. This is the case whatever the presence of AVI. As a consequence, LAS users should have more time and free resources to allocate to the concurrent tasks than BAS users.

Of the few researchers that looked at participants' performance in concurrent tasks most have failed to find a benefit of LAS over BAS regarding participants' performance in such tasks (Sorkin et al., 1988; Wickens & Colcombe, 2007; Wiczorek et al., 2014). In this regard only Wiczorek & Manzey (2014) reported a benefit of LAS over BAS. Comparisons between all these studies are difficult because only Wiczorek & Manzey (2014) and Wiczorek et al. (2014) reported participants' responding behavior. It seems however that the absence of results of Wiczorek et al. (2014) is explained by the fact that participants' overall response rate does not differ between experimental conditions. Difference in the numbers of concurrent tasks might explain why Experiment 1 and 2 could not replicate the benefit reported by Wiczorek and Manzey (2014). In Experiment 1 and 2, the free time and resources that LAS users benefited had to be shared between two concurrent tasks whereas participants in the Wiczorek & Manzey's experiment (2014) could concentrate their entire free resources on one task only. The shared time and resources in our experiments might have been to little to reach a significant improvement of participants' performance in each concurrent task.

Experiment 3 investigated the influence of the number of stages in LAS on participants´ performance. The main conclusion is that more than 4 stages do not seem to improve performance further. Regarding participants´ performance in the concurrent tasks, no difference has been found between 3-stage, 4-stage, and 5-stage LAS.

Participants´ subjective experience was also of interest in this PhD Dissertation. A subjective measure assessed in all experiments was trust. Trust is a very important concept in human-automation interaction since it is an important predictor of operators´ behavior (Lee & Moray, 1994; Lee & See, 2004). The results show that participants trust LAS more than BAS. More

precisely, LAS users reported more trust than BAS users on the subscales reliability and use. This finding is particularly interesting given that the overall reliability of BAS and LAS is exactly the same in our studies. This means that trust does not depend only on the overall reliability of the alarm system. One possible explanation of this effect is that false alarms affect participants´ trust much more than false warnings and that participants might not interpret false warnings emitted by LAS as being errors.

This is an important result as one could speculate that the level of trust can affect participants' behavior; for example regarding their rapidity of response to alarms (Smith, 2008). Operators might require less time to take a decision with LAS than with BAS because they trust LAS outputs more. If additional information is readily available to operators (e.g., a CT scan for a radiologist), this might be particularly important because it could reduce the search time that operators invest before complying with the system and thus liberate more time for the concurrent task. More research is needed to confirm this statement since former results from Wickens and Colcombe (2007) show the opposite trend: LAS users have slower response time than BAS users. The authors interpreted this as the effect of the greater information content of LAS, which takes longer to process in comparison to BAS. However, it is very likely that BAS users and LAS users in the Wickens and Colcombe´s experiment (2007) do not differ regarding their trust ratings. BAS had a very high PPV, over .7, so participants might have highly trusted both BAS and LAS. Hence, trust levels could not have affected performance in this experiment.

One recurrent problem along this work has been the difficulty to compare our results with existing studies. This is (1) because the extend to which LAS has been investigated is limited, and (2) because the variety of experimental settings and measurements make it difficult to compare studies. Most studies used complex measurements of participants´ performance, calculated with two or more parameters (see Part 4.3.4. for some examples). Such performance indicators do not allow gaining a precise understanding of participants' performance with alarm systems. Results of Experiment 1 illustrate this well. By using a calculation of participants´ wrong decision (misses and false alarms), it could be concluded that LAS improves performance over BAS in the low-PPV condition. However, separate analyses of participants´ percentage of misses and false alarms show that this benefit emerged

for false alarms only. Future research on LAS might want to rather use direct measurements of performance in order to gain a more precise understanding the real benefits of LAS.

Moreover, results on participants' behavior were not easily comparable to other studies. Three studies only reported participants' response rates for each stage of LAS (Ragsdale, 2012; Wiczorek & Manzey; Wiczorek & al., 2014). The other studies used measurements, which unfortunately do not provide a precise understanding of participants' behavior for each alert stage. Moreover, it is important to consider participants´ behavior both as a group (mean response rate) as well as at an individual level (participants´ response strategy). This will help to identify any sub-groups among participants and help to speculate on participants´ cognitive processes.

Based on these observations, we would recommend future researchers on LAS to carefully select measurements that allow gaining a more precise and deeper understanding of participants´ performance and behavior.

9.2. Limitations of these studies

All experiments presented as part of this PhD Dissertation are laboratory experiments. This might be perceived as an advantage or disadvantage depending on whether one considers the internal or external validity of research. On the one hand, there are numerous advantages of conducting laboratory experiments with respect to the internal validity of findings. The biggest one is the access to a controlled environment, which reduces the number of confounds that could affect results and make them difficult to interpret. Nonetheless, laboratory experiments are much more simplistic than real-world settings and therefore narrow their external validity. For example, participants in our experiments had to click a button to comply with alarms whereas in real-world settings this usually involves much more work. Consequently, it cannot be excluded that the low amount of effort required to repair containers might have affected participants' response behavior.

Another limitation of these studies is the sample used. Participants were mostly students. For economic reasons it was not possible to investigate real operators' behavior. It cannot be

excluded that this group might have exhibited more heterogeneous response strategies than operators. Participants' strategies might have been influence by their ability to perform the different tasks or by their motivation to get monetary rewards.

Another limitation is that the pay-off used in our experiments does not represent the pay-off found in most of real-world settings. In our experiments false alarms and misses have the same cost. Participants loose two points each time they produce a miss or a false alarm. This pay-off is neutral and should not influence participants' behavior. In other words, the expected value associated with each participants´ behavior (complying with the alarm system vs. ignoring the alarm system) depends only on the PPV of the alarm system. In such conditions the most rationale behavior for participants to maximize performance would be to comply with alerts having a PPV higher than .5 and to ignore alerts having a PPV lower than .5. In most real-world settings the cost of a miss is higher than the cost of a false alarm. This pay-off affects the expected value associated with each action. If the pay-off in our experiment would have been a bit closer to real-world settings, for example -2 cents for misses and -1 cent for false alarms, the expected value associated to each participants´ decisions would have changed. In this case complying with alerts having a PPV of .35 (like the BAS used in Experiment 1) would have led to the best performance outcomes. However, even if the neutral pay-off chosen for our experiments does not correspond to most of real-world settings, it was still the most optimal choice to investigate the effect of PPV on participants' responding behavior. A pay-off, which is not neutral, would have influenced participants' responding behavior and hence made it difficult to interpret any effects (for example of the PPV) observed on participants' behavior. Moreover, the same monetary pay-off has already been used in former studies investigating LAS with the experimental paradigm M-TOPS (Wiczorek, 2012; Wiczorek & Manzey, 2014; Wiczorek et al., 2014). As a consequence, a direct comparison between our results and results of these experiments is possible. Finally, such a pay-off has been rarely used in previous research, yet is relevant for settings in which false alarms and misses incur the same cost.

The last limitation concerns the incongruence between the design chosen for the alarm systems and the neutral pay-off used in the experiments. The alarm systems in our experiments are clearly false alarm-prone. Such alarm systems are implemented in environments in which the cost of missing a critical event is higher that the cost of producing

a false alarm. However, the payoff used in our experiments does not represent such an environment. It is very likely that a false alarm-prone system would not be implemented in a real-world setting in which misses and false alarms have the same cost. There is a clear dissimilarity between the pay-off chosen in these experiments and the chosen placement of criterion. Nevertheless, this should not have a huge impact on the results. Participants actually behaved a bit as if the pay-off would punish misses more strongly than false alarms. For example, BAS users in Experiment 1 complied with more than 70% of alerts having a PPV of .35. In addition, a lot of participants confessed during the debriefing that they tried to avoid misses more than false alarms. Using a chemical plant as cover story in our experiments might have supported such believe. As a consequence, the dissimilarity between the pay-off and the characteristics of the alarm system might not have been so important as it could have been the case with another cover story.

9.3. Practical implications and future of research on LAS

LAS has often been presented as a promising alternative to BAS in order to reduce the cry-wolf effect and to improve operators´ performance, both in the alarm task and in the concurrent tasks. This series of studies has shown that LAS does not systematically lead to an improvement of performance in comparison to BAS, but might, among other variables, depend on the characteristics of the alarm task or of the environment.

The results show that the base rate of critical events is an important moderator to take into account. If the base rate is low, LAS does not improve performance with regards to the number of critical events detected but with regards to false alarms (compared to BAS). If the base rate of critical events is high, BAS helps operators to better detect critical events (compared to LAS). Consequently, LAS should be preferred over BAS in low base rate settings, while BAS should be preferred over LAS in high base rate settings. More generally, this means that LAS should not be implemented if there is the possibility to implement a BAS with a PPV over .6.

Furthermore, our findings show that the cost of consulting AVI does not moderate the beneficial effect of LAS over BAS with regards to the number of critical events detected.

When based only on the conclusion from Experiment 2, BAS should be preferred over LAS if AVI is available, whatever the cost of consulting AVI is.

Another finding is that, even though 4-stage LAS improves performance over 3-stage LAS, additional stages (e.g., 5-stage LAS) does not further improve performance further.

One key takeaway from this PhD Dissertation is that there is no optimal alarm system design or characteristic that can be applied to the diversity of task settings and environments.

Because the optimal characteristics for alarm systems depend, among other things, on the characteristics of a specific situation, which might change over time, one direction for future research on alarm system is to investigate how adaptive alarm systems can improve performance. An adaptive alarm system would be a system able to analyze and comprehend the evolving situation and implement some changes without any action required from humans (Feigh, Dorneich & Hayes, 2012). When using the term 'adaptive', we do not refer exclusively to an adaptive allocation of functions between the operator and the alarm system (i.e., information acquisition, information analysis, decision selection, and action implementation). We rather see adaptive alarm systems as being able to adapt their design and settings to the characteristics of the alarm task (e.g., access to additional information, difficulty of the alarm task), to the characteristics of the environment (e.g., base rate of critical events, presence of concurrent tasks), and to the characteristics of operators (e.g., operators' responding strategies, workload, stress). For example, future adaptive alarm systems could adjust their type (BAS vs. LAS), their placement of their criterion (liberal vs. conservative), and their number of stages (in the case of LAS).

A factor, which adaptive alarm systems could adjust to, is operators' behavior. Operators' behavior can change the base rate of critical events (Meyer & Bitan, 2002). A nurse who demonstrates health preventive behaviors towards patients might avoid the occurrence of critical problems. As a consequence, this nurse will experience a lower base rate of critical events in comparison to a nurse, who does not initiate such preventive behaviors. For the preventive nurse, LAS would lead to better performance than BAS, at least regarding false alarms and correct rejections (as shown in Experiment 1). However, BAS would improve performance for nurses, who do not demonstrate health preventive behavior. This means that

LAS, which can adapt and therefore turn into BAS depending on the user, is more likely to reach optimal performance. Nevertheless, such modifications of type might increase operators' workload and confusion. A more appropriate solution might be to keep LAS design whatever the base rate of critical events is, but to modify the placement of criterion in order to merge the warning stage and the alarm stage of LAS into a single stage if needed. In this case, even though the design in three stages is maintained, LAS would only emit alarms and no-alerts.

Even though adaptive alarm systems might be successful for optimizing human-alarm system performance, more research is needed, for example to investigate how adaptive alarm systems should adapt to the specific characteristics of the situations or the operators (Feigh, Dorneich & Hayes, 2012). An inappropriate adaptation could have unfortunate consequences such as increasing workload and annoyances, decreasing performance, or compromising safety. Future research should further investigate adaptive alarm systems and point towards better design of human-alarm systems interaction.

Another direction, which would be interesting to follow in future research about alarm system is the implication of participants´ response strategies on attention management.

A recurrent problem with alarms is that they interrupt operators in their ongoing tasks. These interruptions can have negative consequences during cognitively demanding times. Latorella (1996) investigated the effects of air traffic management interruptions on pilots´ performance during particularly demanding times, such as takeoff and landing. She reported that participants´ performance in the ongoing task was 53% more likely to contain errors when an interruption occurred. Alerts emitted by alarm systems have probably a similar effect on performance as air traffic management interruptions.

When looking at an alarm system set-up from this point of view, one cannot exclude that some participants ignored alerts only to avoid a decrease of performance related to switch time costs and resumption lag costs (i.e., time required to return to the interrupted task). Task-switching and interruption costs have already been mentioned by Dixon and Wickens (2006) and Maltz and Meyer (2001) as causes of the cry-wolf effect. According to Maltz and Meyer (2001), switching attention away from the concurrent tasks to respond to an alarm signal

incurs high cost on performance. Participants might only accept this cost if it is very likely that the alarm system is correct.

The importance of considering the costs of task-switching and tasks interruptions as causes of the cry-wolf effect is also emphasized by Wickens et al. (2009). In this field study, the authors investigated to what extent the false alarm rate of conflict alerting system in air traffic control affect operators´ response to true alerts and participants´ response time. No evidence of the cry-wolf effect was found. The authors speculate that this was due to the single-task setting since most of the evidence for the cry-wolf effect has been found in dual-task settings or multitask settings. In the air traffic control setting no task switching is required when an alarm goes off. As this study was a field study some confound variables might have influenced the results. The effect of the false alarm rate of alarm systems on the cry-wolf effect should be investigated further in a laboratory experiment comparing a single-task setting to multitask setting in order to confirm this speculation.

Costs related to task-switching and task interruptions have been rarely considered in the literature on human-alarm system interaction. More research is this area is needed as it could provide new knowledge about the reasons why the cry-wolf effect occurs and could help interpreting inconsistent effects.

In conclusion, this PhD Dissertation was successful to identify factors, which can moderate and sometimes inverse the benefit of LAS over BAS. More broadly, this research emphasize the importance of informing practitioners about human-alarm systems interactions principles so that they can make optimal use of them to choose and design alarm systems that have a significant potential to increase safety and productivity in complex working environments.

REFERENCES

Allendoerfer, K., Pai, S., & Friedman-Berg, F. J. (2008). The complexity of signal detection in air traffic control alert situations. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 54-58). Santa Monica, CA: Human Factors and Ergonomics Society.

Andre, A. D., & Cutler, H. A. (1998). Displaying uncertainty in advanced navigation systems. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 31-35). Santa Monica, CA: Human Factors and Ergonomics Society.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Proceedings of INTERACT* (pp. 593-601). Amsterdam: IOS Press.

Balaud, M., & Manzey, D. (2013). Kontrasteffekte bei der benutzung von likelihood alarmsystemen. In E. Brandenburg, L. Doria, A. Gross, T. Günzler & H. Smieszek (Eds.), *Grundlagen und Anwendungen der Mensch-Maschine-Interaktion: 10. Berliner Werkstatt Mensch-Maschine-Systeme* (pp. 264-271). Berlin: Universitätsverlag der TU Berlin.

Balaud, M., & Manzey, D. (2015). The more the better? The impact of number of stages of likelihood alarm systems on human performance. In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.), *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference,* pp. 61-72.

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., & Benight, C. (2007). False alarms and close calls: A conceptual model of warning accuracy. *Weather and Forecasting*, *22*(5), 1140-1147.

Bliss, J. P. (2003a). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology*, *13*(3), 249-268.

Bliss, J. P. (2003b). An investigation of extreme alarm responses of extreme alarm response patterns in laboratory experiments. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 1683-1687). Santa Monica, CA: Human Factors and Ergonomics Society.

Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. Ergonomics, 43(9), 1283–1300.

Bliss, J. P., & Fallon, C. K. (2006). Active warnings II: False alarms. In M. Wogalter (Hrsg.), *Handbook of Warnings* (pp. 232-242). Mahwah, NJ: Lawrence Erlbaum.

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

Bliss, J. P., Jeans, S., & Prioux, H. (1996). Dual-task performance as a function of individual alarm validity and alarm system reliability information. In *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

Borowski, M., Görges, M., Fried, R., Such, O., Wrede, C., & Imhoff, M. (2011). Medical device alarms. *Biomedizinische Technik/Biomedical Engineering*, *56*(2), 73-83.

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

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

Burt, J. L., Bartolome-Rull, D. S., Burdette, D. W., & Comstock, J. R. (1999). A psychophysical evaluation of the perceived urgency of auditory warning signals. In N. A. Stanton & J. Edworthy (Eds.), *Human Factors in Auditory Warnings* (pp. 153-165). Aldershot: Ashgate Publishing Company.

Bustamante, E. A. (2005). *The a-b signal detection theory model* (Unpublished doctoral dissertation). Old Dominion University, Norfolk, Virginia.

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*(2), 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.

Chambrin, M. C., Ravaux, P., Calvelo-Aros, D., Jaborska, A., Chopin, C., & Boniface, B. (1999). Multicentric study of monitoring alarms in the adult intensive care unit (ICU): A descriptive analysis. *Intensive Care Medicine*, *25*(12), 1360-1366.

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

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

Clark, R. M., Ingebritsen, A. M., & Bustamante, E. A. (2010). Differential effects of likelihood alarm technology and false-alarm vs. miss-prone automation on decision-making accuracy and bias. In *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting* (pp. 1508-1512). Santa Monica, CA: Human Factors and Ergonomics Society.

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* 53rd *Annual Meeting* (pp. 349-353). Santa Monica, CA: Human Factors and Ergonomics Society.

Coiera, E., Westbrook, J., & Wyatt, J. (2006). The safety and quality of decision support systems. *Methods Inf Med*, *45*(Suppl. 1), 20-5.

Comstock, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behaviour research. Hampton, VA: NASA Langley Research Center.

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

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(3), 479-487.

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

Edworthy, J., Stanton, N., & Hellier, E. (1995). Warnings in research and practice. *Ergonomics*, *38*(11), 2145-2154.

Edworthy, J., Hellier, E., Walters, K., Clift-Mathews, W., & Crowther, M. (2003). Acoustic, semantic and phonetic influences in spoken warning signal words. *Applied Cognitive Psychology*, *17*(8), 915-933.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems a framework for researchers and system designers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(6), 1008-1024.

Garot, J. M., & Durand, N. (2005). Failures in the automation of air traffic control. Paper presented at *Colloque de l'AAA*. Paper retrieved from http://pom.tls.cena.fr/papers/notice.html?todo=abstract&type=articles&file=garotdurand.txt

Gérard, N. (2012). *Verhaltenseffektivität von alarmen: Experimentelle untersuchungen zum einfluss von reliabilität und prüfmöglichkeit auf die anwendung von heuristiken* (Doctoral dissertation). Technische Universität Berlin, Berlin.

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.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Psychological Bulletin*, *75*, 424-429.

Guenzler, T., & Manzey, D. (2013). Asymmetries in Human Tolerance of Uncertainty in Interaction with Alarm Systems Effects of Risk Perception or Evidence for a General Commission Bias? In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 362-1366). Santa Monica, CA: Human Factors and Ergonomics Society.

Haas, E. C., & Casali, J. G. (1995). Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise. *Ergonomics*, *38*(11), 2313-2326.

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

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267-272.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11-26.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188.

Jacobs, K. W., & Suess, J. F. (1975). Effects of four psychological primary colors on anxiety state. *Perceptual and Motor Skills*, *41*(1), 207-210.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341-350.

Kornblum, S. (1973). Sequential effects in choice reaction time: A tutorial review. In I. Kornblum (Ed.), *Attention and Performance IV*. New York: Academic Press.

Latorella, K. A. (1996). *Investigating Interruptions: implications for flightdeck performance* (Doctoral dissertation, Faculty of Graduate school of the state university of New York, Buffalo). Retrieved from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004291.pdf

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

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153-184.

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

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

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(2), 217-226.

Manzey, D., Gérard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: The impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, 1-23.

McCarley, J. S. (2009). Response criterion placement modulates the benefits of graded alerting systems in a simulated baggage screening task. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (pp. 1106-1110). Santa Monica, CA: Human Factors and Ergonomics Society.

McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, *17*(1), 1-61.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 563-572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 196-204.

Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(3), 343-353.

Meyer, J., Bitan, Y., Shinar, D., & Zmora, E. (1999). Scheduling of actions and reliance on warnings in a simulated control task. In *Proceedings of the Human Factors and Ergonomics Society 43th Annual Meeting* (pp. 251-255). Santa Monica, CA: Human Factors and Ergonomics Society.

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(5), 840-849.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44-58.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *29*(6), 695-706.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381-410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomic Society, 39*, 230-253.

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

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*(3), 286-297.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1-15.

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

Ragsdale, S. A. (2012). Fault diagnosis with multi-state alarms in a nuclear power control simulation. (Unpublished master thesis). University of Idaho, Moscow.

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 56th Annual Meeting* (pp. 2167-2171). Boston, MA: Human Factors and Ergonomics Society.

Rice, S. D. (2009). Examining single- and multiple-process theories of trust in automation. *The Journal of General Psychology*, *136*(3), 303-319.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*(4), 320-331.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207-231.

Roth, E. M., & O'Hara, J. M. (1999). Exploring the impact of advanced alarms, displays, and computerized procedures on teams. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 158-162). Santa Monica, CA: Human Factors and Ergonomics Society.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, *6*(1), 1-36.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 1926-1943). New York: Wiley.

Sheridan, T. B., & Parasuraman, R. (2006). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1, 89-129.*

Shurtleff, M. S. (1991). Effects of specificity of probability information on human performance in a signal detection task. *Ergonomics, 34*(4), 469-486.

Smith, K. (2008). Remote command and control, trust, stress, and soldier performance. In P. A. Hancock & J. L. Szalma (Eds.), *Performance under stress*. (pp. 77-100). Brookfield, VT: Ashgate Publishing Company.

Smith-Jackson, T. L., & Wogalter, M. S. (2000). Applying cultural ergonomics/human factors to safety information research. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 6-150). Santa Monica, CA: Human Factors and Ergonomics Society.

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

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.

Swanson , L. A. (2010). *Response criterion placement modulates the effects of graded alerting systems on human performance and learning in a target detection task* (Doctoral dissertation, University of Illinois at Urbana-Champaign). Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/18377/Swanson_Leah.pdf?sequence=1

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

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

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101-118.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 63-101). New York: Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159-177.

Wickens, C. D., Dixon, S. R., Goh, J., & Hammer, B. (2005). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. Savoy, IL: University of Illinois.

Wickens, C. D., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 839-850.

Wickens, C. D., Hollands, J. G., Parasuraman, R., & Banbury, S. (2013). *Engineering Psychology and Human Performance (4th ed.)*. Upper Saddle River, NJ: Pearson.

Wickens, C. D., & Horrey, W. J. (2008). Models of attention, distraction, and highway hazard avoidance. In M. A. Regan, J. D. Lee & K. L. Young (Eds.), *Driver distraction: Theory, effects, and mitigation* (pp. 41-56). Boca Raton, FL: CRC Press Taylor & Francis Group.

Wickens, C. D., Rice, S., Hutchins, S., Keller, M. D., Hughes, J., & Clayton, K. (2009). False Alerts in the Air Traffic Control Traffic Conflict Alerting System: Is There a "Cry Wolf" Effect. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 51*(4), 446-462.

Wiczorek, R. (2011). Entwicklung und evaluation eines mehrdimensionalen fragebogens zur messung von vertrauen in technische systeme. In S. Schmid, M. Elepfandt, J. Adenauer & A. Lichtenstein (Eds.), *Reflexionen und Visionen der Mensch-Maschine-Interaktion – Aus der Vergangenheit lernen, Zukunft Gestalten, 9. Berliner Werkstatt Mensch-Maschine-Systeme* (pp. 621-626). Berlin: VDI.

Wiczorek, R. (2012). *Verhaltenswirksamkeit von likelihood alarmsystemen* (Doctoral dissertation). Technische Universität Berlin, Berlin.

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*, 1209-1221

Wiczorek, R., Manzey, D. & Zirk, A. (2014). Benefits of decision-support by likelihood versus binary alarm systems: Does the number of stages make a difference? In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting* (pp. 380-384). Santa Monica, CA: Human Factors and Ergonomics Society.

Wilkinson, R. T. (1964). Artificial “signals” as an aid to an inspection. *Ergonomics*, *7*(1), 63-72.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: Rare items often missed in visual searches. *Nature*, *435*(7041), 439-440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*(4), 623-638.

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

Wogalter, M. S., Frederick, L. J., Magurno, A. B., Herrera, O. L. (1997). Connoted hazard of Spanish and English warning signal words, colors, and symbols by native Spanish language users. In *Proceedings of the 13th Triennial Congress of the International Ergonomics Assn.* (pp.353–355). Tampere, Finland: IEA.

Woods, D. D. (1985). Cognitive technologies: The design of joint human-machine cognitive systems. *AI Magazine*, *6*(4), 86-92.

Zirk, A. (2013). *Binäre und Likelihood-Alarmsysteme im direkten Vergleich: Einfluss der Stufenanzahl auf das Verhalten und die Leistung des Operateurs* (Unpublished master's thesis). Institut für Psychologie und Arbeitswissenschaft der Technischen Universität, Berlin.

ANNEX A

DESCRIPTIVE DATA OF EXPERIMENT 1

Experiment 1: Manipulation check

Estimated PPV of alarms

Alarm system	PPV .35	PPV .65
BAS	M = .46	M = .60
	SD = .12	SD = .15
LAS	M = .64	M = .70
	SD = .30	SD = .33

Estimated PPV of warnings

Alarm system	PPV .35	PPV .65
LAS	M = .34	M = .46
	SD = .12	SD = .09

Estimated NPV

Alarm system	PPV .35	PPV .65
BAS	M = .90	M = .87
	SD = .10	SD = .13
LAS	M = .90	M = .92
	SD = .04	SD = .07

Experiment 1: Participants' response rates

Mean compliance rate to alerts (in percentage)

Alarm system	PPV .35	PPV .65
BAS	M = 70.40	M = 97.34
	SD = 41.53	SD = 4.75
LAS	M = 41.41	M = 71.30
	SD = 28.26	SD = 22.55

Mean compliance rate to alarms (in percentage)

Alarm system	PPV .35	PPV .65
BAS	M = 70.40	M = 97.34
	SD = 41.53	SD = 4.75
LAS	M = 97.69	M = 98.83
	SD = 6.47	SD = 1.97

Mean compliance rate to warnings (in percentage)

Alarm system	PPV .35	PPV .65
LAS	M = 20.31	M = 51.11
	SD = 38.28	SD = 38.65

Mean ignoring rate to non-alerts (in percentage)

Alarm system	PPV .35	PPV .65
BAS	M = 97.29	M = 97.55
	SD = 7.35	SD = 6.32
LAS	M = 99.51	M = 99.65
	SD = 1.70	SD = 1.26

Experiment 1: Participants' performance in the Alarm Task

Participants' percentage of hits

Alarm system	PPV .35	PPV .65
BAS	M = 24.59	M = 63.81
	SD = 14.36	SD = 3.19
LAS	M = 24.12	M = 51.18
	SD = 6.82	SD = 10.41

Participants' percentage of misses

Alarm system	PPV .35	PPV .65
BAS	M = 10.26	M = 1.58
	SD = 14.36	SD = 3.19
LAS	M = 12.25	M = 14.20
	SD = 6.82	SD = 10.41

Participants' percentage of false alarms

Alarm system	PPV .35	PPV .65
BAS	M = 45.80	M = 33.53
	SD = 27.27	SD = 2.02
LAS	M = 17.30	M = 20.12
	SD = 21.72	SD = 12.29

Participants' percentage of correct rejections

Alarm system	PPV .35	PPV .65
BAS	M = 19.35	M = 1.08
	SD = 27.27	SD = 2.02
LAS	M = 46.34	M = 14.50
	SD = 21.72	SD = 12.29

Experiment 1: Participants' performance in the concurrent tasks

Number of refilling cycles successfully completed in the Coolant Exchange Task

Alarm system	PPV .35	PPV .65
BAS	M = 14.31	M = 13.31
	SD = 2.84	SD = 1.54
LAS	M = 15.67	M = 13.69
	SD = 1.72	SD = 2.06

Number of correctly sent orders in the Resource Ordering Task

Alarm system	PPV .35	PPV .65
BAS	M = 108.84	M = 94.46
	SD = 52.39	SD = 41.64
LAS	M = 106.33	M = 94.69
	SD = 54.66	SD = 31.47

Proportion of orders participants responded to in the Resource Ordering Task (in percentage)

Alarm system	PPV .35	PPV .65
BAS	M = 96.39	M = 95.93
	SD = 3.7	SD = 5.44
LAS	M = 95.14	M = 95.66
	SD = 5.19	SD = 3.54

Experiment 1: Participants' trust in the alarm system

Participants' mean trust ratings at the FMV questionnaire

Alarm system	PPV .35	PPV .65
BAS	M = 2.57	M = 2.68
	SD = 0.71	SD = 0.51
LAS	M = 2.80	M = 2.86
	SD = 0.51	SD = 0.37

FMV questionnaire: reliability item

Alarm system	PPV .35	PPV .65
BAS	M = 2.02	M = 1.96
	SD = 0.86	SD = 0.83
LAS	M = 2.27	M = 2.42
	SD = 0.67	SD = 0.62

FMV questionnaire: transparency item

Alarm system	PPV .35	PPV .65
BAS	M = 2.92	M = 2.94
	SD = 0.94	SD = 0.76
LAS	M = 3.00	M = 3.04
	SD = 0.95	SD = 0.68

FMV questionnaire: intention item

Alarm system	PPV .35	PPV .65
BAS	M = 2.90	M = 3.08
	SD = 0.75	SD = 0.47
LAS	M = 2.96	M = 3.02
	SD = 0.60	SD = 0.40

FMV questionnaire: use item

Alarm system	PPV .35	PPV .65
BAS	M = 2.44	M = 2.75
	SD = 0.96	SD = 0.66
LAS	M = 2.96	M = 2.96
	SD = 0.42	SD = 0.38

ANNEX B

DESCRIPTIVE DATA OF EXPERIMENT 2

Experiment 2: Manipulation check

Estimated PPV of alarms

Alarm system	Low-cost	High-cost
BAS	M = .59	M = .40
	SD = .12	SD = .11
LAS	M = .79	M = .75
	SD = .15	SD = .16

Estimated PPV of warnings

Alarm system	Low-cost	High-cost
LAS	M = .51	M = .38
	SD = .19	SD = .14

Estimated NPV

Alarm system	Low-cost	High-cost
BAS	M = .74	M = .83
	SD = .21	SD = .14
LAS	M = .79	M = .86
	SD = .21	SD = .09

Experiment 2: Participants' response rates

Participants' compliance rate with alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 1.19	M = 12.22
	SD = 1.69	SD = 21.93
LAS	M = 8.81	M = 20.40
	SD = 12.26	SD = 10.08

Participants' cross-checking rate of alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 93.07	M = 63.85
	SD = 6.74	SD = 40.27
LAS	M = 80.58	M = 60.20
	SD = 19.09	SD = 22.69

Participants' ignoring rate of alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 5.73	M = 23.91
	SD = 6.81	SD = 34.33
LAS	M = 10.60	M = 19.39
	SD = 13.20	SD = 26.54

Participants' compliance rate with alarms (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 1.19	M = 12.22
	SD = 1.69	SD = 21.93
LAS	M = 30.90	M = 71.11
	SD = 42.83	SD = 36.80

Participants' cross-checking rate of alarms (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 93.07	M = 63.85
	SD = 6.74	SD = 40.27
LAS	M = 60.76	M = 16.67
	SD = 43.01	SD = 26.56

Participants' ignoring rate of alarms (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 5.73	M = 23.91
	SD = 6.81	SD = 34.33
LAS	M = 8.33	M = 12.22
	SD = 15.18	SD = 26.49

Participants' compliance rate with warnings (in percentage)

Alarm system	Low-cost	High-cost
LAS	M = 0.52	M = 1.39
	SD = 2.08	SD = 2.03

Participants' cross-checking rate of warnings (in percentage)

Alarm system	Low-cost	High-cost
LAS	M = 88.46	M = 76.52
	SD = 14.10	SD = 28.23

Participants' ignoring rate of warnings (in percentage)

Alarm system	Low-cost	High-cost
LAS	M = 11.45	M = 22.08
	SD = 13.22	SD = 28.32

Participants' compliance rate with non-alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = .00	M = 4.83
	SD = .00	SD = 18.08
LAS	M = 2.21	M = .00
	SD = 8.07	SD = .00

Participants' cross-checking rate of non-alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 66.38	M = 21.00
	SD = 39.12	SD = 35.85
LAS	M = 63.05	M = 11.96
	SD = 41.33	SD = 27.13

Participants' ignoring rate of non-alerts (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 33.61	M = 74.16
	SD = 39.12	SD = 40.48
LAS	M = 34.74	M = 88.04
	SD = 40.00	SD = 27.13

Experiment 2: Participants' performance in the Alarm Task

Participants' percentage of hits

Alarm system	Low-cost	High-cost
BAS	M = 33.65	M = 27.27
	SD = 3.47	SD = 5.00
LAS	M = 33.83	M = 28.35
	SD = 2.19	SD = 5.65

Participants' percentage of misses

Alarm system	Low-cost	High-cost
BAS	M = 2.71	M = 9.09
	SD = 3.47	SD = 5.01
LAS	M = 2.53	M = 8.01
	SD = 2.19	SD = 5.65

Participants' percentage of false alarms

Alarm system	Low-cost	High-cost
BAS	M = 0.97	M = 1.79
	SD = 1.28	SD = 2.86
LAS	M = 2.46	M = 5.35
	SD = 3.54	SD = 2.97

Participants' percentage of correct rejections

Alarm system	Low-cost	High-cost
BAS	M = 62.66	M = 61.85
	SD = 1.28	SD = 2.86
LAS	M = 61.17	M = 58.28
	SD = 3.54	SD = 2.97

Experiment 2: Participants' performance in the concurrent tasks

Number of refilling cycles successfully completed in the Coolant Exchange Task

Alarm system	Low-cost	High-cost
BAS	M = 14.36	M = 18.14
	SD = 1.55	SD = 1.79
LAS	M = 14	M = 17.47
	SD = 1.55	SD = 2.20

Number of correctly sent orders in the Resource Ordering Task

Alarm system	Low-cost	High-cost
BAS	M = 88.29	M = 118.14
	SD = 19.75	SD = 32.83
LAS	M = 82.25	M = 101.57
	SD = 39.46	SD = 24.60

Proportion of orders participants responded to in the Resource Ordering Task (in percentage)

Alarm system	Low-cost	High-cost
BAS	M = 96.82	M = 91.16
	SD = 2.59	SD = 9.37
LAS	M = 92.89	M = 91.59
	SD = 7.20	SD = 8.66

Experiment 2: Participants' trust in the alarm system

Participants' mean trust ratings at the FMV questionnaire

Alarm system	Low-cost	High-cost
BAS	M = 2.66	M = 2.56
	SD = 0.69	SD = 0.36
LAS	M = 2.86	M = 3.03
	SD = 0.46	SD = 0.46

FMV questionnaire: reliability item

Alarm system	Low-cost	High-cost
BAS	M = 1.98	M = 1.75
	SD = 0.73	SD = 0.67
LAS	M = 2.38	M = 2.6
	SD = 0.49	SD = 0.57

FMV questionnaire: transparency item

Alarm system	Low-cost	High-cost
BAS	M = 3.11	M = 3.34
	SD = 0.76	SD = 0.74
LAS	M = 2.98	M = 3.13
	SD = 0.87	SD = 0.93

FMV questionnaire: intention item

Alarm system	Low-cost	High-cost
BAS	M = 3.11	M = 3.09
	SD = 0.85	SD = 0.71
LAS	M = 3.27	M = 3.43
	SD = 0.47	SD = 0.44

FMV questionnaire: use item

Alarm system	Low-cost	High-cost
BAS	M = 2.43	M = 2.07
	SD = 0.96	SD = 0.50
LAS	M = 2.8	M = 2.97
	SD = 0.71	SD = 0.59

Trust item

Alarm system	Low-cost	High-cost
BAS	M = 4.39	M = 3.42
	SD = 2.32	SD = 1.92
LAS	M = 5.24	M = 5.93
	SD = 2.31	SD = 2.06

Experiment 2: Participants' workload

Participants' mean workload ratings at the NASA-TLX

Alarm system	Low-cost	High-cost
BAS	M = 5.25	M = 5.20
	SD = 1.43	SD = 1.13
LAS	M = 5.70	M = 5.27
	SD = 1.21	SD = 1.35

NASA-TLX: mental demand item

Alarm system	Low-cost	High-cost
BAS	M = 6.23	M = 6.25
	SD = 2.20	SD = 2.42
LAS	M = 6.95	M = 6.57
	SD = 2.02	SD = 1.78

NASA-TLX: physical demand item

Alarm system	Low-cost	High-cost
BAS	M = 3.37	M = 2.83
	SD = 2.35	SD = 2.21
LAS	M = 3.54	M = 3.13
	SD = 2.38	SD = 2.04

NASA-TLX: temporal demand item

Alarm system	Low-cost	High-cost
BAS	M = 7.65	M = 7.15
	SD = 1.43	SD = 1.72
LAS	M = 8.34	M = 6.88
	SD = 1.42	SD = 1.99

NASA-TLX: performance item

Alarm system	Low-cost	High-cost
BAS	M = 3.57	M = 4.31
	SD = 2.11	SD = 1.63
LAS	M = 3.73	M = 4.39
	SD = 1.88	SD = 2.32

NASA-TLX: effort item

Alarm system	Low-cost	High-cost
BAS	M = 6.62	M = 6.36
	SD = 1.90	SD = 1.87
LAS	M = 7.20	M = 6.65
	SD = 1.94	SD = 2.11

NASA-TLX: frustration item

Alarm system	Low-cost	High-cost
BAS	M = 4.09	M = 4.31
	SD = 2.64	SD = 2.44
LAS	M = 4.44	M = 4.00
	SD = 2.62	SD = 2.40

ANNEX C

DESCRIPTIVE DATA OF EXPERIMENT 3

Experiment 3: Manipulation check

Estimated PPV and NPV of alerts

Alarm system	Alarm	Orange warning	Yellow-orange warning	Yellow warning	Non-alert
LAS3	M = .87 SD = .17	M = .30 SD = .13			M = .93 SD = .05
LAS4	M = .81 SD = .20	M = .65 SD = .19		M = .26 SD = .11	M = .89 SD = .10
LAS5	M = .76 SD = .23	M = .61 SD = .18	M = .41 SD = .22	M = .29 SD = .20	M = .83 SD = .22

Experiment 3: Participants' response rates

Mean compliance rate to alerts (in percentage)

Alarm system	Alerts
LAS3	M = 57.53 SD = 15.77
LAS4	M = 69.45 SD = 11.65
LAS5	M = 56.66 SD = 16.23

Mean compliance rate to the different kind of alert emitted by LAS (in percentage)

Alarm system	Alarm	Orange warning	Yellow-orange warning	Yellow warning	Non-alert
LAS3	M = 98.56 SD = 3.10	M = 16.51 SD = 7.64			M = 99.82 SD = 0.74
LAS4	M = 96.63 SD = 5.60	M = 97.16 SD = 5.47		M = 14.58 SD = 31.53	M = 99.63 SD = 1.00
LAS5	M = 95.19 SD = 9.68	M = 86.36 SD = 10.95	M = 35.71 SD = 41.24	M = 9.37 SD = 26.14	M = 99.45 SD = 2.21

Experiment 3: Participants' performance in the Alarm Task

Participants' percentage of hits, misses, false alarms and correct rejections

Alarm system	Hit	Miss	False alarm	Correct rejection
LAS3	M = 17.64 SD = 2.49	M = 18.72 SD = 2.49	M = 9.29 SD = 12.27	M = 54.34 SD = 18.03
LAS4	M = 26.33 SD = 3.95	M = 10.04 SD = 3.95	M = 18.18 SD = 17.01	M = 45.45 SD = 17.01
LAS5	M = 26.42 SD = 4.39	M = 9.94 SD = 4.39	M = 18.28 SD = 13.90	M = 45.36 SD = 13.90

Experiment 3: Participants' performance in the concurrent tasks

Number of refilling cycles successfully completed in the Coolant Exchange Task

Alarm system	
LAS3	M = 16.31 SD = 2.59
LAS4	M = 16.13 SD = 2.78
LAS5	M = 16.00 SD = 2.66

Number of correctly sent orders in the Resource Ordering Task

Alarm system	
LAS3	M = 118.38 SD = 35.82
LAS4	M = 119.44 SD = 61.04
LAS5	M = 105.53 SD = 40.92

Proportion of orders participants responded to in the Resource Ordering Task (in percentage)

Alarm system	
LAS3	M = 130.62 SD = 31.22
LAS4	M = 133.63 SD = 52.50
LAS5	M = 119.40 SD = 33.41

Experiment 3: Participants' trust in the alarm system

Participants' mean trust ratings at the FMV questionnaire

Alarm system	Overall	Reliability	Transparency	Intention	Use
LAS3	M = 2.97 SD = 0.38	M = 2.38 SD = 0.62	M = 3.09 SD = 0.64	M = 3.27 SD = 0.42	M = 3.16 SD = 0.58
LAS4	M = 2.91 SD = 0.41	M = 2.25 SD = 0.71	M = 3.19 SD = 0.69	M = 3.16 SD = 0.55	M = 3.03 SD = 0.43
LAS5	M = 3.04 SD = 0.37	M = 2.34 SD = 0.60	M = 3.16 SD = 0.60	M = 3.55 SD = 0.37	M = 3.10 SD = 0.65

Experiment 3: Participants' workload

Participants' mean workload ratings at the NASA-TLX

Alarm system	Overall	Mental demand	Physical demand	Temporal demand	Perfor-mance	Effort	Frustration
LAS3	M = 11.24 SD = 4.26	M = 15.25 SD = 4.73	M = 7.88 SD = 5.22	M = 14.22 SD = 4.61	M = 8.28 SD = 4.72	M = 13.03 SD = 5.44	M = 8.78 SD = 5.42
LAS4	M = 10.76 SD = 3.35	M = 13.59 SD = 4.71	M = 5.81 SD = 3.52	M = 13.81 SD = 5.05	M = 8.34 SD = 4.54	M = 14.38 SD = 4.38	M = 8.59 SD = 4.63
LAS5	M = 12.40 SD = 1.80	M = 15.75 SD = 2.42	M = 8.00 SD = 5.33	M = 15.16 SD = 3.62	M = 10.90 SD = 5.07	M = 14.38 SD = 3.95	M = 10.19 SD = 5.00

ANNEX D

INSTRUCTIONS

Instructions from Experiment 2, for participants using LAS in the low-cost condition

HERZLICH
WILLKOMMEN!

WEITER

Vielen Dank, dass du an dieser Untersuchung teilnimmst!

Die Teilnahme an diesem Versuch erfolgt freiwillig. Es ist zu jedem Zeitpunkt möglich, den Versuch ohne Angabe von Gründen abzubrechen. In diesem Fall können wir jedoch keine Aufwandsentschädigung ausbezahlen.

Alle erhobenen Daten werden anonymisiert gespeichert und ausgewertet. Es erfolgt keine Weitergabe an Dritte.

VOR

Du wirst zunächst erfahren, was genau deine Aufgabe in diesem Versuch sein wird. Anschließend wirst du die Gelegenheit haben, jede Aufgabe kurz zu üben. Außerdem wird es einen kurzen Probedurchlauf geben.

Danach findet der eigentliche Versuchsdurchlauf statt, in dem du dieselbe Aufgabe wie zuvor bearbeiten sollst. Nach dem Probedurchlauf und dem eigentlichen Versuchsdurchlauf werden wir dich jeweils bitten, uns einige Fragen zu beantworten.

Bitte lies dir die Aufgabenbeschreibung **sorgfältig** durch! Dort wird dir alles erklärt, was du wissen musst. Mit den Schaltflächen „VOR" und „ZURÜCK" kannst du durch die Seiten blättern. Während der Durchgänge wird es nicht möglich sein, Fragen zu stellen. Falls du nach dem Versuch noch Fragen hast, werde ich sie dir gerne beantworten.

ZURÜCK VOR

Instruktion

ZURÜCK VOR

Hintergrund

In vielen Industrien, wie zum Beispiel im Bereich der Chemie- und Energiebranche, werden Produktionsanlagen von Leitwarten (Kontrollräumen) aus gesteuert und überwacht. In diesen Leitwarten fallen verschiedene Arbeitsaufgaben an. Sie bilden den Ausgangspunkt für diese Untersuchung. Hierbei ist interessant, wie Menschen mit mehreren verschiedenen Aufgaben und den daraus entstehenden Belastungen in Leitwarten umgehen.

VOR

Im Rahmen der heutigen Untersuchung arbeitest du als Schichtmitarbeiter in der Leitwarte einer chemischen Anlage. Deine Aufgabe ist die Steuerung der chemischen Anlage.

ZURÜCK VOR

Bei der Steuerung einer so komplexen Anlage gibt es viele verschiedene Aufgaben. Du bist für drei Aufgaben zuständig:

- Bereitstellung der Katalysatoren
- Aufbereitung des Kühlwassers
- Überwachung der Reaktionscontainer

ZURÜCK VOR

Alle drei Aufgaben sind wichtig für die Steuerung der Anlage. Die Aufgaben werden dir im Folgenden genau erklärt.

ZURÜCK VOR

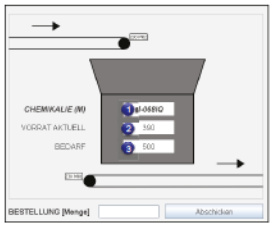
Bereitstellung der Katalysatoren

Bei dieser Aufgabe geht es darum, dafür zu sorgen, dass immer genügend Katalysatoren vorhanden sind. Es handelt sich dabei um Chemikalien, die für die Reaktion benötigt werden.

Es wird dir die aktuelle Menge an Katalysatoren sowie ihr aktueller Bedarf angezeigt. Deine Aufgabe ist es, die fehlende Menge zu ermitteln und entsprechend viele Katalysatoren zu bestellen.

ZURÜCK VOR

Bereitstellung der Katalysatoren



Diese Aufgabe wird sich auf deinem Bildschirm in der linken oberen Ecke befinden:

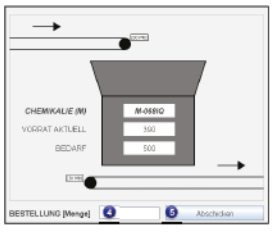
Im obersten Feld ist der Name der Chemikalie angezeigt. ❶

Darunter der aktuelle Vorrat ❷ und der Bedarf. ❸

Für die Bestellung muss die Differenz zwischen Bedarf und Vorrat gebildet werden.

ZURÜCK VOR

Bereitstellung der Katalysatoren



Die benötigte Menge wird über die Tastatur in das untere Feld ❹ eingetragen.

Die Bestellung wird durch klicken auf den „Abschicken"-Button ❺ ausgeführt.

Nach 1 Sekunde erscheint die nächste Chemikalie zur Bearbeitung.

Wird eine Bestellung nicht innerhalb von 15 Sekunden abgeschickt, gilt sie als nicht bearbeitet. Die nächste Chemikalie erscheint automatisch.

ZURÜCK VOR

Bereitstellung der Katalysatoren

Für jede einzelne korrekte Bestellung bekommst du in den späteren Versuchsdurchgängen **1,5 Punkte** auf dein Lohn-Konto.

ZURÜCK VOR

Bereitstellung der Katalysatoren

Jetzt kannst du die Aufgabe 2 Minuten lang üben.

Klicke bitte auf „ÜBEN" und bearbeite die Aufgabe bis die Übung von selbst endet.

ÜBEN

Wenn du die Übung beendet hast, klicke bitte auf „VOR".

ZURÜCK VOR

Aufbereitung des Kühlwassers

Die Reaktionsbehälter der Anlage müssen gekühlt werden. Überhitzte Anlagen stellen eine Gefahr für die Sicherheit der Produktion dar. Damit immer eine optimale Kühlung gewährleistet werden kann, muss das Kühlwasser in den Behältern regelmäßig ausgetauscht werden. So wird vermieden, dass sich das Kühlwasser zu stark erhitzt und die Kühlfunktion nicht mehr erfüllen kann.

Deine Aufgabe besteht darin, das Kühlwasser in den Behältern verschiedener Bereiche innerhalb der Anlage auszutauschen. Dieser Austausch sollte so schnell wie möglich erfolgen.

ZURÜCK VOR

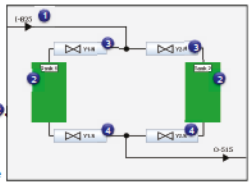
Aufbereitung des Kühlwassers

Diese Aufgabe wird sich auf deinem Bildschirm in der rechten oberen Ecke befinden:

Oberhalb der Behälter wird die Nummer des aktuellen Bereichs angezeigt 1.

Ein Bereich besteht aus zwei Behältern 2, zwei Zulaufventilen 3 und zwei Ablaufventilen 4.

Durch Anklicken mit der Maus können die Ventile geöffnet und geschlossen werden.



ZURÜCK VOR

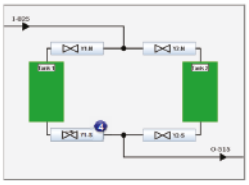
Aufbereitung des Kühlwassers

Die **grüne** Farbe der Behälter zeigt an, dass das Kühlwasser **verbraucht** ist und ein Austausch stattfinden muss.

Dazu muss das verbrauchte Wasser zunächst abgelassen werden. Dies geschieht über das Öffnen eines Ablaufventils mit Anklicken auf das Ventil (4 mit Pfeil: offen).

Es kann jeweils nur **ein** Ablaufventil zur gleichen Zeit geöffnet sein.

Das Ablaufen des Wassers kann einen Moment dauern. In dieser Zeit zeigt sich keine sichtbare Veränderung.



ZURÜCK VOR

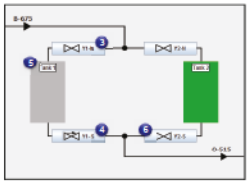
Aufbereitung des Kühlwassers

Ist das verbrauchte Wasser abgelaufen, erscheint der Behälter **grau** 5. Dies kann einen Moment dauern.

Dann kann für diesen Behälter das Ablaufventil 4 geschlossen und danach das Zulaufventil 3 geöffnet werden.

Anschließend kann das Ablaufventil des anderen Behälters 6 geöffnet werden.

So kann aus dem zweiten Behälter das alte Wasser (grün) ablaufen, während der erste gleichzeitig mit frischen Wasser befüllt wird.



Merke: Es können nie gleichzeitig beide Zulaufventile oder beide Ablaufventile geöffnet sein!

ZURÜCK VOR

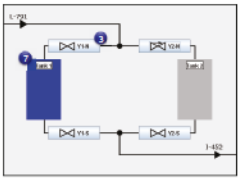
Aufbereitung des Kühlwassers

Ist der Behälter mit frischem Wasser befüllt, erscheint er blau 7.

Dann kann für diesen Behälter das Zulaufventil 3 geschlossen werden.

Sobald der andere Behälter leer (grau) ist, kann sein Ablaufventil geschlossen und sein Zulaufventil geöffnet werden.

Ist dieser Behälter gefüllt (blau), kann das Zulaufventil geschlossen werden.



Wenn beide Behälter mit frischem Wasser gefüllt und beide Ventile geschlossen sind, wird ein neuer Bereich zum Kühlwasseraustausch angezeigt.

ZURÜCK VOR

Aufbereitung des Kühlwassers

Du erhältst in den späteren Versuchsdurchgängen für ein Paar frisch befüllte Kühlwasserbehälter **7,5 Punkte** auf dein Lohnkonto.

Der Punkteunterschied zwischen der Bestellaufgabe und der Kühlwasseraufgabe erklärt sich dadurch, dass man länger braucht um die Kühlwasseraufgabe durchzuführen.

ZURÜCK VOR

Aufbereitung des Kühlwassers

Jetzt kannst du die Aufgabe 2 Minuten lang üben.

Klicke bitte auf „ÜBEN" und bearbeite die Aufgabe bis die Übung von selbst endet.

ÜBEN

Wenn du die Übung beendet hast, klicke bitte auf „VOR".

ZURÜCK VOR

Überwachung der Reaktionscontainer

Das chemische Endprodukt wird in Reaktionscontainer abgefüllt. Bevor diese Container an den Kunden ausgeliefert werden, muss überprüft werden, ob das Produkt fehlerfrei ist.

Eine zu hohe Temperatur im Container führt zu einem zu hohen Molekulargewicht. Wenn das **Molekulargewicht** im Behälter zu **hoch** ist, kann das Produkt **beschädigt** werden.

Deine Aufgabe ist es, bei Bedarf die Temperatur zu senken, indem du den Behälter bearbeitest.

ZURÜCK VOR

Überwachung der Reaktionscontainer

Da zur Steuerung der chemischen Anlage mehrere Aufgaben gleichzeitig bearbeitet werden müssen, steht dir zur Unterstützung ein Alarmsystem für die Kontrolle der Reaktionscontainer zur Verfügung.

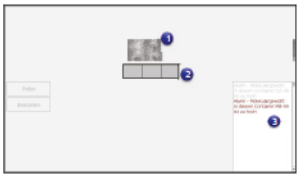
ZURÜCK VOR

Überwachung der Reaktionscontainer

Diese Aufgabe wird sich auf deinem Bildschirm in der rechten unteren Ecke befinden:

Die Container ❶ wandern nacheinander über den Monitor.

Das Alarmsystem analysiert den Containerinhalt und zeigt seine Diagnose über ein entsprechendes Farbsignal unterhalb des Containers ❷ und in einem Zustandmonitor ❸ an.



ZURÜCK VOR

Alarmsystem

Das Alarmsystem zeigt dir an, wenn das Molekulargewicht in einem Container zu hoch ist. Das System arbeitet mit drei Anzeigestufen:

Wenn das Alarmsystem „Rot" anzeigt, lautet seine Diagnose im Zustandsmonitor ❸: das Molekulargewicht in diesem Container ist zu hoch.

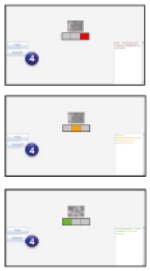
Wenn das Alarmsystem „Orange" anzeigt, lautet seine Diagnose im Zustandsmonitor ❸: das Molekulargewicht in diesem Container eventuell zu hoch ist.

Wenn das Alarmsystem „Grün" anzeigt, lautet seine Diagnose im Zustandsmonitor ❸: das Molekulargewicht in diesem Container ist in Ordnung.

ZURÜCK VOR

Überwachung der Reaktionscontainer

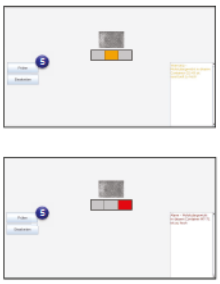
Wenn du eine bestimmte Meldung bekommst, muss du entscheiden, was zu tun ist.
Wenn das Molekulargewicht im Container **zu hoch** ist, muss die Temperatur gesenkt werden. Du kannst die Temperatur senken, indem du auf den **„Bearbeiten"**-Button 4 klickst.
Ist das Molekulargewicht im Container ok, musst du nichts tun.
Beachte dabei, dass das Alarmsystem grundsätzlich zuverlässig ist, aber nicht perfekt arbeitet. Es kann auch Fehler machen.



ZURÜCK VOR

Überwachung der Reaktionscontainer

Du kannst die Diagnose des Prüfsystems überprüfen, indem du auf den **„Prüfen"**-Button 5 klickst. Es öffnet sich ein Bild der Innenansicht des Containers.



ZURÜCK VOR

Überwachung der Reaktionscontainer

Ist der Inhalt des Containers 6 **ok**, so ist **kein „K"** im Container enthalten. Falls der Inhalt des Containers **verunreinigt** ist, ist **ein „K"** enthalten.
Ist ein „K" erkannt worden, solltest du den Container **„Bearbeiten"** 7 . Ist der Inhalt des Containers ok, beende die Überwachung mit **„Weiter"** 8 . Während des Prüfvorgangs können keine anderen Aufgaben bearbeitet werden.

ZURÜCK VOR

Überwachung der Reaktionscontainer

In den späteren Versuchsdurchgängen werden dir für jeden **Fehler bei der Überwachung 2 Punkte** von Deinem Lohnkonto **abgezogen**.

Es ist sowohl ein Fehler, wenn Du einen Container nicht bearbeitest, dessen Inhalt verunreinigt ist, als auch wenn Du einen Container bearbeitest, dessen Inhalt ok ist.

Wenn Du die Diagnose des Prüfsystems überprüfst (auf den „Prüfen"-Button klickst), werden dir für die Verwendung **keine Punkte** von Deinem Lohnkonto **abgezogen**.

Während des Prüfvorgangs können keine anderen Aufgaben bearbeitet werden und die Aufgaben laufen währenddessen weiter.

ZURÜCK VOR

Überwachung der Reaktionscontainer

Jetzt kannst du die Aufgabe 1 Minute lang üben.

Klicke bitte auf „ÜBEN" und bearbeite die Aufgabe bis die Übung von selbst endet.

ÜBEN

Wenn du die Übung beendet hast, klicke bitte auf „VOR".

ZURÜCK VOR

Steuerung der chemischen Anlage

Nachdem du nun die drei Aufgaben kennen gelernt hast, wirst du die Möglichkeit haben alle Aufgaben zusammen 2 Minuten lang zu üben.

Klicke bitte auf „ÜBEN" und bearbeite die Aufgaben bis die Übung von selbst endet.

ÜBEN

Wenn du die Übung beendet hast, klicke bitte auf „VOR".

ZURÜCK VOR

Fragen?

Hast du alles verstanden? Ist alles klar?

Bitte beantworte nun den **Fragebogen 1**, die auf deinem Tisch liegt.

Es ist wichtig, dass du alles verstanden hast, bevor du das Experiment weitermachst. Falls du noch Fragen hast oder bei etwas nicht sicher bist, wende dich bitte an den Versuchsleiter.

Während des Experiments wird es nicht möglich sein, Fragen zu stellen.

ZURÜCK VOR

Kennenlerndurchgang für das Alarmsystem

Wie bereits erwähnt arbeitet das Alarmsystem sehr gut, aber nicht perfekt. Es kann vorkommen, dass es **Fehler** macht.

Für deine Leistung im Experiment wird es wichtig sein, das Alarmsystem gut zu kennen. Dafür wirst du jetzt die Gelegenheit bekommen, dich mit dem Alarmsystem vertraut zu machen. Mit **demselben System** wirst du auch im anschließenden Durchgang arbeiten. Über die Kopfhörer bekommst du eine Rückmeldung, ob die Entscheidung, die du getroffen hast, richtig oder falsch war. Wenn die getroffene Entscheidung falsch war, hörst du einen Hup-Ton. Eine falsche Entscheidung bedeutet, dass du einen Container bearbeitet hast, dessen Molekulargewicht ok war, oder dass du einen Container nicht bearbeitet hast, dessen Molekulargewicht zu hoch war.

ZURÜCK VOR

Kennenlerndurchgang für das Alarmsystem

Bitte setze nun die Kopfhörer auf

ZURÜCK VOR

Kennenlerndurchgang für das Alarmsystem

Jetzt kannst du mit dem Alarmsystem etwa 8 Minuten lang üben.

Klicke bitte auf „Kennenlernen" und bearbeite die Aufgabe bis der Kennenlerndurchgang von selbst endet. Der Durchgang endet, wenn 50 Container über deinen Bildschirm gelaufen sind.

Kennenlernen

Wenn der Probedurchgang beendet ist, klicke bitte auf „VOR".

ZURÜCK VOR

Fragebogen 2

Bitte setze nun die Kopfhörer wieder ab – du brauchst sie nun nicht mehr, da es in den kommenden Versuchsdurchgängen keine akustische Rückmeldung geben wird.

Bitte beantworte nun den Fragebogen 2, der auf deinem Tisch liegt. Bitte folge der angegebenen Reihenfolge.

Wenn du den Fragebogen ausgefüllt hast, klicke bitte auf „VOR".

ZURÜCK VOR

Auszahlungssystem

Du bekommst auf jeden Fall die Basisbezahlung von 5 Euro für deine Teilnahme. Zusätzlich kannst du eine leistungsabhängige Bezahlung von bis zu 4 Euro dazuverdienen.

So funktioniert das Auszahlungssystem:
Im folgenden Durchgang bekommst du

-1,5 Punkte für jede richtige Bestellung von **Katalysatoren**
-7,5 Punkte für jedes Paar frisch befüllter **Kühlwasserbehälter**
-2 Punkte Abzug für jede falsche Reaktion bei der **Überwachung der Container**

Am Ende wird jeder Punkt mit **1 Cent** vergütet.

ZURÜCK VOR

Versuchsdurchgang

Bitte denke daran, dass du für alle drei Aufgaben zugleich zuständig bist:
Katalysatorenbereitstellung,
Kühlwasseraufbereitung,
Reaktionscontainerüberwachung.

Deine Aufgabe ist es, so viele Chemikalien wie möglich zu bestellen und in so vielen Behältern wie möglich das Kühlwasser auszutauschen.
Deine Aufgabe ist es außerdem keine Fehler bei der Containerüberwachung zu machen.
Alle drei Aufgaben sind wichtig.

ZURÜCK VOR

Versuchsdurchgang

Der Versuchsdurchgang dauert ca. 20 Minuten und endet von selbst.

Klicke bitte auf „START" und bearbeite die Aufgaben bis der Durchgang von selbst endet. Du kannst nun Punkte für den Leistungszuschlag sammeln.

START

Wenn du den Durchgang beendet hast, klicke bitte auf „VOR".

ZURÜCK VOR

Fragebogen 3

Wir möchten dir nun gerne einige weitere Fragen stellen. **Bitte beantworte nun den Fragebogen 3, der auf deinem Tisch liegt.** Bitte folge der angegebenen Reihenfolge.

Wenn du mit der Beantwortung der Fragen fertig bist, klicke bitte auf „VOR".

ZURÜCK VOR

Fragebogen 4

Wir möchten dir nun gerne einige weitere Fragen stellen. **Bitte beantworte nun den Fragebogen 4, der auf deinem Tisch liegt.** Bitte folge der angegebenen Reihenfolge.

Wenn du mit der Beantwortung der Fragen fertig bist, klicke bitte auf „VOR".

ZURÜCK VOR

Fragebogen 5

Wir möchten dir nun gerne einige weitere Fragen stellen. **Doppelklicke** daher nun bitte auf „FRAGEBOGEN". Es wird sich eine Internetseite mit dem Fragebogen öffnen.

FRAGEBOGEN

Wenn du mit der Beantwortung der Fragen fertig bist, klicke bitte auf „VOR".

ZURÜCK VOR

Danke für deine Teilnahme!

Danke, dass du mitgemacht hast. ☺

ANNEX E

PPV QUESTIONNAIRE

Deine Einschätzung der Zuverlässigkeit

Du hast gerade 50 Container überprüft, wobei du jeweils die Container bearbeiten solltest, die fehlerhaft waren. Dabei hat dich das automatische Alarmsystem unterstützt, dessen Anzeige entweder grün, orange oder rot werden konnte. Bitte gib nun zunächst an, wie oft die Anzeige deiner Einschätzung nach jeweils grün, orange und rot war. Bitte achte darauf, dass die Summe der drei Zahlen 50 ergeben muss.

- In wie vielen der 50 Fälle war der Alarm **rot**? _____
- In wie vielen der 50 Fälle war der Alarm **orange**? _____ } Summe: **50**
- In wie vielen der 50 Fälle war der Alarm **grün**? _____

Vielleicht ist dir aufgefallen, dass ein Container nicht immer fehlerhaft war, wenn ein Alarm angezeigt wurde. Auf der anderen Seite musste ein Container nicht unbedingt in Ordnung sein, wenn kein Alarm erfolgte. Das liegt daran, dass das Alarmsystem mit seiner Diagnose nicht immer vollkommen richtig liegt.

Bitte schätze nun ein, wie oft ein Container bei den unterschiedlichen Anzeigenfarben jeweils fehlerhaft war, das heißt wie viele der Container jeweils hätten bearbeitet werden müssen. Bitte fülle dazu die Lücken unten aus. Schreibe in die erste Lücke jeweils dieselbe Zahl, die du oben als Häufigkeit für die jeweilige Anzeigenfarbe angegeben hast.

- **Rote Anzeige**

  In wie vielen von den insgesamt _____ Fällen, in denen der Alarm rot war,

  war der Container fehlerhaft, hätte also bearbeitet werden müssen? _____

- **Orangene Anzeige**

  In wie vielen von den insgesamt _____ Fällen, in denen der Alarm orange war,

  war der Container fehlerhaft, hätte also bearbeitet werden müssen? _____

- **Grüne Anzeige**

  In wie vielen von den insgesamt _____ Fällen, in denen der Alarm grün war,

  war der Container fehlerhaft, hätte also bearbeitet werden müssen? _____

ANNEX F

TRUST QUESTIONNAIRE (FMV)

Fragebogen

Es werden dir nun unterschiedliche Aussagen gezeigt, die das Alarmsystem in der unteren rechten Ecke betreffen, mit dem du eben gearbeitet hast. Deine Aufgabe wird es sein, die Aussagen auf einer 4-stufigen Skala von „stimme nicht zu" bis „stimme zu" zu beurteilen.

Lies dir nun bitte zunächst das Beispiel gut durch:

	Stimme nicht zu	Stimme eher nicht zu	Stimme eher zu	Stimme zu
Die farbliche Gestaltung des Systems gefällt mir sehr gut	○	○	◉	○

Hast du diese Antwortoption ausgewählt, bist du der Meinung, dass die farbliche Gestaltung gut ist und nur noch geringe Verbesserungen nötig wären.

Bitte lies dir die Aussagen in Ruhe durch und gib das Ausmaß deiner Zustimmung an. Wenn du dir nicht sicher bist, kreuze diejenige Antwortoption an, die am ehesten deine Meinung widerspiegelt.

Denke bitte daran alle Aussagen zu beantworten und pro Aussage nur eine Antwortmöglichkeit anzukreuzen.

Auf der nächsten Seite geht es los!

Umfrage verlassen und Antworten löschen | Später fortfahren | Weiter ›

LimeSurvey

Fragebogen

Bitte gib nun an, wie stark du den folgenden Aussagen zustimmst.

* Bitte beachte, dass sich die folgenden Fragen immer auf das gesamte Alarmsystem beziehen (nicht auf die einzelnen Anzeigen).

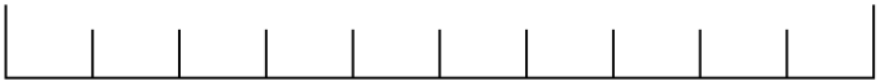
	Stimme nicht zu	Stimme eher nicht zu	Stimme eher zu	Stimme zu
Das Alarmsystem arbeitet sicher	○	○	○	○
Das Alarmsystem verlangsamt meine Arbeit	○	○	○	○
Die Intention des Alarmsystems ist positiv	○	○	○	○
Das Alarmsystem funktioniert gut	○	○	○	○
Die Arbeitsweise des Alarmsystems ist mir klar	○	○	○	○
Das Alarmsystem erschwert meine Arbeit	○	○	○	○
Das Alarmsystem soll helfen, die Gesamtleistung zu verbessern	○	○	○	○
Das Alarmsystem arbeitet genau	○	○	○	○
Ich weiß gut über das Alarmsystem Bescheid	○	○	○	○
Der Zweck des Alarmsystems ist es, mir Arbeit abzunehmen	○	○	○	○
Ich verfüge über alle erforderlichen Informationen zur Bedienung des Alarmsystems	○	○	○	○
Das Alarmsystem ist nützlich für meine Arbeit	○	○	○	○
Das Alarmsystem arbeitet unzuverlässig	○	○	○	○
Ich verstehe, wie das Alarmsystem funktioniert	○	○	○	○
Das Alarmsystem wurde mit dem Ziel implementiert, mir zu helfen	○	○	○	○
Ich empfinde das Alarmsystem als eine Unterstützung bei meiner Arbeit	○	○	○	○

ANNEX G

TRUST ITEM

Gib bitte an, wie stark dein **Vertrauen** zu dem GESAMTEN Alarmsystem ist, mit dem du gerade gearbeitet hast. Bitte setze dafür ein Kreuz an eine beliebige Stelle auf der Linie.

Kein Vertrauen **Starkes Vertrauen**



ANNEX H

NASA-TLX

Fragebogen 4
Erklärung der Dimensionen

Geistige Anforderungen

Wie viel geistige Anstrengung war bei der Informationsaufnahme und bei der Informationsverarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen …)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erfordert sie hohe Genauigkeit oder ist sie fehlertolerant?

Körperliche Anforderungen

Wie viel körperliche Aktivität war erforderlich (z.B. ziehen, drücken, drehen, steuern, aktivieren …)? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erholsam oder mühselig?

Zeitliche Anforderungen

Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam oder geruhsam oder schnell und hektisch.

Ausführung der Aufgaben

Wie erfolgreich haben Sie ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?

Anstrengung

Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?

Frustration

Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?

Fragebogen 4

Geben Sie bitte an, wie hoch Ihre Beanspruchung in den einzelnen Dimensionen war. Die Erklärung dafür finden Sie auf einem separaten Blatt. Markieren Sie dazu auf den folgenden Skalen bitte, in welchem Maße Sie sich in den 6 Dimensionen von der Aufgabe beansprucht oder gefordert gesehen haben:

Beispiel:

Geistige Anforderungen

Körperliche Anforderungen

Zeitliche Anforderungen

Ausführung der Aufgaben

Anstrengung

Frustration