Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation, System Experience, and Operator Functional State

vorgelegt von
Dipl.-Psych.
Juliane Reichenbach
geb. in Berlin

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

Doktor der Philosophie
– Dr. phil. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Matthias Rötting
Gutachter: Prof. Dr. Dietrich Manzey
Gutachter: Prof. Dr. Hartmut Wandke

Tag der wissenschaftlichen Aussprache: 17. Juni 2015

Berlin 2015
D83

Für Friedrich

This dissertation is based on research that has been published before by SAGE Publications. Parts of the methodology and results sections of studies I-III are verbatim reproductions of these publications. All figures in the results sections of studies I-III were created by the author based on the data. A figure in this dissertation is marked with an asterisk * if the result was presented in a figure in these prior publications.

Manzey, D., Reichenbach, J., & Onnasch, L. (2008). Performance consequences of automated aids in supervisory control: The impact of function allocation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52,* 297-301. Santa Monica, CA: Human Factors and Ergonomics Society.
DOI: 10.1177/154193120805200421
http://pro.sagepub.com/cgi/content/abstract/52/4/297

Manzey, D., Reichenbach, J., & Onnasch, L. (2009). Human performance consequences of automated decision aids in states of fatigue. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53,* 329-333. Santa Monica, CA: Human Factors and Ergonomics Society.
DOI: 10.1177/154193120905300435
http://pro.sagepub.com/cgi/content/abstract/53/4/329

Reichenbach, J., Onnasch, L., & Manzey, D. (2010). Misuse of automation: The impact of system experience on complacency and automation bias in interaction with automated aids. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54,* 374-378. Santa Monica, CA: Human Factors and Ergonomics Society.

DOI: 10.1177/154193121005400422

http://pro.sagepub.com/cgi/content/abstract/54/4/374

Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human performance consequences of automated decision aids in states of sleep loss. *Human Factors, 53,* 717-728.

DOI: 10.1177/0018720811418222

http://hfs.sagepub.com/cgi/content/abstract/53/6/717

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

DOI: 10.1177/1555343411433844

http://edm.sagepub.com/cgi/content/abstract/6/1/57

Zusammenfassung

Mit der Einführung von Automation sollen Leistungsvorteile erzielt, eine Erhöhung der Systemsicherheit und Effizienz erreicht und Kosten reduziert werden. Leider ist sie oft auch von negativen Effekten wie *complacency* und *automation bias*, Verlust des Situationsbewusstseins und Fertigkeitsverlusten begleitet. Automatisierung ist keine Alles-oder-Nichts Entscheidung, und der Grad der Automatisierung beeinflusst sowohl die Vorteile als auch die Nachteile. Während Leistungsvorteile sich besonders bei hohen Automationsstufen zeigen, gibt es keine so klaren Ergebnisse darüber, welche Automationsstufe die optimale ist, um negative Automatisierungsfolgen zu vermeiden.

In dieser Arbeit wurde in einer ersten Studie der Einfluss des Automationsgrades (*degree of automation*, DOA) auf die Leistung bei zuverlässiger Automationsunterstützung wie auch bei Automationsversagen untersucht. Die Ergebnisse zeigen, dass sowohl Primär- und Sekundäraufgabenleistung als auch *workload* von der Automationsunterstützung profitieren. Die Leistungsvorteile waren höher bei höherer DOA. Allerdings wurden bei der höchsten DOA auch Fertigkeitsverluste gefunden, wenn wieder ohne Unterstützung der Automation gearbeitet werden musste. Es wurde kein Einfluss von DOA auf *automation bias* gefunden.

In einer zweiten Studie wurde der Einfluss der Systemerfahrung untersucht. Die Erfahrung von Automationsfehlern führte zu einem drastischen Einbruch des Vertrauens und einer stärkeren Überprüfung der Automation. Frühe Fehlererfahrung reduzierte das Risiko von *complacency* und *automation bias*, konnte es aber nicht

komplett verhindern. Desweiteren wurden in der Studie drei mögliche Ursachen von *commission errors* identifiziert:

a) die Automation wird unvollständig überprüft,

b) die Automation wird vollständig überprüft, aber widersprechende Information nicht bewusst verarbeitet,

c) widersprechende Information wird bei der Entscheidung nicht berücksichtigt.

In der dritten Studie wurden Effekte von Bedienerzustand und DOA untersucht. Mit Unterstützung höherer Automation konnte die Leistung während der Nacht besser erhalten werden. Allerdings waren die Leistungseinbußen bei Ausfall der Automation stärker als bei niedrigerem Automationsgrad. Die Automation wurde nachts mehr überprüft als tagsüber, und das Risiko eines *commission error* war nachts geringer.

Abstract

Introducing automation intends to yield performance benefits, increase system safety and efficiency, and decrease costs. However, the intended benefits are often offset by negative effects such as complacency and automation bias, loss of situation awareness, and skill degradation. Automating systems is not an all-or-none decision, and degree of automation (DOA) has been shown to influence both intended performance benefits as well as performance decrements. While performance benefits more from higher automation, results about the optimal degree of automation in terms of preventing negative effects are not so clear-cut.

In a first study, we investigated effects of degree of automation on routine performance and failure performance using a simulated process control task. Results show that primary and secondary task performance as well as workload benefit from providing automation, and performance gains were higher for higher DOA. However, skill degradation was observed for the highest DOA when returning to manual performance. Regarding automation bias, there was no effect of DOA.

The second study focused on the effects of system experience. Failure experience led to a strong decrease in trust, also reflected in more intense automation verification. Early failure experience reduced the risk of complacency and automation bias but did not prevent it completely. Moreover, we identified three possible causes of commission errors:

a) incomplete automation verification,

b) complete automation verification without attentive processing of contradictory information, analogous to a looking-but-not-seeing effect,

c) discounting of contradictory system information.

The third study looked into effects of operator functional state and DOA. Higher DOA could better protect performance after extended wakefulness, but return-to-manual performance suffered more. More information was sampled for automation verification during the night, and the risk of commission errors was lower compared to daytime performance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the past decades, automation support in complex work environments has increased consistently. Especially operation of systems that support decision making, such as cockpit warning systems in aviation, navigation systems in automobiles, image-based assistance systems in medicine, or diagnostic support systems in process control, has proliferated.

Moray, Inagaki, and Itoh (2000, p. 44) define automation as "any sensing, detection, information-processing, decision-making, or control action that could be performed by humans but is actually performed by machine". Parasuraman and Riley (1997) define automation as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" (Parasuraman & Riley, 1997, p. 231). There are different reasons for automating tasks that could be performed by humans or have been performed by humans in the past. Some tasks are dangerous or difficult to perform, others are extremely monotonous and repetitive. Some task are automated just because it is technically feasible. Automating tasks is intended to increase efficiency and reduce costs, increase system

reliability and performance, and to decrease the operator's workload. However, automation might not only lead to the intended benefits but possible benefits might be offset by unwanted performance consequences (Bainbridge, 1983; Endsley & Kiris, 1995; Sarter, Woods, & Billings, 1997).

Automation does not only reduce the human's work but changes the nature of the work. An executive task becomes a supervisory control task (Sheridan & Verplank, 1978; Sheridan, 1992). The human operator supervises the automation when it works reliably, has to detect errors of the automation and take over manual control in case of malfunctions. Early on, Bainbridge (1983) reported on possible adverse effects of automation. The operator can loose manual skills needed to perform the automated functions as he only takes over manual control when the automation fails, but during normal operation merely supervises the automation. However, during abnormal operation or automation failure the operator has to take appropriate action instantly. This is aggravated by the fact that system errors are often obscured by the automation's attempt to handle the error, so the error might become apparent only late in the process when it is already hardly controllable.

Other possible performance consequences include overreliance on the automation. In the context of supervisory control, overreliance behaviorally shows in inappropriate monitoring of automated functions, a phenomenon which has been referred to as automation-induced complacency (Parasuraman, Molloy, & Singh, 1993). Complacency can lead to loss of situation awareness and loss of skills (Endsley & Kiris, 1995; Parasuraman et al., 1993). In the context of automated decision aids, overreliance is associated with automation bias (Mosier & Skitka, 1996). Automation bias can lead to omission and commission errors.

Decisions about automating tasks are not all-or-none decisions. Not only can

different subtasks be automated while others are not, also those automated subtasks can be automated to different degrees. Benefits and costs of automation depend on function allocation. The key criteria for evaluating automation design should be the human performance consequences for specific degrees of automation (Parasuraman, Sheridan, & Wickens, 2000; Parasuraman, 2000). Those performance consequences include workload, situation awareness, complacency, and skill degradation. Secondary evaluative criteria include, among others, automation reliability, costs of action outcomes, implementation costs, and liability.

Stressors like noise or sleep deprivation affect the interaction with automation. Under stress, operators show a preference for higher levels of automation (Sauer, Kao, Wastell, & Nickel, 2011; Sauer, Nickel, & Wastell, 2013). Automation support can help reduce the negative effects of stressors like noise on performance and workload (Sauer et al., 2011; 2013). For sleep-deprived operators in supervisory control tasks, a shift towards less demanding system management strategies has been shown (Hockey, Wastell, & Sauer, 1998; Sauer, Wastell, Hockey, & Earle, 2003). In highly automated workplaces a lot of accidents happened during night shifts when operators are sleepy (Folkard, Lombardi, & Tucker, 2005). Despite its practical relevance, there is only little research on the moderating effects of operator functional state on human performance in interaction with automation, with effects of stress and fatigue being the most important in this respect.

In the following sections the related work is reviewed. In chapter 2 the simulation of a supervisory control task that was used in the experiments is introduced. The three experiments that were conducted are presented in chapters 3 - 5. A general discussion of the findings follows in chapter 6.

1.1 Function Allocation & Operator Performance

How should tasks be allocated to human or automation? Which functions should be automated, which tasks should be worked on by humans? After deciding which subtasks should be automated, it has to be decided to which degree it should be automated, ranging from manual to fully automated. There are different approaches to allocate functions to human or automation. Performance benefits and costs of automation depend on function allocation.

1.1.1 Function Allocation

The technology-centered approach focuses the technical feasibility, technical reliability and costs. Functions are automated if cost reducing technical solutions can be built, other functions are left to the human. Obviously, this approach will not always provide the best solution. Related subtasks are split up between human and automation if one can be automated and one cannot.

In competence-centered views like Fitts MABA MABA lists (Fitts, 1951) tasks are allocated to this part of the human-machine system that is better able to master a task. Tasks that a machine can do more efficiently are allocated to the machine, tasks humans are better at are left to the human. Problematic with this approach is that it might not be clearly decided which part of the system is better able to do a certain task; also abilities might change over time, especially technology constantly becomes more capable. In addition, the human-machine interaction is disregarded, but a lot of problems occur in the interaction between human and machine.

Human-centered automation (Billings, 1997) intends to support the human op-

erator, keep the human engaged and informed about the system state and ongoing activities, and regards human and machine as part of one system.

Decisions about automation are not all-or-none decisions. Some subtasks might be automated while others are not, and those automated subtasks can be automated to different degrees. Benefits and costs of automation depend on how functions are allocated to human or machine. The primary evaluation criteria should be the human performance consequences for specific degrees of automation (Parasuraman et al., 2000; Parasuraman, 2000). Probability and costs of automation failure should play an important role in the decision about automation (Sheridan & Parasuraman, 2000).

Different models of human-automation interaction have been proposed in the past (e.g., Sheridan & Verplank, 1978; Endsley & Kaber 1999; Parasuraman et al., 2000; Wandke, 2005). In an early seminal work, Sheridan & Verplank (1978) proposed a 10-level model of human-automation interaction, with 10 levels of automation ranging from manual performance to full automation. The 10-level model is shown in Table 1.1, in an adaptation by Parasuraman et al., (2000). Since they did not only rephrase the original list but slightly changed the meaning, the original wording from Sheridan & Verplank (1978) can be found in appendix A.

Endsley and Kaber (1999) presented a level of automation taxonomy assigning four functions to either human or machine. Those four functions are (1) monitoring display information to perceive system status, (2) generating options or strategies for achieving goals, (3) selecting an option or strategy, and (4) implementing the chosen option. The 10 levels are shown in Table 1.2. Endsley and Kaber (1999) note that the order is not necessarily ordinal.

Parasuraman et al. (2000) took this idea one step further and presented a

Table 1.1: The 10-Level Model of Human-Automation Interaction (Sheridan & Verplank, 1978), adapted from Parasuraman, Sheridan, & Wickens (2000, p. 287). For original wording from Sheridan & Verplank (1978), see appendix A.

LOW	1	The computer offers no assistance: human must take all decision and actions.
	2	The computer offers a complete set of decision/action alternatives, or
	3	narrows the selection down to a few, or
	4	suggests one alternative
	5	executes that suggestion if the human approves, or
	6	allows the human a restricted time to veto before automatic execution, or
	7	executes automatically, then necessarily informs humans, and
	8	informs the human only if asked, or
	9	informs the human only if it, the computer, decides to.
HIGH	10	The computer decides everything and acts autonomously, ignoring the human.

model that distinguish four types of functions, and each function can be automated to a different level (see Figure 1.1). The four types of functions that can be automated, corresponding to a four-stage model of human information processing, are (1) information acquisition, (2) information analysis, (3) decision & action selection, and (4) action implementation. Examples of resulting automation profiles are depicted in Figure 1.1. In system A, information acquisition and information analysis are highly automated while decision & action selection and action implementation are manual. Such a system could be an alarm system that acquires information via sensors, analyzes them, and raises an alarm when certain values are below or above a predefined threshold. Another system B could additionally have a highly automated action implementation while decision & action selection is automated on a low level. Such a system would implement the actions after the

Table 1.2: Level of Automation Taxonomy, adapted from Endsley & Kaber (1999, p. 466).

LOA	Monitoring	Generating	Selecting	Implementing
Manual Control	Human	Human	Human	Human
Action Support	H/C	Human	Human	H/C
Batch Processing	H/C	Human	Human	Computer
Shared Control	H/C	H/C	Human	H/C
Decision Support	H/C	H/C	Human	Computer
Blended Decision Support	H/C	H/C	H/C	Computer
Rigid System	H/C	Computer	Human	Computer
Automated Decision Making	H/C	H/C	Computer	Computer
Supervisory Control	H/C	Computer	Computer	Computer
Full Automation	Computer	Computer	Computer	Computer

Note. H/C: Human/Computer

operator selected an action.

Wickens, Li, Santamaria, Sebok, & Sarter (2010) suggested that a higher *degree of automation* (DOA) can be reached by automating a later stage and / or implementing a higher level of automation within one stage. So system B from the example in Figure 1.1 would have a higher degree of automation than system A. Onnasch, Wickens, Li, & Manzey (2014) differentiate four cases of varying DOA, pure level, pure stage, aggregation, and confound. They argue that in the first three cases it can clearly be said which of two automated systems is automated to a higher degree. For example, in the *pure level* case, two systems are automated at the same stage at different levels. The system automated at a higher level has a higher DOA. The forth case, the *confound* case, compares a system in which an early stage is automated at a high level with a system in which a later stage is automated at a lower level. In this case, it cannot clearly be decided which system

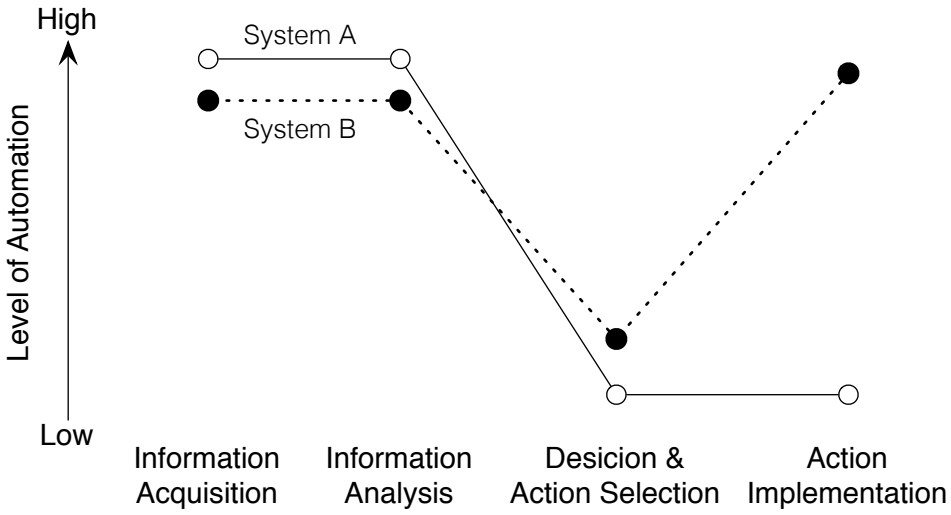

Figure 1.1: Types and Levels of Automation, adapted from Parasuraman, Sheridan, & Wickens (2000, p. 288), and examples with different automation profiles.

has a higher DOA.

Since the effect of automation on human performance and joint human-system performance is different when automation works reliably versus when the automation fails, it is important to look at both routine performance and failure performance. Figure 1.2 displays the hypothesized routine-failure trade-off, the relationship between degree of automation, routine performance, failure performance, workload and situation awareness proposed by Wickens and colleagues (Wickens et al., 2010; Onnasch et al., 2014). The figure shows that routine performance and workload benefit from higher degrees of automation while situation awareness and failure performance decline with increasing degree of automation. If failure performance starts decreasing only at the DOA marked at point (A) in the graph, this would be the optimal DOA with the best routine performance combined with the lowest possible workload, just before detrimental effects on failure performance come into effect. However, Wickens et al. (2010) remark that the form of the

curves is hypothetical as there is not yet sufficient empirical evidence.

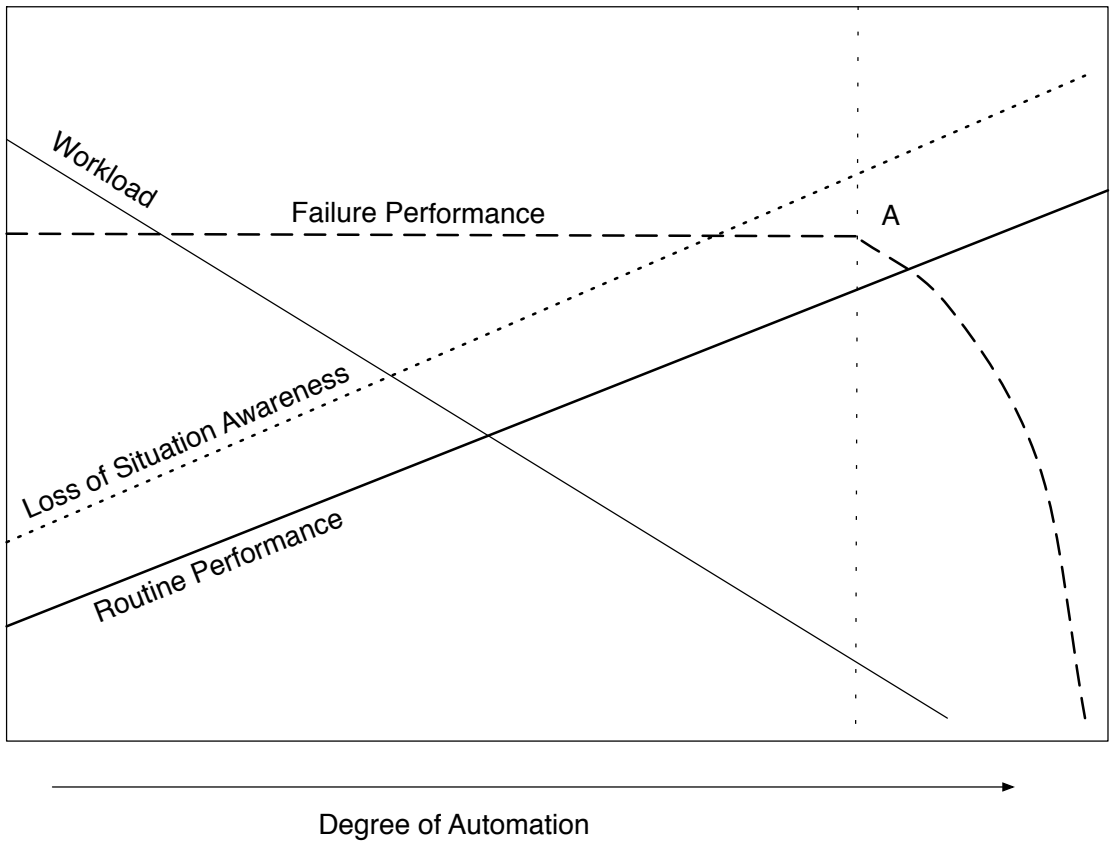

Figure 1.2: Trade-off of Routine and Failure Performance, Workload and SA, with Degree of Automation, adapted from Wickens, Li, Santamaria, Sebok, & Sarter (2010, p. 390)

1.1.2 Degree of Automation and Operator Performance

A number of studies examined the influence of degree of automation on performance benefits and costs of automation. While performance benefits increase with increasing degree of automation, findings on performance costs are more ambiguous. Some studies found benefits of medium DOA for situation awareness and manual skills, others found favorable effects of higher DOA. Note that in most

studies the term *level of automation* (LOA) was used, not the term *degree of automation* (DOA). As a higher level of automation also represents a higher degree of automation, we use both terms in this work. Both terms imply increased automation at a higher level or degree of automation.

Endsley and Kiris (1995) investigated effects of level of automation on situation awareness and decision performance in an automobile navigation task that was assisted by an expert system. Participants were presented a problem description and three possible actions to choose from. Five levels of automation were compared. In the manual condition, participants were only shown the possible actions. In the decision support condition, they were additionally presented the system's assigned probabilities for each alternative that this was the correct solution for the problem. In the consensual AI condition, the system preselected one option that the participant could confirm or select another option. In the monitored AI condition, the participant could only veto the preselected option within 30 seconds. In the full automation condition the participants merely monitored the automation. After automation breakdown, participants had to make all decisions manually. Results show that decision accuracy remained at a very high level also after automation failure, but decision time was longer immediately after automation breakdown. Participants with prior full automation had longer decision times and poorer situation awareness. Participants with prior medium levels of automation showed medium levels of situation awareness.

In another study by Endsley and colleagues (Endsley & Kaber, 1999), the ten levels of automation (LOA) from the automation taxonomy described earlier were studied. In a complex control task involving object collision avoidance, participants either worked manually or with the support of one of nine automations ranging

from action support to full automation. Performance under reliable automation was better in the supported groups than the manual group. Participants profited most from implementation assistance. Performance was worse when generating options was shared than when it was done solely by the human or the machine. Automation failure performance was not different from performance of the manual group, except for some levels on different performance measures, but no general trend was observed favoring a certain level. Time to recovery was better for LOA that required human action in the implementation. Higher LOA for which the selecting function was allocated to the machine or was shared by human and machine was associated with better situation awareness and lower workload.

Extending this line of research, Kaber & Endsley (2004) studied the object collision avoidance task with participants additionally performing a secondary monitoring task. Manual and automation supported phases alternated. LOA had no effect on primary or secondary task performance, situation awareness or workload in manual trials. In automated trials, low and intermediate automation were associated with better performance, intermediate LOA supported higher situation awareness.

In a study by Sarter & Schroeder (2001) pilots were provided with an automated decision aid for in-flight icing events. Two decision aids of different degrees of automation were compared. The lower automated aid provided information about icing events but left the action selection to the pilot, reflecting an automation support at the stage of information acquisition and analysis (Parasuraman et al., 2000). The higher automated aid provided support for decision making and action selection in addition to providing information about icing conditions. Both aids increased the number of correct decisions about icing events compared to a baseline

condition without an automated aid, given the aid provided correct information. However, pilots also followed the aid's recommendation if it was wrong, which led to performance decrements compared to the baseline condition in case the aid's information was inaccurate. This adverse effect was even stronger for the higher automated aid.

Rovira, McGarry, & Parasuraman (2007) studied the influence of degree of automation of an automated aid on the performance in a command and control task using an information automation and three levels of decision automation. They found that reaction times decreased compared to manual condition when the automation provided correct information but increased when it provided inaccurate information. Decision-making accuracy was also impaired by inaccurate automation advice. The performance costs were greater for the higher DOA decision aids than the information automation when the overall automation reliability was high (80%). When overall automation reliability was lower (60%), performance suffered in both information and decision automation when the automation provided inaccurate advice.

Performance benefits of higher DOA in case of automation failure were found by Lorenz and colleagues (Lorenz, Di Nocera, Roettger, & Parasuraman, 2002). They studied degree of automation effects in a simulated process control task, comparing three automated aids that supported fault identification and management. The first aid was a fault finding guide, the second aid automatically provided diagnoses and action recommendations for a given system fault, the third aid additionally implemented the appropriate actions if not vetoed by the operator. Benefits in diagnostic accuracy and fault identification time were higher for medium and high DOA with reliable automation support. Return-to-manual performance after an

automation failure was better for the higher automated aid. Analysis of the verification behavior of the operators showed that in the higher automated condition, operators sampled more information which helped them keep up system awareness. However, time pressure differed between the conditions: operators in the veto condition were shown a countdown until the fault management implementation would automatically start. A sensible way to use this waiting time was to cross-check other system information.

Metaanalyses by Wickens et al. (2010) and Onnasch et al. (2014) which aggregated data from 14 (Wickens et al., 2010) and 18 studies (Onnasch et al., 2014), respectively, (some of which were described earlier in this work) show that an increasing degree of automation benefits routine performance, but on the downside failure performance suffers. Effects on workload are not so clear-cut, but several studies show that higher DOA decreases workload. Onnasch et al. (2014) suggest that the human operator should be kept actively involved in the decision making process as that supports situation awareness and leaves the operator less vulnerable to the negative effects on performance when the automation fails.

1.2 Trust, Distrust, and Overtrust in Automation

The reliability of a system is one of the most important factors influencing trust in automation (Sheridan, 1988; Moray et al., 2000). Other important factors are robustness, familiarity, understandability, usefulness, dependability, trust in one's own capabilities, and individual differences in attitude towards technology

(Sheridan, 1988; Lee & Moray, 1992, 1994).

Trust only changes slowly as long as the automation works reliably and no errors occur (Moray et al., 2000). However, when an automation failure occurs, trust decreases instantly and only recovers slowly (Lee, 1991; Lee & Moray, 1992; Moray et al., 2000).

In the ideal case, trust in automation is appropriately calibrated (Lee & See, 2004). Appropriate trust is a prerequisite for appropriate use and monitoring of automation in accordance with the automation's characteristic features like reliability. Presenting dynamic system confidence information can help calibrate trust appropriately and enhance performance (McGuirl & Sarter, 2006).

On the other hand, trust can be too low which can lead to disuse of an automation, or it can be too high which can lead to misuse of an automation (Parasuraman & Riley, 1997; Lee & See, 2004). Both cases of inappropriate reliance on automation can lead to adverse effects in the human-automation interaction.

Distrust often develops when the reliability of an automation is underestimated after the experience of automation errors, especially when the automation makes errors while performing an easy task (Madhavan, Wiegmann, & Lacson, 2006) or when the perceived or real costs of automation errors are high. If automation errors can be explained and are predictable, they have less effect on trust (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003).

Distrust and disuse of automation is often seen in relation with warning systems. A fatal accident happened at a railway crossing when a driver ignored the warning about an approaching train and did not stop at the crossing. Prior to the accident, the driver repeatedly experienced an alarm warning about an approaching train, but there was no train coming (S. R. Dixon, personal communication).

Warning systems are often used in contexts where the consequences of missing a critical situation are associated with high costs. Accordingly, thresholds are set low in order to not miss critical states. High sensitivity, however, also leads to high false alarm rates. When at the same time the base rate of system errors or critical states is low, as it is in highly reliable technical systems, the probability that a critical state is present when a warning systems indicates so, is very low (Parasuraman, Hancock, & Olofinboba, 1997). The repeated experience of alarms emerging in situations with no underlying critical system state can lead to a slowed reaction or a complete disregard of alarms. This phenomenon is known as the cry wolf effect (Breznitz, 1983).

As a consequence of distrust and disuse of automation the intended support of the operator and reduction of workload may not be achieved. This may also have adverse effects on system safety.

Just as distrust in automation, overtrust can have adverse effects. The operator fails to recognize the system's limitations and relies on the automation more than system capabilities would justify. The operator monitors and controls the automation insufficiently or uses the automation as a decision heuristic (Parasuraman & Riley, 1997). This can compromise system safety.

Overtrust may develop in interaction with systems that have a high and constant reliability (Parasuraman et al., 1993), when operators have a positive attitude towards technology (Singh, Molloy, & Parasuraman, 1993a, b), or under multiple-task conditions when the operator's workload is high (Parasuraman et al., 1993).

1.3 Complacency & Automation Bias

1.3.1 Complacency

Complacency has been identified as an issue in cockpit automation, and it has been a contributing factor to a number of incidents and accidents in aviation (Billings, Lauber, Funkhouser, Lyman, & Huff, 1976; Wiener, 1981; Funk et al., 1999). Complacency is related to overtrust in an automated system, behaviorally shows in insufficient monitoring of an automation, and can result in problems such as missing or delayed reaction to automation failures, loss of situation awareness, and loss of skills (Parasuraman et al., 1993; Moray & Inagaki, 2000; Moray, 2003; Parasuraman & Manzey, 2010).

Previous research suggests that complacency depends on a number of factors of the automation, the situation, and the person, such as reliability and consistency of the automation, demands of the operator's concurrent tasks (Parasuraman et al., 1993; Molloy & Parasuraman, 1996; Singh, Molloy, & Parasuraman, 1997), complexity of the monitoring task itself (Thackray & Touchstone, 1989; Bailey & Scerbo, 2007; Kerstholt, Passenier, Houttuir, & Schuffel, 1996), training and failure experience (Manzey, Bahner, & Hüper, 2006; Bahner, Hüper, & Manzey, 2008; Bahner, Elepfandt, & Manzey, 2008; Aust, Moehlenbrink, & Jipp, 2011), and characteristics of the individual (Prinzel, DeVries, Freeman, & Mikulka, 2001; Szalma & Taylor, 2011).

Parasuraman et al. (1993) studied the effects of automation reliability and task characteristics. They found no difference between high and low reliability systems, but complacency was lower under variable reliability than under constant reliability, with a higher probability of detecting an automation failure when the

system's reliability was changing. However, complacency was not found when the operator's only task was a monitoring task, it was only found under multiple task conditions. Similar results were found by Molloy & Parasuraman (1996) and Singh et al. (1997) working with similar tasks (MAT battery). When an automation failure occurred late in the simulation, more participants missed the error than when it occurred early in the simulation, again only under multiple task conditions (Molloy & Parasuraman, 1996). In the Singh et al. (1997) study, the automated task that had to be supervised was centrally located on the screen instead of peripheral. Central location did not prevent complacency.

Similar results were found with a simulated driving task with automated steering and lateral control (de Waard, van der Hulst, Hoedemaeker & Brookhuis, 1999). A single automation failure at the end of the simulation was not detected by most participants, or detected too late to take over manual control in time to prevent a collision.

While the previously mentioned studies (Parasuraman et al., 1993; Molloy & Parasuraman, 1996; Singh et al., 1997) measured complacency as detection rate of automation failure, Moray and colleagues suggest that complacency should be measured by examining monitoring behavior, not detection rate (Moray & Inagaki, 2000; Moray, 2003). In a gedankenexperiment they demonstrate that even a perfect observer can miss a signal if he has to monitor two screens with only one screen being visible at a time. If a signal occurs on both screens at the same time, he can only see one and will miss one signal. Another problem with using detection rate for measuring complacency is that, in case of a multiple task scenario, an operator might notice an automation failure but be too busy with the concurrent tasks or find them to be more important than dealing with the automation failure. So

missing an automation failure does not necessarily indicate insufficient monitoring of an operator, and not reacting to an automation failure does not necessarily mean it was not detected. Payoffs and importance of different tasks must be explicit to the operator. According to Moray (Moray & Inagaki, 2000; Moray, 2003), a complacent operator is an operator who samples less information than would be optimal. As a consequence, optimal sampling needs to be defined when examining complacency. Depending on the task, this might not always be possible.

Bagheri & Jamieson (2004a) replicated the Parasuraman et al. (1993) study and in addition tracked eye movements to allow studying monitoring behavior. They found that in the condition with constant high reliability, participants looked at the monitoring zone only rarely – and detection performance was poor. Also, monitoring and detection rate was lower when the participant's trust in the automation was higher. In a second study, Bagheri & Jamieson (2004b) explicitly informed the participants about the automation's reliability. When provided with this additional context information, participants fixated the monitoring zone more often and longer. Allocating more attention to the monitoring task led to improved detection rates without compromising secondary task performance.

In a study with air traffic controllers, eye movement was recorded in addition to conflict detection (Metzger & Parasuraman, 2005). Controllers who detected an automation failure (missing to signal a potential conflict) in the manual and automation supported condition showed no difference in fixations of the radar display. In contrast, controllers who missed the automation failure and did not detect the conflict had fewer fixations in the automated condition than in the manual condition.

Manzey and colleagues (Manzey et al., 2006; Bahner, Hüper et al., 2008) stud-

ied the effects of experiencing an automation failure as opposed to merely being informed about the possibility of automation failures. In addition to measuring detection of automation failure, they measured automation verification in accordance with a predefined normative model of information sampling, following Moray's criticism (Moray & Inagaki, 2000; Moray, 2003). This allowed quantifying complacency independent of possible performance consequences in case of automation failure. They found that the experience group was less complacent than the information group; the experience group spent more time on fault identification and sampled more relevant information. However, they followed the wrong diagnosis just as often, there was no difference between experience and information group with respect to commission errors. Contrasting those participants who followed the aid's wrong diagnosis (commission error) with those participants who did not (no commission error) showed that participants with commission error were more complacent already in the preceding trials with correct diagnosis, spending less time on fault identification and sampling less relevant information. In addition, they had a weaker secondary task performance (simple reaction time task) when a system error was present and they had to validate the proposed diagnosis. Secondary task performance (simple reaction time task) did not suffer in the group of participants with no commission error. This suggests that the effect cannot be explained as a tradeoff between primary and secondary task performance.

Other studies showed that failure experience can affect detection rate. Pilots inspecting flight plans detected more errors if they had experienced erroneous flight plans repeatedly than if they had only encountered error free plans before (Aust et al., 2011).

The effects of error experience can be very specific, as Bahner and colleagues

showed (Bahner, Elepfand, et al., 2008; Manzey et al., 2006; Bahner, Hüper et al., 2008). In one study already described before (Manzey et al., 2006; Bahner, Hüper et al., 2008) the alerting function of the aid worked properly, but the diagnostic function failed. This only affected the sampling behavior when system errors where present and a potential commission error. It did not affect sampling behavior in fault-free states. In another study (Bahner, Elepfand et al., 2008), the alerting function of the automated aid failed, so participants were not warned about a system error that had occurred. The diagnostic function of the aid worked properly. The experience of the fallible aid led to increased sampling in error-free phases and less omission errors. It did not affect validation of the aid's diagnosis and commission errors. During training, participants were either informed about the possibility of automation misses or they experienced the automation failing to indicate a system error. During the experiment, the automation failed. This was the first failure experience for the information group. 80% of the information group did not detect the failure or detected it only very late, whereas in the experience group, 18% of the participants failed to detect the error. When a second automation failure occurred, the performance of the information group was similar to that of the experience group. At the end of the simulation, the diagnostic function of the aid failed and the aid provided a wrong diagnosis. 74% of the participants followed the wrong recommendation, making an error of commission, independent of their previous training experience. Out of the participants who made an error of commission, 82% showed varying degrees of complacency, having checked none or only part of the necessary parameters to validate a diagnosis. The remaining 18% followed the aid's advice despite full validation of other system information which contradicted the aid's diagnosis. Participants who committed

an error of omission were not more likely to commit an error of commission.

Also other factors such as the complexity of the monitoring task itself has been found to influence complacency. Thackray and Touchstone (1989) showed that monitoring and detection of critical signals degraded when multiple information sources had to be integrated. Bailey and Scerbo (2007) compared three monitoring tasks that varied in complexity. They found that operator performance decreased with increasing complexity. In a study by Bahner (2008; Bahner, Hüper et al., 2008), 79% of the participants detected an automation failure when it was easy to detect. In this case, checking one parameter was sufficient to realize that the aid provided a wrong diagnosis. In a second experiment (Bahner, 2008; Bahner, Elepfand et al., 2008), the validation procedure was more complex and also involved time-consuming control actions. Only 26% of the participants detected the failure in this case.

1.3.2 Automation Bias

In the context of automated decision aids that support operators with detection and diagnosis of critical systems states, Mosier and Skitka (1996) described another possible consequence of overtrust, namely automation bias. They define automation bias as "the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing" (Mosier & Skitka, 1996, p. 205). Performance consequences of automation bias are errors of omission and errors of commission. Omission errors can happen during seemingly uncritical phases when the warning aid does not notify about malfunctions or critical system states. Trusting the warning aid, the system is not monitored sufficiently during "uncrit-

ical" system state, and a system error or malfunction can be missed or detected too late if not indicated by the aid. On the other hand, commission errors can happen during seemingly critical system states when the aid warns about a system error and proposes a diagnosis. The proposed diagnosis and instructions for dealing with the error are accepted and followed uncritically without examining other available sources of information that could verify or falsify the aid's diagnosis, or despite contradictory information from other sources. In case the aid's diagnosis is wrong, the operator follows a wrong diagnosis and recommendation.

Factors contributing to the occurrence of automation bias are the tendency to minimize cognitive effort in decision making (the cognitive miser hypothesis, Fiske & Taylor, 1991), diffusion of responsibility or social loafing (Domeinski, Wagner, Schoebel, & Manzey, 2007), and trust in automation (Lee & See, 2004) which might lead to an overestimation of the automation's capabilities.

According to the cognitive miser hypothesis (Fiske & Taylor, 1991) humans tend to use cognitive resources economically and prefer strategies that save time and effort in decision making, using simple heuristics (Tversky & Kahneman, 1974; Kahneman, Slovic, & Tversky, 1982; Gigerenzer & Todd, 1999) instead of an effortful analysis of all available information to make a decision. Mosier and Skitka (1996) proposed that automated cues are used as heuristics in place of more demanding information processing.

When sharing responsibility for a task with another human or automation, humans tend to reduce their effort on tasks that are worked on redundantly (Karau & Williams, 1993; Domeinski et al., 2007). The perceived competence of the partner influences the performance. When the partnering automation is believed to be more competent, the own performance suffers whereas performance increases

when the partnering automation is believed to be less competent (Domeinski et al., 2007).

Humans are often positively biased and tend to trust automation. Capabilities of automation are often overestimated and performance is believed to be superior. Participants in a study by Dzindolet et al. (2003) believed the automation would show a good performance even with little information about the automation.

Galletta, Durcikova, Everard, & Jones (2005) examined the influence of an automatic spell checker when editing documents. Results show that the spell checker improved performance when it worked correctly but gave rise to omission and commission errors when it failed to mark typing or grammar errors or falsely marked correct typing as errors. Participants trusted the spell checker more than their own abilities, they overlooked more errors and left more errors uncorrected than when spellchecking a document unaided. The authors attribute the effect to a) the software's credibility (analogous to the trust in automation (Lee & See, 2004) explanation of automation bias), b) to a tendency to avoid effort (analogous to the cognitive miser hypothesis (Fiske & Taylor, 1991), and c) to yielding the responsibility for finding errors to the aid (diffusion of responsibility).

In the area of computer-aided detection (CAD) in radiology, Alberdi and colleagues found evidence for omission and commission errors (Alberdi, Povyakalo, Strigini, & Ayton, 2004; Alberdi, Povyakalo, Strigini, Ayton, & Given-Wilson, 2008; Alberdi, Povyakalo, Strigini, & Ayton, 2009). CAD tools highlight areas on digitized mammograms that the radiologist should attend to more closely. It is thought of as a detection aid that presents radiologists an attention cue, but in reality they are also used as diagnostic cues (Alberdi et al., 2009). More cases were falsely categorized as cancerous if the automation marked a non-pathological

mammogram compared to unaided performance (Alberdi et al., 2008). Detection rates for breast cancer dropped in cases when the automation did not mark mammograms that actually contained signs for cancer or marked it at a wrong position (Alberdi et al., 2004, 2008; Taplin, Rutter, & Lehman, 2006; Zheng et al., 2004).

Mosier, Palmer, & Degani (1992) studied professional pilots' decision making in an engine fire situation in a flight simulator. When the automation falsely recommended to shut down an engine that caught fire, 75% of the participants followed the recommendation when they worked with an automated electronic checklist. In addition, less information was taken into account when making the decision. When working with standard paper checklists, only 25% followed the wrong recommendation.

Also other studies found high rates of omission and commission errors. Mosier, Skitka, Heers, & Burdick (1998) studied omission and commission errors with professional pilots in a part-task flight simulation. The rate of omission errors was related to the importance of the task. Safety-critical errors were more likely to be detected, but there was still a very high omission error rate: almost half of the pilots did not detect the automation error. A communications-related automation error was undetected by 71% of the pilots. When pilots were confronted with an erroneous engine fire warning, all participants followed the wrong recommendation and shut down the engine, despite contradictory information from other cockpit instruments (normal engine parameters, no aural warnings or warning lights). In a debriefing interview 67% of the participants reported seeing cues indicative of the warning proposed by the automation which actually were not there, a phenomenon Mosier et al. (1998) termed "phantom memory". Participants remembering phantom cues were also reported by Mosier, Skitka, Dunbar, & McDonnel (2001).

Skitka, Mosier, Burdick, & Rosenblatt (2000) investigated if specific automation bias training or working in teams vs. working solo could prevent automation bias. They found that participants made less commission errors after explicitly being informed about the problem of automation bias during training. However, the training had no effect on omission errors. Working in teams did not reduce automation bias. In a follow up study with professional pilot crews (Mosier et al., 2001), the automation bias training did not effect omission or commission errors. As in the study before, working in teams did not reduce automation bias and associated errors.

Effects of social accountability on automation verification and automation bias when using an automated monitoring aid were studied by Skitka, Mosier and Burdick (2000). They found that making participants socially more accountable led to an increased rate of automation verification and a decreased rate of automation bias. Also Mosier et al. (1998) found that pilots who felt accountable for their performance were more likely to check other information to validate the automation and were less prone to errors.

Presenting confidence levels could also help lower automation bias. In a study by McGuirl & Sarter (2006), pilots were provided with an automated decision aid for in-flight icing events. Along with the information about icing condition and action recommendations, the automated aid presented dynamically updated information about its level of confidence that its decision is correct. This extra information about level of system confidence led to less deterioration of pilot performance in case of a false recommendations.

1.3.3 Integrated Model of Complacency and Automation Bias

Automation bias as described by Skitka and colleagues (Mosier & Skitka, 1996; Skitka, Mosier, & Burdick, 1999) can partly be seen as a decision bias. If information from different sources are available to judge a situation, computer-generated cues are trusted more and the decision is biased towards the information obtained from the automation. If, however, commission errors occur as a "result of not seeking out confirmatory or disconfirmatory information" (Skitka et al., 1999, p. 993), this is parallel to complacency in supervisory control tasks (Parasuraman et al., 1993; Moray & Inagaki, 2000). The automation is overtrusted, attention is reallocated to other concurrent tasks, and the automation is not validated against other sources of information. Also, omission errors can happen when monitoring is insufficient given a high but not perfect reliability of an automation. System errors or malfunctions can be missed or detected too late if not indicated by the aid.

Based on a model of complacency by Bahner and Manzey (Bahner & Manzey, 2004; Manzey & Bahner, 2005), Bahner (2008) presented an integrated model of complacency and automation bias. Both the model of complacency and the integration of complacency and automation bias are shown in the Appendix B.

Bahner's (2008) model was further developed by Parasuraman & Manzey (2010). The Parasuraman & Manzey (2010) model is depicted in Figure 1.3. It points to the central role of attentional factors contributing to complacency and some forms of automation bias. Two aspects of complacency and automation bias are differentiated, complacency potential and attentional bias in information process-

ing. Complacency potential is the tendency to over-rely on an automation. It is assumed to be affected by characteristics of the individual (technology-related attitudes, self-efficacy, personality traits) and the system (reliability, consistency, degree of automation), and by the experience the operator gained with the specific automation. By itself, complacency potential is not necessarily manifested in an attentional bias in information processing. Aspects of the task context (concurrent tasks, workload) and operator state contribute to a reallocation of attention and selective information processing. This less attentive information processing results in a loss of situation awareness. During normal automation operation there are no performance consequences. However, when the automation fails, it can lead to omission and commission errors. Complacency and automation bias are conceived to be dynamic and adaptive, and develop in accordance with the experience the operator made with the specific automation. This is modeled through two feedback loops. The positive feedback loop promotes an increase in complacency potential over time, as in highly reliable systems negative performance consequences will only rarely be experienced. However, if automation failure and resulting performance consequences are experienced, complacency potential decreases, triggered by a negative feedback loop. The drastic effect of failures has been shown in research on trust in automation (e.g., Moray et al., 2000; Lee, 1991; Lee & Moray, 1992).

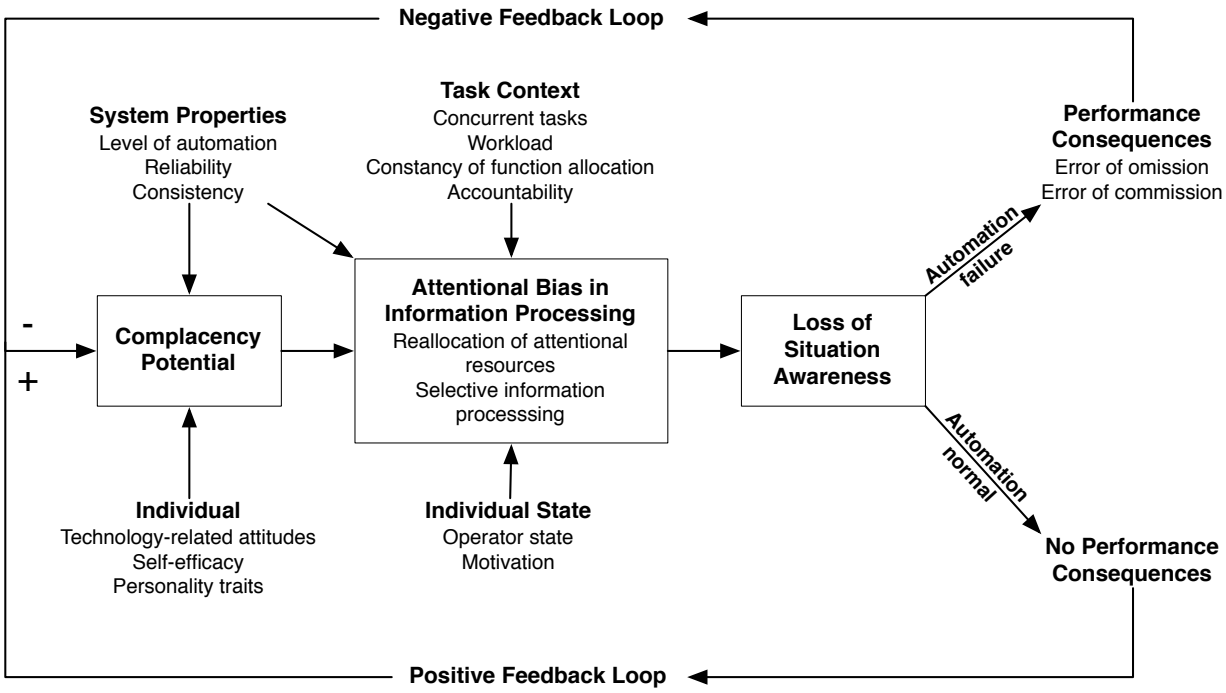

Figure 1.3: Integrated Model of Complacency and Automation Bias, adapted from Parasuraman & Manzey (2010, p. 404)

1.4 Loss of Situation Awareness & Loss of Manual Skills

Situation Awareness

Endsley (1988, 1995) defines situation awareness (SA) as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." (Endsley, 1995, p. 36) Level 1 SA comprises perception of relevant information and is the basis for level 2 and level 3 SA. Depending on goals and objectives, attention is directed to different cues. Level 2 SA is about comprehending the current situation; perceived information is interpreted and integrated to understand the situation. With knowledge of the critical elements and understanding of the situation, level 3 SA can be achieved, the elements' future status can be predicted.

Automation can influence situation awareness through changes in vigilance and complacency, a shift from active to passive information processing and system control, and altered feedback provided to the operator (Endsley, 1996). Lack of situation awareness can cause out-of-the-loop performance problems. Carmody & Gluckman (1993) found detrimental effects on level 2 SA (understanding the situation) when tasks were automated. Also Endsley & Kiris (1995) found negative effects on level 2 SA under full automation but no effect on level 1 SA. They attribute this effect to a shift from active to passive information processing.

Loss of Manual Skills

Automating manual or cognitive functions can lead to degradation or loss of those skills in the human operator. However, the human operator has to be able to step in whenever the automation fails, so skills have to be preserved. Endsley & Kiris (1995) found that loss of skills is smaller when the operator is kept in the loop, when automation is lower. On the other hand, Lorenz et al. (2002) showed that even with highly automated support, skills can be preserved when the operator mentally follows the automated functions. Loss of skill might be difficult to show in laboratory experimentation because it usually covers only short time spans. Bahner, Hüper et al. (2008) found no decrease of fault identification performance after working with an automated aid that provided diagnoses for system faults for some hours. Sauer, Hockey, & Wastell (2000) reported that in a test session eight months after an extensive training on the task no performance decrement was found. To counteract loss of skill, automated functions can be executed manually on a regular basis, like pilots do in simulator flights, or, in the context of adaptive automation, functions are allocated to the human or machine depending on situational needs.

1.5 Automation and Operator Functional State

Sleepiness is often involved in incidents and accidents (Dinges, 1995). Increased risk of incidents and accidents caused by sleepiness have been reported in car driving (Horne, & Reyner, 1999), railway traffic (Härmä, Sallinen, Ranta, Mutanen, & Müller, 2002), aviation (Samel, Wegmann, & Vejvoda, 1995), medicine (Rogers, Hwang, Scott, Aiken, & Dinges, 2004), and industry (Philip, & Akerstedt, 2006).

Increased risk of fatal occupational accidents is associated with difficulties in sleeping and non-daytime work (Akerstedt, Fredlund, Gillberg, & Jansson, 2002). A number of infamous accidents at highly automated workplaces happened during night shifts, such as Three Mile Island, Exxon Valdez, Bhopal, and the Estonia ferry accident (Folkard et al., 2005).

Despite its relevance for automation design, so far only few studies focused on the influence of the operator functional state on performance consequences of automation. The operator functional state can be defined as "the variable capacity of the operator for effective task performance in response to task and environmental demands, and under the constraints imposed by cognitive and physiological processes that control and energise behaviour" (Hockey, 2003a, p.3).

Under conditions of stress, fatigue or high workload, one would expect performance to decrease. However, if decrements are found, they are small and affect less important tasks (Hockey, 1997; Hockey, 2003b). Navon and Gopher (1979) talk about "graceful degradation" in this context.

In the compensatory control model, Hockey (1997, 2003b) postulates an adaptive regulatory process that aims at keeping up performance in high priority tasks, but at the expense of detrimental effects on lower priority tasks. Primary task performance may not give rise to concerns that the operator is at his limits because, as part of a compensatory process, resources may be reallocated to protect the primary task performance. Thus, no performance problems are apparent. However, decrements show in less important secondary tasks, effort and workload are increased, and psychophysiological activation is elevated. Control strategy shifts are made towards less demanding cognitive operations, reduction in redundancy, reduction of working memory load, and increased selectivity. The operator may

only be able to manage routine tasks, and performance can break down in case of unpredictable events. In the long run, it may have negative effects on the operator's health (Hockey, 1997; Hockey, 2003b).

Adaptive mechanisms are used in response to stress, noise, sleep deprivation and shift work (Hockey, 2003a). Sauer and colleagues studied the influence of noise as a stressor and the effect of different levels of adaptive and adaptable automation. (Sauer et al., 2011; Sauer, Kao, & Wastell, 2012; Sauer et al., 2013). Participants worked with AutoCAMS (for a detailed description see the following chapter) under different levels of noise. They found that noise exposure caused performance decrements and increased workload. Automation support reduced the negative effects of noise. Participants preferred higher DOA support under noise (Sauer et al., 2011; 2013). Adaptable automation showed advantages over adaptive automation (Sauer et al., 2012; 2013).

Effects of sleep deprivation on fault management in a supervisory control task were studied by Hockey and colleagues (Hockey et al., 1998; Sauer et al., 2003). Participants in the first study (Hockey et al., 1998) worked with CAMS after normal sleep and after one night of sleep loss. In the human-centered condition, they had access to the system raw data at all times. In the machine-centered condition, they could only access system parameters in case of system malfunctions. No effect on primary task performance under sleep deprivation was found, but slower reaction times in the secondary task (albeit only in the machine-centered condition), increased effort and a shift in system management strategies towards less monitoring. Sauer et al. (2003) also had participants work with CAMS in a day and a night shift. Again, they found no decrease in primary task performance after sleep deprivation but slower reaction times in the secondary task. Also, participants re-

ported increased perceived mental load and use of simplified strategies after sleep deprivation. The results of those studies are explained within the framework of the compensatory control model proposed by Hockey (1997). Under stress and high workload, attentional resources are allocated to what is perceived as the primary task in order to protect primary task performance. This can result in decreased secondary task performance, use of less demanding problem solving strategies, and increased effort and fatigue (Hockey, 1997; Hockey, 2003b).

1.6 Current Research

As part of this research, three studies were conducted examining the impact of degree of automation, system experience and operator functional state on human performance in interaction with automated aids.

In the first study, we examine the influence of degree of automation on human-automation interaction. Three automation supported groups and a manually working group are compared. We expect advantages of automation support compared to manual performance during routine performance. Differences between different degrees of automation are studied, with reliable automation support as well as when automation fails and operators have to return to manual performance. Complacency and automation bias are examined during reliable automation support and in case of automation failure. Following Moray's criticism (Moray & Inagaki, 2000; Moray, 2003), we not only measure detection rate of automation failure in our studies but also monitoring behavior, with optimal sampling predefined.

The second study focuses on the performance consequences of system experience. Effect of failure experience and duration of failure-free automation support

on trust and automation bias are studied. In addition, we analyze possible causes of commission errors in more detail.

The third study addresses the impact of operator functional state and DOA on human-automation interaction. Performance during the day is compared with performance during the night after prolonged wakefulness, simulating a first night shift. We look at how automation can help protect performance in sleep-deprived operators and how automation use affects complacency and automation bias and return to manual performance in case of automation failure.

In all three studies, we use a simulated supervisory control task, AutoCAMS. The same experimental task was used in the studies by Bahner and colleagues (Bahner, 2008; Manzey et al., 2006; Bahner, Hüper et al., 2008; Bahner, Elepfandt et al., 2008). However, some changes were made to get more detailed information about verification sampling behavior. AutoCAMS as well as the improvements made for the current studies are described in detail in the following chapter.

Chapter 2

AutoCAMS 2.0

A revised version of AutoCAMS (Hockey et al., 1998; Lorenz et al., 2002) was used for the experiments (AutoCAMS 2.0; Manzey, Bleil et al., 2008). This task was developed as a small-scale simulation of a typical supervisory control task of control room operators. It simulates an autonomously running life-support system consisting of five subsystems that are critical to maintain atmospheric conditions in a remote space capsule: oxygen, nitrogen (needed to maintain stable cabin pressure), carbon dioxide, temperature, and humidity. During nominal operation, all parameters are automatically kept within target range. However, due to malfunctions in the system, parameters can go out of range. A total of nine malfunctions can occur in either the oxygen or the nitrogen subsystem, including a blockage of a valve, a leak of a valve, a stuck-open valve, a defective sensor, or a defective mixer valve. The system faults are described in detail in Section 2.5. The user interface of AutoCAMS 2.0 is shown in Figure 2.1.

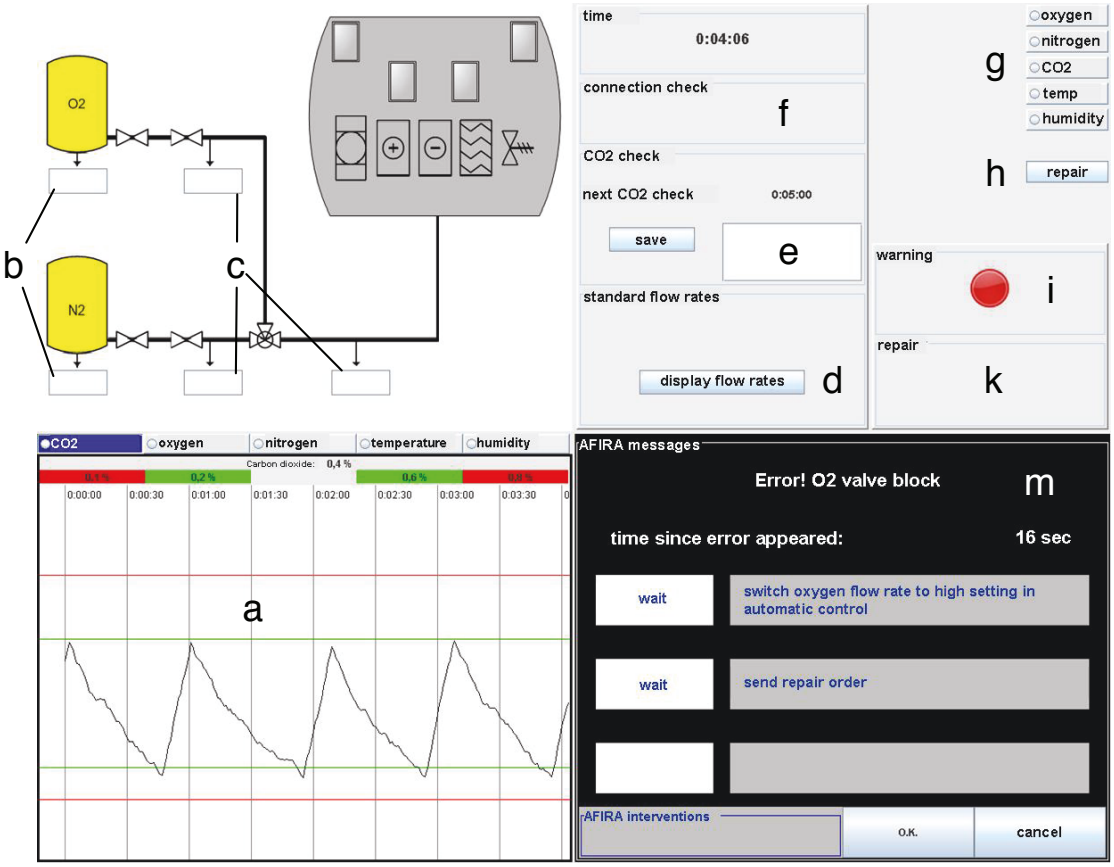

Figure 2.1: AutoCAMS 2.0 User Interface with activated Action Implementation Support. The figure shows the following elements of the AutoCAMS 2.0 User Interface: (a) monitors / history graphs for each subsystem, (b) O_2 and N_2 tank level readings, (c) O_2, N_2 and mixer valve flow readings, (d) standard flow rates, (e) input field for CO_2 readings (secondary task), (f) connection check icon (secondary task), (g) menu for manual control of each parameter, (h) menu for repair orders, (i) master alarm, (k) repair order information, (m) automated aid (Automated Fault Identification and Recovery Agent, AFIRA).

2.1 Operator Primary Task

The primary task of the operator involves supervisory control of the subsystems, including diagnosis and management of system faults. Whenever a fault is detected in the system, a master alarm turns on (Figure 2.1 i), and a time counter starts displaying how much time has elapsed since the fault occurred.

To have the malfunction fixed, its specific cause has to be identified, and an appropriate repair order has to be selected from a maintenance menu (Figure 2.1 h). The repair itself takes 60 seconds. During this time, the operator is required to control the affected subsystem manually. For this purpose, a manual control menu can be activated that allows for manual control of the different system parameters (Figure 2.1 g). If the repair order sent was correct, the warning signal turns green and all subsystems run autonomously again. In case of a wrong repair order, the warning light stays red and the operator is required to manually control the system by selecting appropriate actions from the control menu until a correct repair is initiated and completed.

2.2 Concurrent Tasks

In addition to the primary task, two concurrent secondary tasks have to be performed. The first one is a prospective memory task, which requires participants to check and record the carbon dioxide values every 60 seconds (Figure 2.1 e). The other secondary task is a simple probe reaction time task. This task is introduced to the participants as a check of a proper connection with the spacecraft. Participants have to click on a "communication link" icon (Figure 2.1 f) as fast as

possible. This icon appears in random intervals roughly twice per minute.

2.3 Automation Support

Depending on the specific version of AutoCAMS 2.0, participants have to perform fault diagnosis and management manually (manual control) or with the support of an automated aid (Automated Fault Identification and Recovery Agent AFIRA; Figure 2.2). There are three degrees of automation, Information Analysis Support, Action Selection Support, and Action Implementation Support. In case of *Information Analysis Support* (Figure 2.2b), the master alarm is accompanied by a message providing a specific diagnosis for the given system fault. However, action planning and implementation is left to the operator. In case of *Action Selection Support* (Figure 2.2c), the diagnosis is supplemented by a list of appropriate actions that the operator has to implement manually. In case of *Action Implementation Support* (Figure 2.2d), AFIRA does not only display a diagnosis and a list of necessary actions but also implements all steps autonomously if confirmed by the operator. In case of *manual control* (Figure 2.2a), the AFIRA message field merely shows the time that has elapsed since the error occurred.

2.4 Access to Raw Data

To be able to identify faults in the manual condition or verify proposed diagnoses in conditions with automation support, operators have independent access to all important parameters (see Figure 2.1 a, b, c, d). These include relevant system parameters and a history graph for each of the five subsystems. However, this

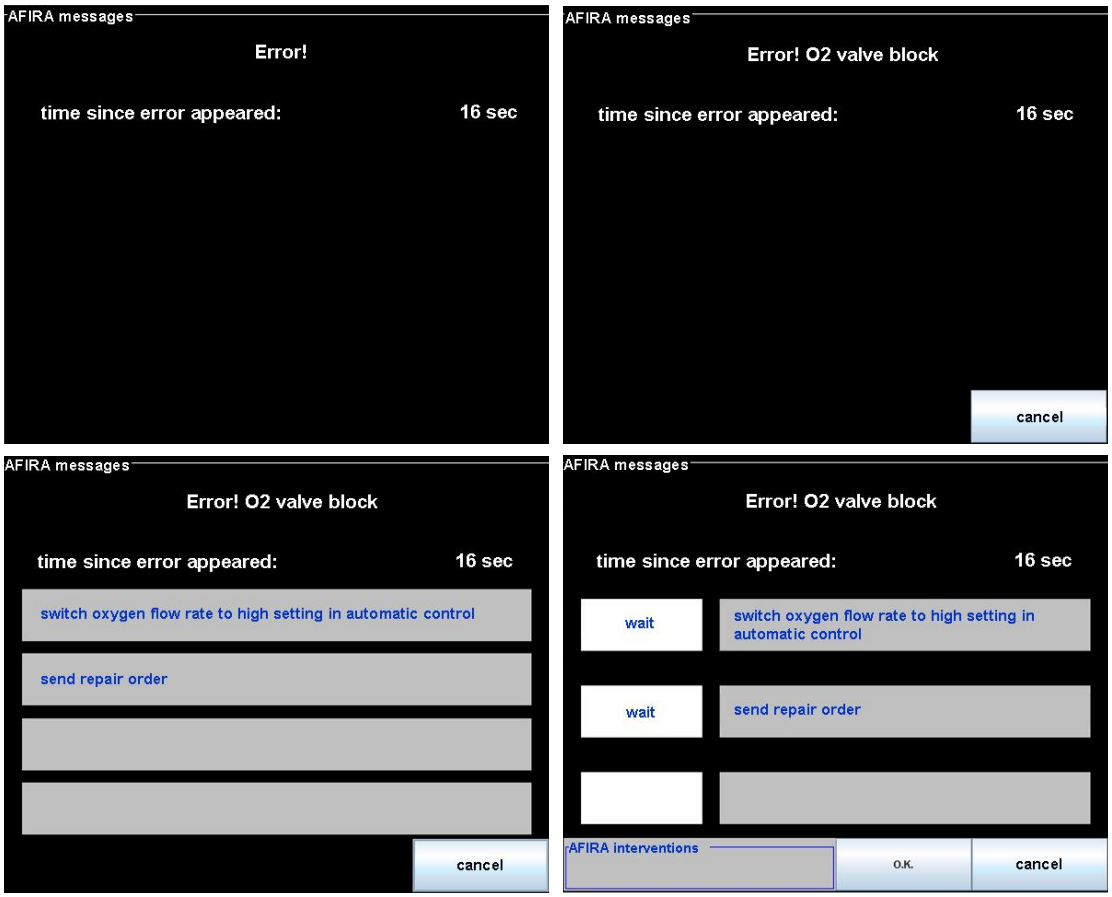

Figure 2.2: Automated Fault Identification and Recovery Agent AFIRA. Error messages for a) Manual Diagnosis, b) Information Analysis Support, c) Action Selection Support, and d) Action Implementation Support

information is not always visible but has to be activated for a 10-seconds view by a mouse click on the tank, flow meter, or history graph, respectively. Every system malfunction has specific symptoms such that it is possible for the operator to identify the malfunction or to verify the diagnosis provided by AFIRA by accessing two to four specific parameters, depending on the complexity of the fault. Identifying most complex faults unambiguously additionally requires interventions in the system. For each malfunction, there is one or more other system faults that has some symptoms in common, so the operator has to check all necessary parameters to be able to disambiguate the diagnosis. For details, see the description of system faults below.

2.5 System Faults, Diagnosis, and Recovery

The following system faults can occur in either the nitrogen or oxygen system: blockage of a valve, leak of a valve, stuck-open valve, defective sensor. In addition, there can be a defective mixer valve. For each malfunction, there is one or more other system faults that have some symptoms in common, so the operator has to check all necessary parameters to be able to disambiguate the diagnosis. In addition to the necessary parameters, there are relevant parameters that are useful to check but not necessary for unambiguous diagnosis. Without checking necessary information, it is not possible to diagnose a malfunction unambiguously. Necessary and relevant parameters were predefined for each malfunction according to a normative model, and accessed parameters were compared to the normative model. Participants were trained to check all necessary and relevant parameters. Table 2.1 lists all necessary and relevant parameters for each system malfunction,

the distinctive symptoms, and similar malfunctions that have to be ruled out.

Table 2.1: AutoCAMS 2.0. Necessary and Relevant Parameters to Identify System Malfunctions

System Malfunction	Necessary Parameters	Relevant Parameters	Symptoms	Similar Malfunction
O_2 Valve Leak		O_2 Monitor	O_2 level is falling below the lower boundary	O_2 Valve Blockage
	O_2 Tank			
	O_2 Valve Flow		tank level (t_2) - tank level (t_1) > O_2 flow	
O_2 Valve Blockage		O_2 Monitor	O_2 level is falling below the lower boundary	O_2 Valve Leak
		Pressure Monitor		Mixer Valve Blockage
	O_2 Tank			
	O_2 Valve Flow		O_2 flow is reduced compared to standard flow rates	
	N_2 Valve Flow			
	Standard Flow Rates			
N_2 Valve Leak		Pressure Monitor	pressure is falling below the lower boundary	N_2 Valve Blockage
	N_2 Tank			
	N_2 Valve Flow		tank level (t_2) - tank level (t_1) > N_2 flow	
N_2 Valve Blockage		Pressure Monitor	pressure is falling below the lower boundary	N_2 Valve Leak
		O_2 Monitor		Mixer Valve Blockage
	N_2 Tank			
	N_2 Valve Flow		N_2 flow is reduced compared to standard flow rates	
	O_2 Valve Flow			
	Standard Flow Rates			
Mixer Valve Blockage		O_2 Monitor	O_2 level is falling below the lower boundary	O_2 Valve Blockage
		Pressure Monitor	pressure level is falling below the lower boundary	N_2 Valve Block
	O_2 Valve Flow		O_2 flow is reduced compared to standard flow rates	
	N_2 Valve Flow		N_2 flow is reduced compared to standard flow rates	
	Standard Flow Rates			
Defective O_2 Sensor (oxygen falling)	O_2 Monitor		O_2 level is falling below the lower boundary	
	O_2 Valve Flow		O_2 flow = 0	
Defective O_2 Sensor (oxygen rising)		O_2 Monitor	O_2 level is rising above the upper boundary	O_2 Valve Stuck Open
	Turn O_2 Flow Off Manually			
	O_2 Monitor		O_2 level is falling back to normal	
	Turn O_2 Flow Control Back To Automatic			
		O_2 Monitor	O_2 level keeps falling	
	O_2 Valve Flow		O_2 flow = 0	
O_2 Valve Stuck Open		O_2 Monitor	O_2 level is rising above the upper boundary	Defective O_2 Sensor
	Turn O_2 Flow Off Manually			
	O_2 Monitor		O_2 level is falling back to normal	
	Turn O_2 Flow Control Back To Automatic			
		O_2 Monitor	O_2 level is rising	
	O_2 Valve Flow		O_2 flow > 0	
Defective Pressure Sensor (pressure rising)	Pressure Monitor		pressure is rising above the upper boundary	N_2 Valve Stuck Open
	O_2 Monitor		O_2 decreases	
N_2 Valve Stuck Open	Pressure Monitor		pressure oscillates just below the upper boundary	Defective Pressure Sensor
	O_2 Monitor		O_2 decreases	

2.5.1 Decreasing Oxygen Level: Leak of an Oxygen Valve versus Blockage of an Oxygen Valve versus Mixer Valve Blockage

Leak of an O_2 valve, blockage of an O_2 valve, and a defective mixer valve are system faults that lead to a reduced flow of oxygen. The decreasing oxygen level can be seen in the oxygen monitor.

In case of an O_2 valve leak, the amount of oxygen that flows from the tank is bigger than the amount that flows through the valve. In case of an O_2 valve blockage, the amount of oxygen that flows from the tank is reduced, thus also the amount that flows through the valve is reduced compared to standard flow rates. However, in contrast to an O_2 valve leak, the amount of oxygen that flows from the tank is the exact same amount that flows through the valve. In case of a blocked mixer valve, both oxygen and nitrogen flows are reduced compared to standard flow rates. The amounts of oxygen and nitrogen that flow from the tanks are the exact same amount that flow through the O_2 valve and N_2 valve, respectively. The decision tree is shown in Figure 2.3.

Fault management for a leak or blockage of the oxygen valve requires the operator to set oxygen flow from "standard" to "high". This way enough oxygens flows into the cabin even though a part of the oxygen is lost. In case of a defective mixer valve, both oxygen and nitrogen flow are to be set from "standard" to "high" so enough oxygen and nitrogen can flow into the cabin. After the repair is completed (60 seconds after sending the correct repair order), the flow is automatically set back to "standard".

2.5.2 Decreasing Pressure: Leak of a Nitrogen Valve versus Blockage of a Nitrogen Valve versus Defective Mixer Valve

Leak of an N_2 valve, blockage of an N_2 valve, and a defective mixer valve are system faults that lead to a reduced flow of nitrogen. The decreasing cabin pressure can be seen in the pressure/nitrogen monitor. Diagnosis and fault management are equivalent to these faults in the oxygen system, see detailed description in the preceding section. The decision tree is shown in Figure 2.4.

2.5.3 Increasing Oxygen Level: Stuck-open Oxygen Valve versus Defective Oxygen Sensor

A stuck-open O_2 valve and a defective O_2 sensor are system faults that lead to an increased oxygen level. The increasing oxygen level can be seen in the oxygen monitor.

In case of an stuck-open O_2 valve, the valve cannot be closed which leads to a continuous flow of oxygen into the cabin and a resulting high oxygen concentration in the cabin. Also cabin pressure increases to the upper limit, a drain valve prevents pressure from going above the upper limit. The same symptoms are seen in case of a defective O_2 sensor when the fault occurs while oxygen is rising. The oxygen level rises above the upper limit, the automation does not stop oxygen flow as the sensor fails to signal that the upper limit has been reached.

In order to disambiguate the two system faults, oxygen flow has to be stopped manually. After the oxygen level has reached target range again, control of oxygen

flow is given back to the automation. Next, the oxygen flow has to be checked: in case of an stuck-open O_2 valve, oxygen flows again. In case of a defective O_2 sensor, oxygen does not flow, the oxygen flow reading is zero. The decision tree is shown in Figure 2.5.

If the defective O_2 sensor error occurs while oxygen is falling, O_2 keeps falling below the lower limit. The O_2 flow reading is zero. No further checking is needed as in the case of a defective O_2 sensor while oxygen is rising. The complexity of the verification procedure is comparable to that of an oxygen leak.

Fault management of both system malfunctions requires the operator to manually control oxygen flow until the repair is completed.

2.5.4 Increasing Pressure: Stuck-open Nitrogen Valve versus Defective Pressure Sensor

A stuck-open N_2 valve and a defective pressure sensor are system faults that lead to increased cabin pressure. The increasing cabin pressure can be seen in the pressure / nitrogen monitor.

In case of a stuck-open N_2 valve, the valve cannot be closed which leads to a continuous flow of nitrogen into the cabin and a resulting high pressure in the cabin. Cabin pressure increases to the upper limit and oscillates just below the upper limit, a drain valve prevents pressure from going above the upper limit. Oxygen level decreases which can be seen in the oxygen monitor. Similar symptoms are seen in case of a defective pressure sensor. However, in case of a defective pressure sensor, pressure rises above the upper limit, the automation does not stop nitrogen flow as the sensor fails to signal that the upper limit has been reached. The drain

valve stays closed. Oxygen level decreases which can be seen in the oxygen monitor.

Fault management of both errors requires the operator to manually control nitrogen flow until the repair is completed.

2.5.5 Methodological Improvements

Previous studies by Bahner and colleagues (Bahner, 2008; Manzey et al., 2006; Bahner, Hüper et al., 2008; Bahner, Elepfandt et al., 2008) used an older version of AutoCAMS. In the current studies, the newer version AutoCAMS 2.0 was used which implemented some changes to get more detailed information about verification sampling behavior as described earlier in this chapter.

In Bahner's (2008) studies, only one button had to be pressed to display all tank level and flow information at the same time. This way, it could not be distinguished which of the 5 different pieces of information (tank level oxygen and nitrogen, gas flow through oxygen, nitrogen, and mixer valve) a participant was actually requesting and looking at. For the current studies we improved this methodology and implemented a separate access to each parameter by clicking directly on the requested parameter. This way, participants could access information only by clicking on the desired item, so we could see which information they actually sought. This approach allowed a more detailed analysis of information sampling behavior.

A further improvement was the distinction between relevant and necessary parameters. Relevant parameters include system information that helps identify a system fault, but diagnosis is also possible without this information. Necessary parameters are essential for diagnosis, unambiguous diagnosis is not possible without

checking all necessary parameters.

In addition, the system fault and proposed diagnosis in case of a false diagnosis were altered such that a false diagnosis was not obvious after sampling one piece of information. In Bahner's study (Bahner, Hüper, et al., 2008) the automated aid AFIRA suggested an oxygen blockage when the actual system fault was an oxygen valve stuck open. The same system (oxygen) was affected, but the effect on the oxygen concentration that is displayed in the oxygen monitor was opposite. In case of an oxygen blockage, the oxygen concentration falls while for an oxygen valve stuck open, the oxygen concentration rises. Thus, the participant only needed to check the oxygen monitor to detect that the aid provided a wrong diagnosis. (The oxygen monitor is often the first thing to check to get an overview of the situation before checking details of flow in the different valves. In fact, all the participants that committed a commission error in Bahner's study (5 out of 24) did check the oxygen monitor but followed the wrong diagnosis despite the contradicting information from the oxygen monitor.) That means that even a participant who is complacent in the sense that he does not check all the necessary information could easily detect that the aid provided a wrong diagnosis. After seeing that the trend in the oxygen monitor was opposite to what had to be expected according to the aid's suggestion, he could already conclude that the aid erred and only had to find the real underlying system fault. In addition, the errors were not parallel in the sense that the same system raw data had to be sampled to see that the automated aid provided a wrong diagnosis. For oxygen blockage, the oxygen monitor, actual gas flow and standard gas flow had to be checked. For oxygen valve stuck open, the oxygen monitor was the only parameter to be checked that both errors shared; the oxygen valve stuck open error would then have to be disambiguated against

a defective sensor by controlling the oxygen flow (turning off automatic control, turning automatic control back on).

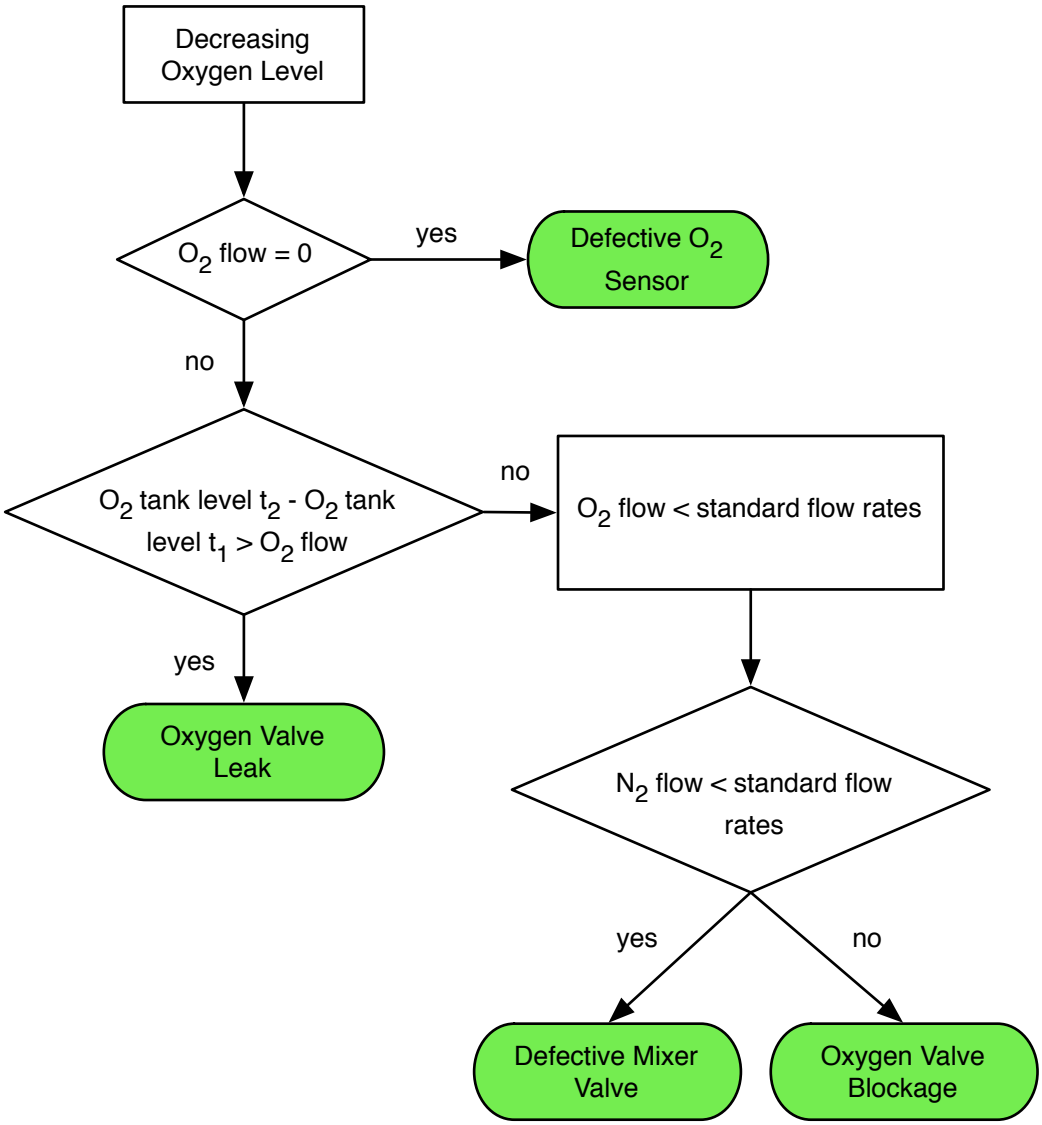

Figure 2.3: Decision Tree in Case of Decreasing Oxygen Level

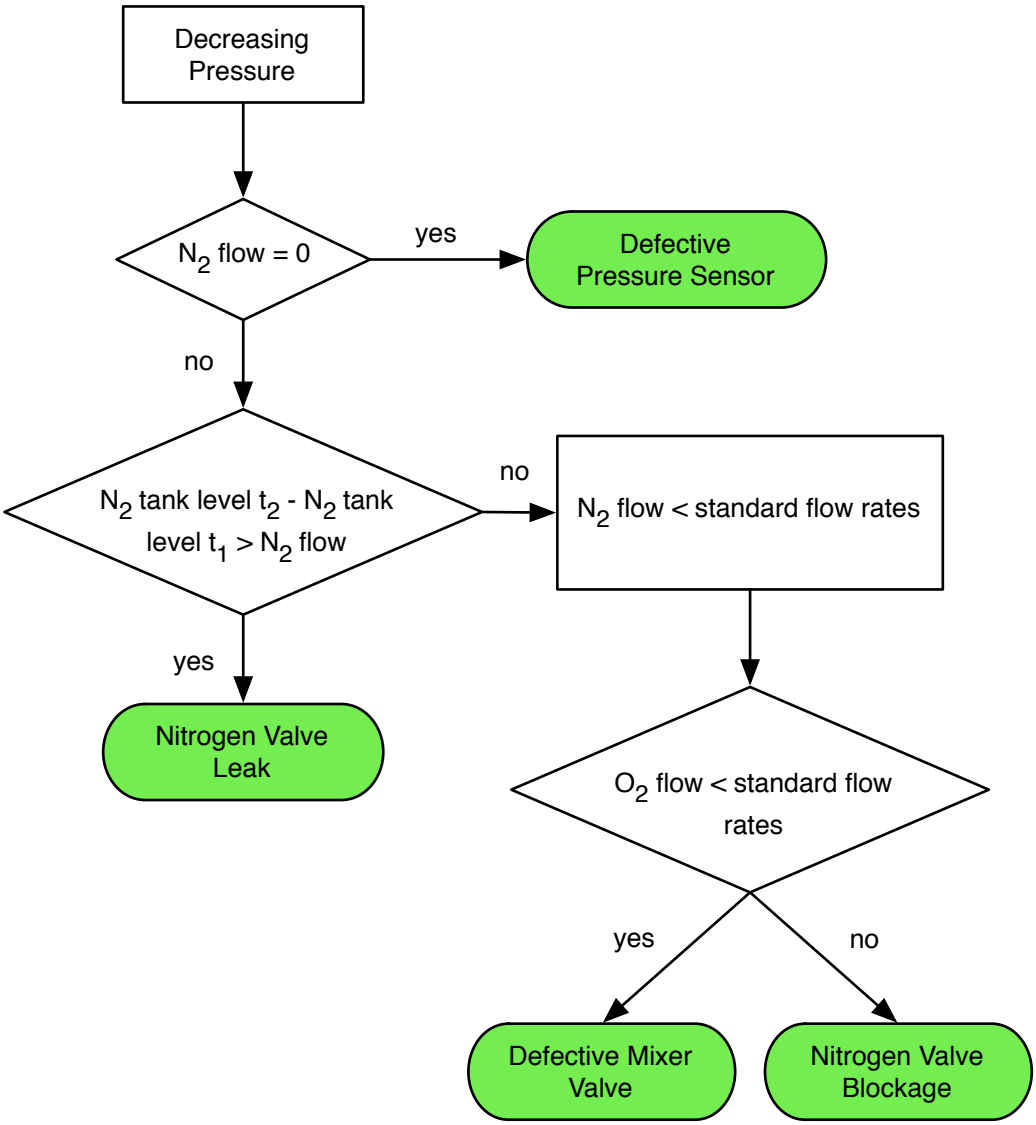

Figure 2.4: Decision Tree in Case of Decreasing Pressure

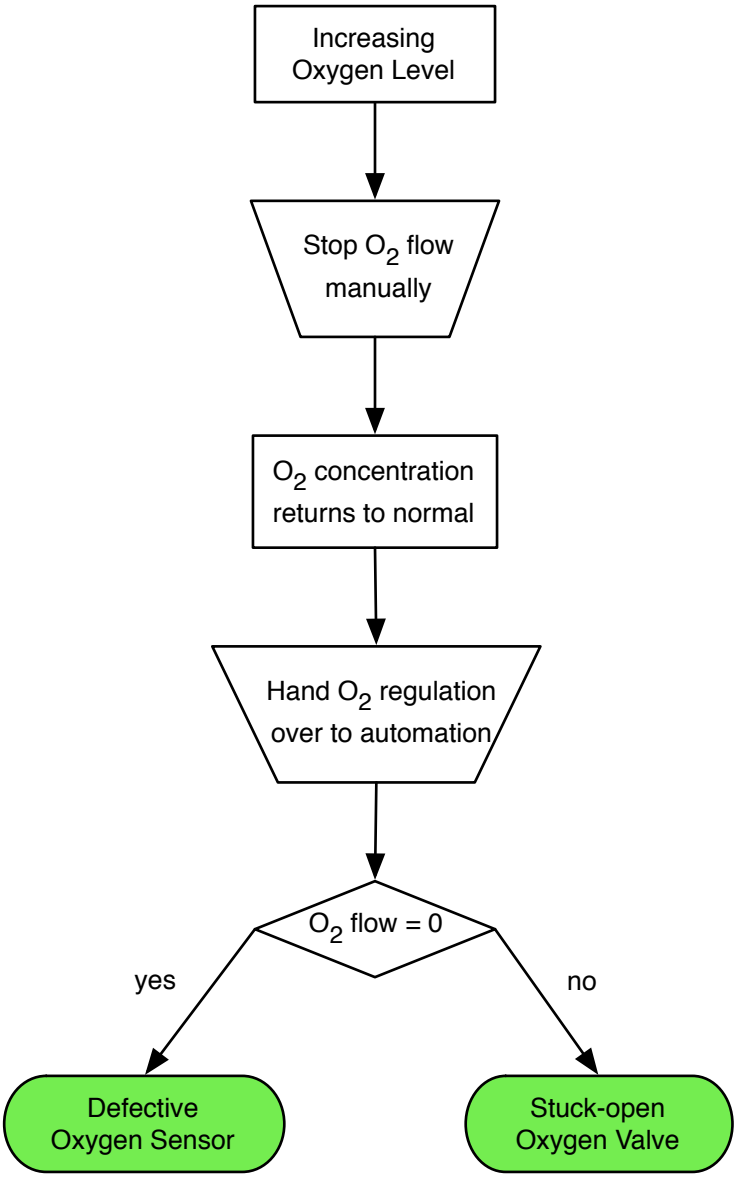

Figure 2.5: Decision Tree in Case of Increasing Oxygen Level

Chapter 3

Study I: The Impact of Degree of Automation

In the first study we explored the effects of degree of automation on human performance in interaction with an automated aid in a simulated process control task. We compared three automated aids that offered varying degrees of support for fault diagnosis and management (Information Analysis Support, Action Selection Support, and Action Implementation Support; for details see previous chapter on AutoCAMS). A manual condition in which participants diagnosed and managed system faults unaided served as a control condition. We looked into intended positive effects on routine performance and workload as well as failure performance. Negative performance effects studied included return to manual performance after automation breakdown as well as effects of complacency and automation bias.

During reliable automation support we expected performance benefits reflected in better primary and secondary task performance, and decreased workload compared to manual performance. With increasing degree of automation, performance

benefits were expected to increase and workload was expected to decrease. Previous research results are not clear about which DOA could best prevent performance losses during automation failure.

As higher DOA has been shown to support information sampling for automation verification (Lorenz et al., 2002) we expected more information sampling in higher DOA conditions, preventing commission errors in case of a false diagnosis provided by the automated aid.

Information sampling was expected to decrease with increasing complexity of the verification procedure. We used three levels of complexity with varying numbers of parameters that were necessary to verify a given diagnosis unambiguously.

3.1 Methodology

3.1.1 Participants

In the first experiment, 56 engineering students (40 male, 16 female) participated, ranging in age from 20 to 31 years ($M = 24.2$). None of them had prior experience with the simulated process control task used in the study. Participants were paid 70 Euro for completing the study.

3.1.2 Apparatus: AutoCAMS 2.0

AutoCAMS 2.0 was used for this experiment. For a detailed description see chapter 2 AutoCAMS 2.0.

3.1.3 Design

The study used a 4 (Degree of Automation, DOA) x 5 (Block) design with DOA (Manual Control, Information Analysis (IA) Support, Action Selection (AS) Support, Action Implementation (AI) Support) defined as between-subjects factor and block defined as within-subjects factor. The study design is illustrated in Figure 3.1. During the first block, all participants worked manually, without the assistance of AFIRA. During Blocks 2, 3, and 4, the three AFIRA groups were supported by AFIRA, whereas the manual control group continued working without automation support. In Block 5, participants of all experimental groups had to return to manual performance, i.e., diagnose and manage all system faults manually again without automation support.

In each block, six system faults occurred. Faults in all blocks were matched with respect to type and complexity. Thus, it was ensured that the fault identification and management procedures were equivalent for all blocks. All groups worked with the same set and distribution of faults (see Table 3.1). In the AFIRA groups, the six faults in Blocks 2, 3, and 4 were all correctly indicated and diagnosed by the automated aid. However, in Block 4, an additional seventh fault occurred for which AFIRA provided a wrong diagnosis (AFIRA proposed an Oxygen Valve Blockage when the actual system malfunction was an Oxygen Valve Leak). This failure of AFIRA was implemented to simulate a “first automation failure effect”.

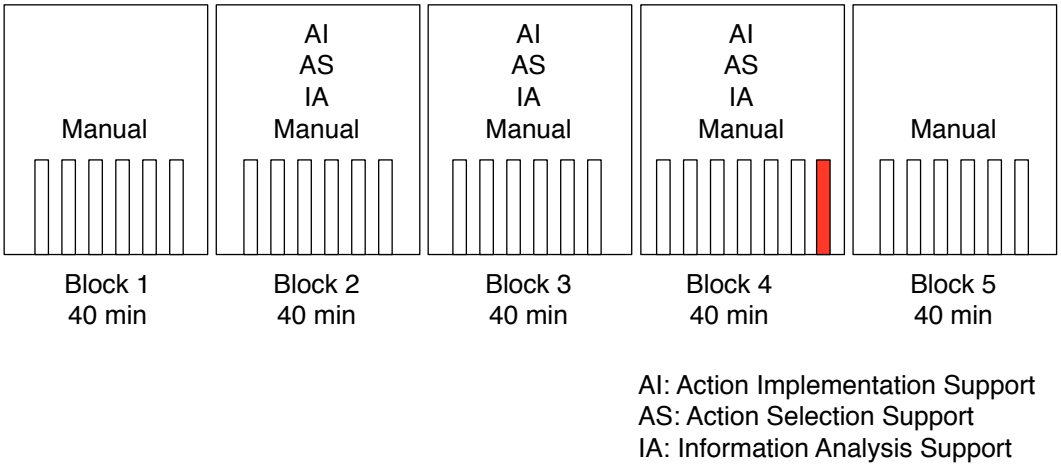

Figure 3.1: Study I. Experimental Design. The figure shows the distribution of system faults and automation failures across blocks, and the available automation support for each block. Each column represents one system fault. The red column represents the critical automation failure at the end of Block 4.

Table 3.1: Study I. Distribution and Timing of System Faults Across Blocks

Block	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6	Fault 7
Block 1	Valve Leak N_2 142s	Defective Sensor N_2 462s	Valve Blockage Mixer 841s	Defective Sensor O_2 (falling) 1253s	Valve Stuck Open O_2 1629s	Valve Blockage O_2 2014s	
Block 2	Defective Sensor O_2 (rising) 89s	Valve Blockage O_2 486s	Valve Stuck Open N_2 858s	Valve Blockage N_2 1238s	Valve Leak O_2 1624s	Valve Leak N_2 2052s	
Block 3	Valve Blockage N_2 124s	Defective Sensor N_2 479s	Valve Blockage O_2 863s	Valve Leak N_2 1224s	Valve Stuck Open O_2 1610s	Valve Leak O_2 2049s	
Block 4	Valve Leak N_2 102s	Valve Blockage O_2 414s	Valve Leak O_2 742s	Valve Stuck Open N_2 1087s	Valve Blockage N_2 1474s	Defective Sensor O_2 (rising) 1820s	Valve Leak O_2 2151s
Block 5	Valve Leak N_2 142s	Defective Sensor N_2 462s	Valve Blockage Mixer 841s	Defective Sensor O_2 (falling) 1253s	Valve Stuck Open O_2 1629s	Valve Blockage O_2 2014s	

Note. Fault 7 in Block 4 (O_2 Valve Leak) was falsely diagnosed by AFIRA as O_2 Valve Blockage.

3.1.4 Procedure

The study comprised two 4.5 hours sessions distributed across 2 days. The first session included a familiarization and practice session with the AutoCAMS 2.0 system. Participants were introduced to the different subsystems and trained to manually identify and manage all possible system faults that could occur either in the oxygen or nitrogen subsystem. The training was concluded by a questionnaire-based test which served to assess the participants' acquired knowledge about system fault identification and fault management procedures. All participants passed this proficiency test successfully and were accepted for the experiment. The training material can be found in Appendix C.

On the second day, participants were randomly assigned to one of four experimental groups. Participants of the automation supported groups were introduced to their automated aid and practiced using it for several trials. During this training, AFIRA always provided correct diagnoses. However, participants were informed that its reliability is high but not perfect and were warned explicitly to check the proposed diagnoses before initiating a repair. To keep the amount of training the same for all groups, participants of the manual control group performed the same practice trials with AutoCAMS 2.0 but without any automation support. The experiment started after this training session. Participants worked with AutoCAMS for five blocks of 40 minutes each. Subjective ratings of workload were collected during short breaks between blocks and after the last block.

3.1.5 Dependent Measures

Dependent measures were derived from questionnaires and from data that were logged during the experiment, including participants' mouse-clicks and AutoCAMS system dynamics. For all measures, Fault 7 in Block 4 (wrong diagnosis is proposed by decision aid) was analyzed separately.

Three **primary task performance** measures were calculated for each block:

(a) *Percentage of correct diagnoses* was the percentage of the six faults occurring per block for which the first repair order sent was correct, a measure of quality of fault identification performance.

(b) *Fault identification time* was defined as time (in seconds) from appearance of the master alarm until the correct repair order was issued. This measure was used to assess speed of fault identification performance.

(c) *Out-of-target error* was defined as the time (in seconds) the most critical system parameter (oxygen) was out of target range when a system fault was present, a measure of quality of fault management performance.

Secondary task performance was assessed by two measures:

(a) *Simple Reaction Time* defined as mean response time (in milliseconds) to the appearance of the "communication link" icon ("connection check" task) and

(b) *Prospective Memory Performance* defined as proportion of entries of carbon dioxide records that were provided within the correct time interval (at every full minute with a tolerance of 5 seconds).

Only performance during periods when a participant had to deal with a system fault was considered for secondary task performance.

Subjective workload was assessed by the *NASA Task Load Index* (NASA-

TLX; Hart & Staveland, 1988) and was defined as the mean of the ratings provided for the six subscales.

Measures used to assess the effort invested in **automation verification** included

(a) *Automation Verification Time* (AVT), defined as the time interval (in seconds) from the appearance of the master alarm until sending a first repair order, regardless of whether this repair order was correct or wrong.

(b) *Automation Verification Information Sampling of Relevant System Parameters* (AVIS-R), defined as the proportion of all system parameters accessed that were considered useful (relevant) to verify the automatically generated diagnosis for a given malfunction.

(c)*Automation Verification Information Sampling of Necessary System Parameters* (AVIS-N), defined as the proportion of all system parameters accessed that were necessary to verify a given diagnosis unambiguously. Note that necessary parameters are a subset of relevant parameters.

Necessary and relevant parameters were determined by means of a task analysis that was conducted to define a normative model of *eutactic* operator information sampling (Moray & Inagaki, 2000). For relevant and necessary parameters for each system fault, see Table 2.1. The number of necessary parameters that were needed to verify a given diagnosis unambiguously varied as a function of the complexity of a given system fault and included two parameters (low complexity), three to four parameters (medium complexity), or two parameters combined with two additional active interventions in the system (high complexity). The number of parameters actually accessed and interventions correctly performed when needed was then related to this normative model. Only parameters accessed between the occurrence

of the master alarm and sending the first repair order were considered for this measure. This approach to operationally define the level of complacency has first been described and used by Bahner and colleagues (Manzey et al., 2006; Bahner et al., 2008).

Automation bias was analyzed by the proportion of participants committing a *commission error*, defined as percentage of participants who followed the diagnosis of the automated decision aid for Fault 7 in Block 4 although it was wrong. As a control measure, it was assessed how many participants of the manual control group provided a wrong diagnosis for this fault.

Return-to-manual performance was assessed for the automation supported groups by comparing performance in Block 1 and Block 5, based on primary and secondary task performance measures as defined above.

3.2 Results

3.2.1 Primary Task Performance

Analysis of primary task performance measures was based on a 4 (DOA) x 5 (Block) ANOVA.

For percentage of correct diagnoses, a significant Block effect was found, $F(4, 208) = 25.09$, $p < .01$, moderated by a significant DOA x Block interaction, $F(12, 208) = 1.95$, $p < .03$. Already in Block 1, general level of performance was comparatively high for all experimental groups (Manual Control: 87%; IA Support: 86%; AS Support: 81%; AI Support: 86% correct diagnoses). As expected, providing automated support in Blocks 2 - 4 improved performance to about 100% correct

diagnoses in all automation supported groups, whereas the manual control group showed only slight improvement across blocks (M = 91%).

For fault identification time, the DOA effect, $F(3, 52) = 3.45$, $p < .03$, the Block effect, $F(4, 208) = 48.25$, $p < .01$, and the DOA x Block interaction, $F(12, 208) = 2.68$, $p < .01$, were significant. This is illustrated in Figure 3.2. As becomes evident, fault identification times profited considerably from providing automated aids in Blocks 2 - 4 compared with manual performance in Blocks 1 and 5, as well as compared with the performance of participants in the manual control group.

In addition, mean fault identification time in blocks with automation support (Blocks 2-4) differed between the automation supported groups. Fault identification time was shorter for AI Support ($M = 20.9$ s) than for the two groups with lower automation support (IA Support: $M = 28.3$ s; AS Support: $M = 28.5$ s). This effect was confirmed by a separate 3 (DOA) x 3 (Block) ANOVA comparing fault identification times for the three automation supported groups in Blocks 2 - 4, which showed a significant main effect of DOA, $F(2, 39) = 4.98$, $p < .02$, and post hoc contrasts (Bonferroni), both contrasts $p < .05$.

Similar results were found for the quality of fault management performance as reflected in out-of-target error. Effects of out-of-target error are displayed in Figure 3.3. One participant of the Information Analysis group and three participants of the Action Selection group were excluded from the analysis. They had the oxygen level drop down to zero or close to zero after one system fault. Once the oxygen level is extremely low, it takes some minutes until it reaches the normal range again, so we excluded those extreme outliers from the analysis.

For out-of-target error, the DOA effect, $F(3, 48) = 3.60$, $p < .03$, the Block effect, $F(4, 192) = 62.27$, $p < .01$, and the DOA x Block interaction, $F(12, 192)$

= 2.48, p < .01 were significant. For all experimental groups, fault management performance was better in automation supported blocks (Blocks 2 - 4) than in the manual Blocks 1 and 5. Improvements developed over time but in different ways for the different groups. The smallest improvements are observed in the Manual group. The most highly automated aid (AI Support) yielded the highest improvement, right from the first automation supported block (Block 2). The two groups with less automated aids showed less improvement.

Contrasting the three automation-supported groups in Blocks 2 - 4, performance in the AI group (M = 103.10 s) was better than in the IA (M = 142.51 s) and AS group (M = 137.59 s). This was also indicated by a separate 3 (DOA) x 3 (Block) ANOVA contrasting the performance in the three automation-supported groups in Blocks 2 - 4 which yielded a main effect of DOA, $F(2, 37) = 4.37$, p < .03. Post hoc contrasts (Bonferroni) showed that the AI Support group was significantly better than the IA Support group, p < .03, and tended to be better than the AS Support group, p < .07.

3.2.2 Secondary Task Performance

Performance in both secondary tasks was analyzed by a 4 (DOA) x 5 (Block) ANOVA.

No significant effects emerged for simple reaction times in the connection check task.

A significant Block effect, $F(4, 208) = 21.17$, p < .01, and a DOA x Block interaction, $F(12, 208) = 2.94$, p < .01, were found for prospective memory performance. As becomes evident from Figure 3.4, prospective memory performance

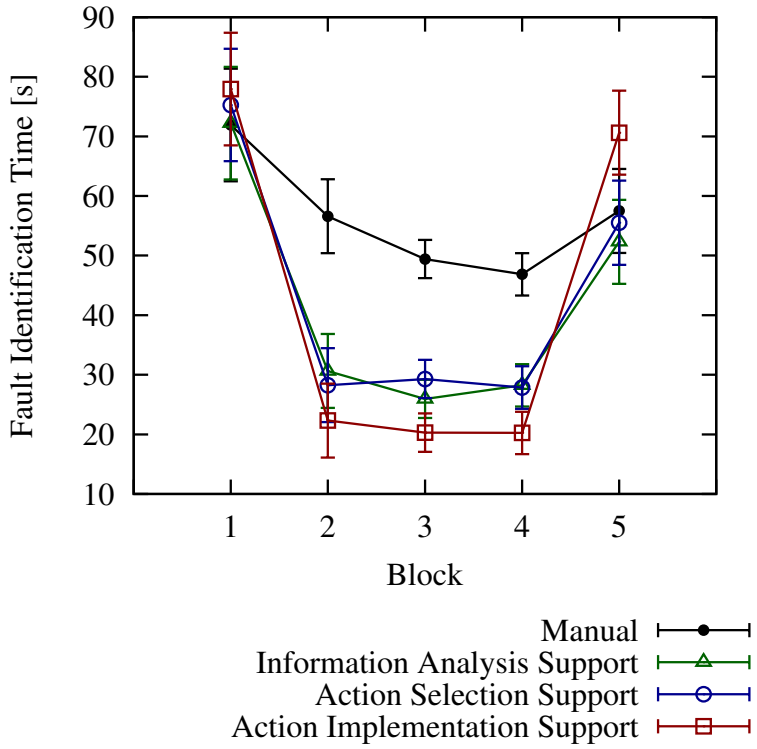

Figure 3.2: * Study I. Primary Task Performance: Fault Identification Time

improved immediately with the introduction of automation support (Blocks 2 - 4) for participants supported by the most highly automated aid (AI Support). Performance improvements were also found for IA and AS Support, but they developed slowly across the automation supported blocks. No performance changes across blocks were found for the Manual group.

3.2.3 Subjective Workload

Analysis of subjective workload was based on a 4 (DOA) x 5 (Block) ANOVA.

A significant Block effect, $F(4, 208) = 24.99$, $p < .01$, was found, moderated by a significant DOA x Block interaction, $F(12, 208) = 2.33$, $p < .01$. This

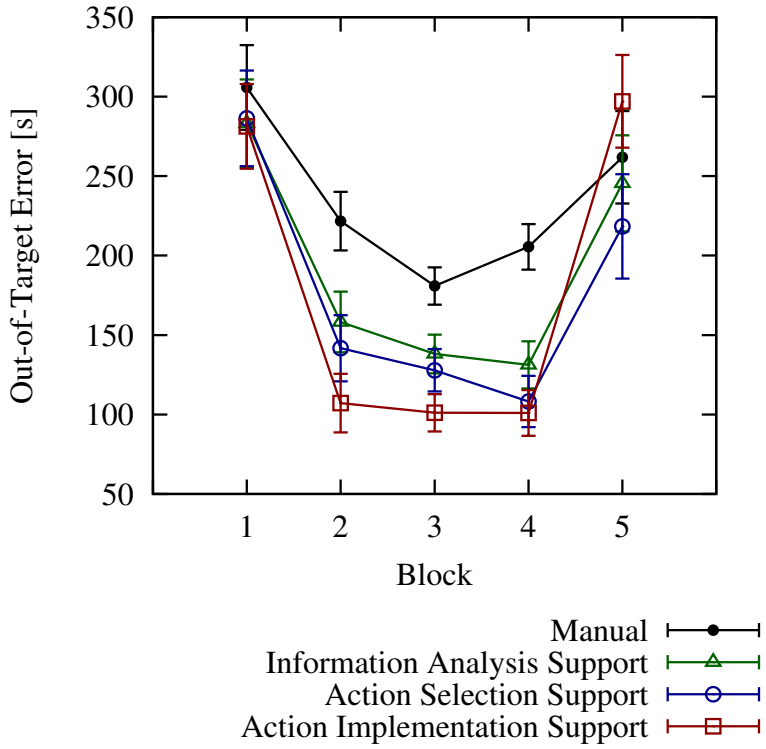

Figure 3.3: * Study I. Primary Task Performance: Out-Of-Target Error

effect is illustrated in Figure 3.5. All groups started at about the same level in Block 1. While in Blocks 2 - 4, workload decreased for all groups, it decreased the most for the AI group. In Block 5, which demanded manual control again, subjective workload increased for the automation supported groups, with the most pronounced increase in the AI group.

3.2.4 Return-to-Manual Performance

Assessment of return-to-manual performance for automation supported groups was based on a contrast of performance in Blocks 1 and 5 by a 3 (DOA) x 2 (Block) ANOVA.

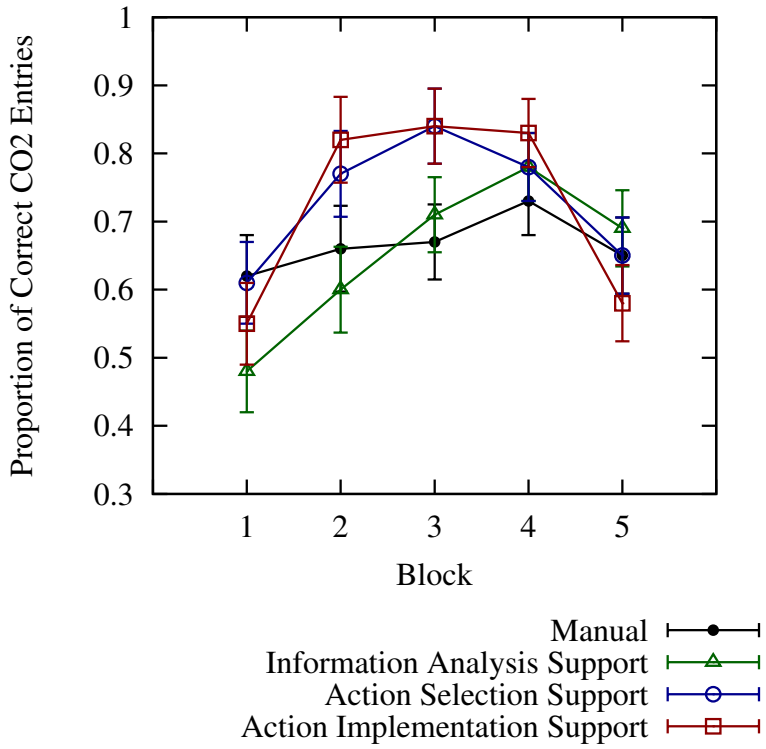

Figure 3.4: * Study I. Secondary Task Performance: Prospective Memory Performance

Whereas no significant effects were found for percentage of correct diagnoses, fault identification time improved across blocks, $F(1, 39) = 6.24$, $p < .02$, probably reflecting effects of practice.

With respect to secondary task performance, no significant effects emerged for simple reaction time. Prospective memory performance showed a significant improvement across blocks only (Block 1: 54.5% correct entries; Block 5: 64.1%), $F(1, 39) = 7.23$, $p < .01$.

Some indications of DOA effects on return-to-manual performance emerged for the out-of-target error, reflecting fault management performance. Whereas

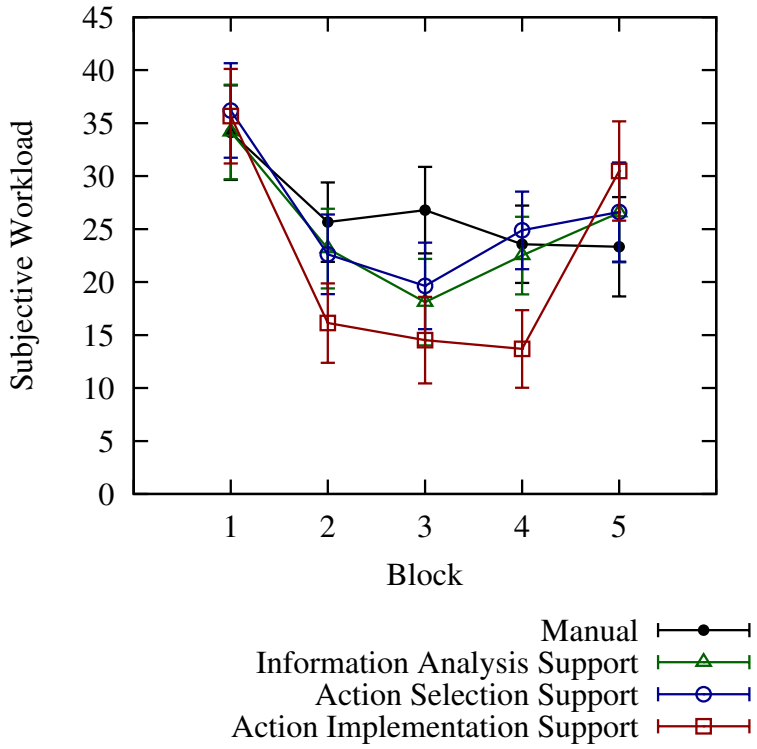

Figure 3.5: * Study I. Subjective Workload

manual performance of the participants in the IA group and the AS group improved considerably from Block 1 to Block 5, a slight performance decrement was observed for the AI group. This effect was evaluated by aggregating the data of the IA and AS groups that did not have any automation support for fault management implementation, contrasting it with the AI group. A 2 (DOA: IA & AS vs. AI) x 2 (Block) ANOVA revealed a significant DOA x Block effect, $F(1, 38) = 4.46$, $p < .05$. This effect is illustrated in Figure 3.6.

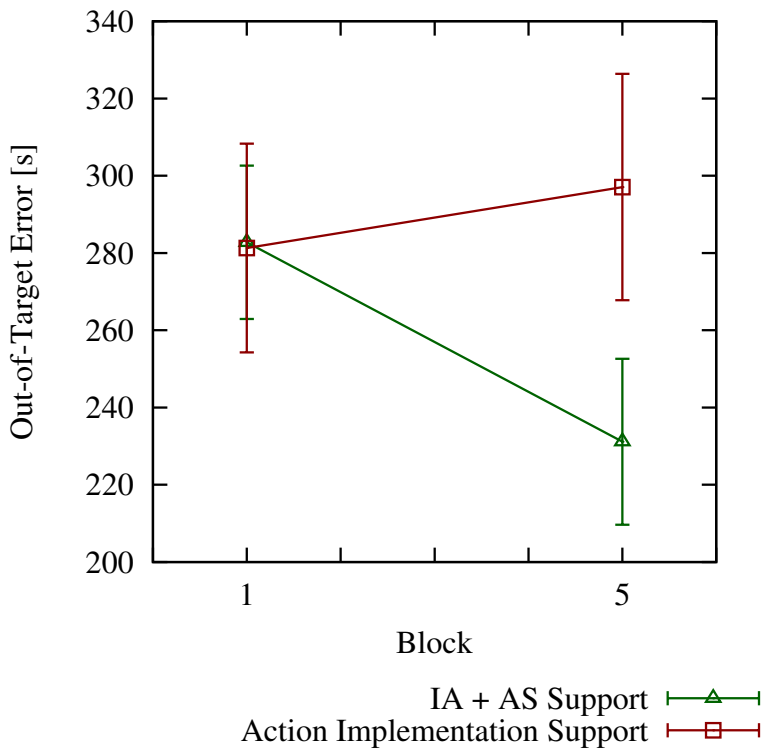

Figure 3.6: * Study I. Return to Manual Performance: Out-of-Target Error

3.2.5 Automation Verification During Reliable Automation Support

Automation verification behavior was analyzed by a 3 (DOA) x 3 (Block) ANOVA for the automation supported groups in Blocks 2, 3, and 4.

Analysis of the time spent to verify the recommendation of the aid (AVT) revealed a significant effect of DOA, $F(2, 39) = 4.32$, $p < .02$. Post hoc analysis revealed that the group supported by the highest DOA aid (AI Support) invested significantly less time to verify the automatically provided diagnosis than did the group working with the least automated aid (IA support), $p < .03$. Neither the

Block effect nor the interaction were significant. This is surprising as the aid's support in diagnosing the system fault is the same in all three DOA groups. In each case the automation provides a diagnosis which the operator is trained to validate before accepting it. Only the support for error management differs between the DOA groups. Thus, we further analyzed the verification behavior.

The total number of diagnostic clicks decreased significantly across blocks, $F(2, 78) = 6.55$, $p < 0.01$, but there was no difference between DOA, $F(2, 39) = 1.17$. In contrast, regarding regulatory control clicks needed for error management and for disambiguating two high complexity errors, the DOA effect was significant, $F(2, 39) = 10.53$, $p<0.01$. The highest DOA aid (AI Support) performed less regulatory clicks than the other two DOA groups (IA Support and AS Support).

For parameters that were *relevant* for verifying a diagnosis (AVIS-R), we found a significant Block effect, $F(2, 78) = 9.43$, $p < .01$. The DOA effect was not significant. Sampling of *relevant* parameters decreased over time, independent of DOA. However, this seemed to be an optimization of information sampling since relevant parameters included parameters that were useful but not necessary for diagnosis. When only regarding sampling of those parameters that were *necessary* for verifying the automatically generated diagnoses of the aid (AVIS-N), it showed that information sampling stayed at a constantly high but not quite perfect level ($M = 93.5\%$) across blocks, independent of DOA. Neither the DOA effect nor the Block effect nor the DOA x Block interaction were significant. These results are depicted in Figure 3.7.

The effect of complexity of fault diagnosis on automation verification was analyzed by a 3 (DOA) x 3 (Block) x 3 (Complexity) ANOVA for AVIS-N. Two parameters were necessary for verification of low complexity errors, four parame-

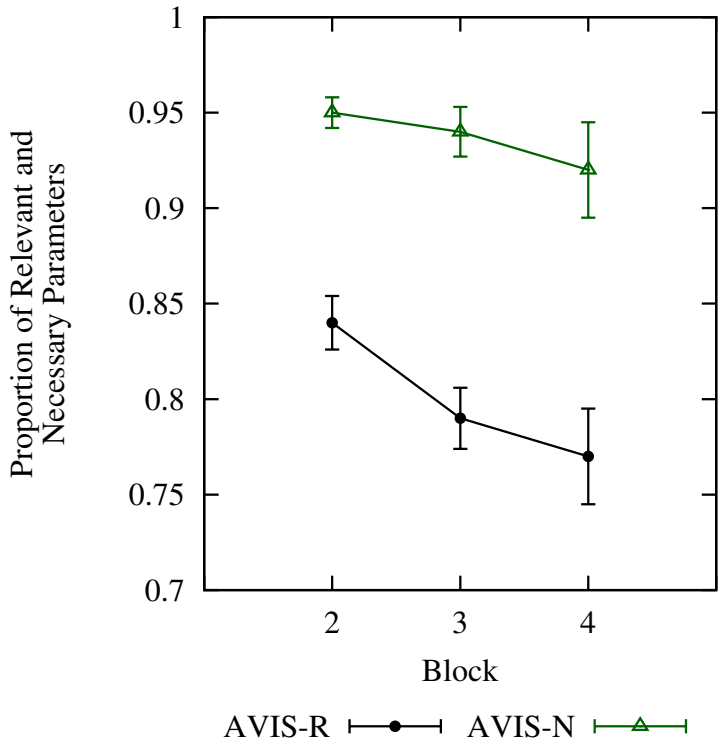

Figure 3.7: Study I. Automation Verification During Reliable Automation Support: Relevant and Necessary Parameters

ters were necessary for medium complexity errors. Verification of high complexity errors demanded accessing two parameters and in addition, the participant had to implement two control actions to be able to disambiguate two possible diagnoses (defective oxygen sensor versus stuck open oxygen valve).

For verification of necessary parameters, we found a significant main effect of Complexity, $F(2, 78) = 7.50$, $p < .01$, and a significant Complexity x DOA interaction, $F(4, 78) = 3.92$, $p < .01$. This effect is illustrated in Figure 3.8. While verification was almost complete for low complexity errors ($M = 97.6\%$), it decreased for medium complexity errors ($M = 93.1\%$) and high complexity errors

(M = 91.1%). Most interestingly, the extent of verification for low and medium complexity errors did not differ between the three DOA groups. However, for high complexity errors, which also demanded implementing control actions as part of the verification procedure, the Action Implementation group completed significantly less verification steps than did the Information Analysis or Action Selection group. One-way ANOVAs contrasting the automation verification behavior for the complexity levels separately revealed a significant effect of DOA only for high complexity errors, $F(2, 39) = 3.74$, $p < .05$. There were no significant effects for low and medium complexity errors. When looking at which part of the verification procedure is omitted, it is noticeable that only participants who were supported by the most highly automated aid (AI Support) omitted control actions that were needed to disambiguate two possible diagnoses. However, they did access the two necessary system parameters.

3.2.6 Automation Bias and Automation Verification in Case of Automation Failure

Clear evidence for automation bias leading to a commission error was found in all automation supported groups by analyzing fault identification performance for Fault 7 in Block 4 when the automation provided a false diagnosis. Up to half of the participants in the automation supported groups followed the automatically generated diagnosis for this fault even though it was wrong. However, no significant difference was found for the different degrees of automation support (IA: 42.9%; AS: 50%; AI: 35.7%), $F < 1.0$. In contrast, 13 out of 14 participants in the manual control group (92.9%) working on the same fault identified this fault correctly and

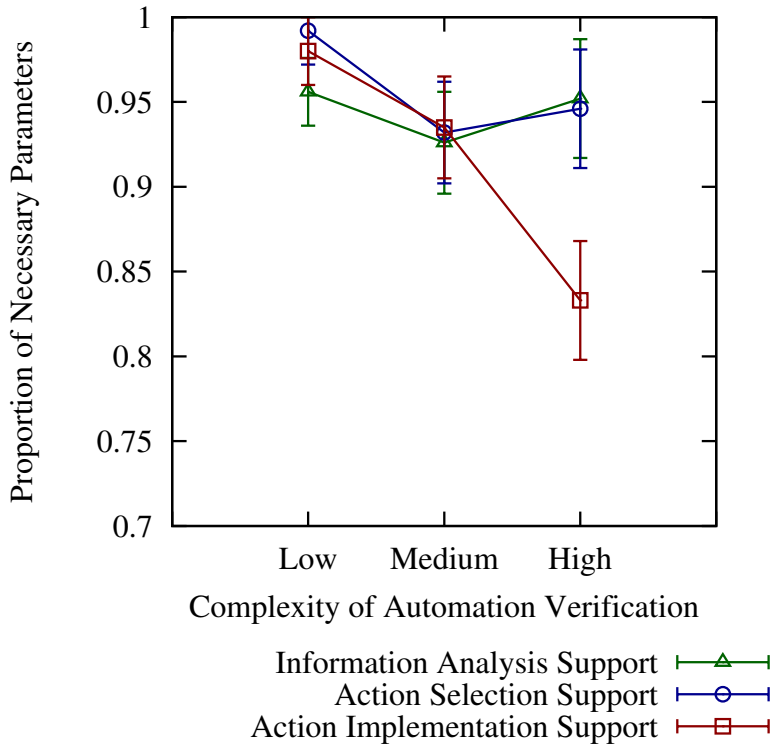

Figure 3.8: * Study I. Automation Verification During Reliable Automation Support: Effect of Complexity

sent a correct repair order.

In order to investigate whether the observed automation bias was due to a lack of automation verification or to a discounting of contradictory information from other available sources, we contrasted the information sampling behavior of participants who committed an error of commission with that of participants who did not. A 3 (DOA) x 2 (Commission Error: aid's false diagnosis detected vs. not detected) ANOVA revealed no significant effects for sampling of necessary information (AVIS-N).

For the automation failure, the proposed diagnosis was an oxygen valve block-

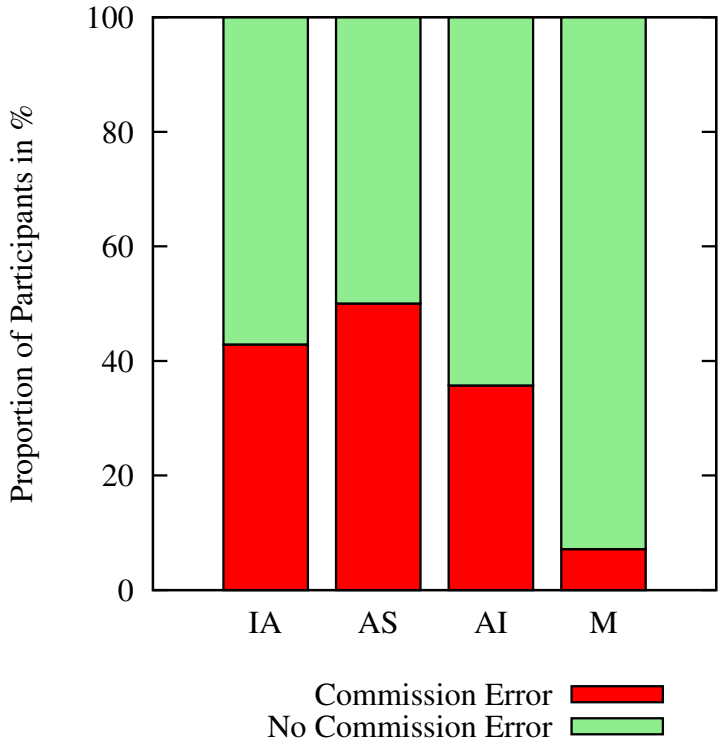

Figure 3.9: Study I. Automation Bias

age while the true system fault was an oxygen leak. To verify the aid's diagnosis, it was necessary to cross-check four parameters (oxygen tank level, oxygen flow, nitrogen flow, standard flow rates). Two of those parameters contained contradictory information (oxygen tank level and oxygen flow) that falsified the aid and should have led the operator to disagree with the aid's proposed diagnosis, and that enabled the operator to diagnose the true system malfunction. If those two pieces of information were not attended to, it was not possible to find the aid's diagnosis to be wrong and identify the true underlying system fault.

A detailed analysis of information sampling within the group of participants who committed a commission error revealed that out of the 18 participants who

followed the wrong recommendation, 11 had checked all necessary information before making a decision, i.e, they accessed all system parameters needed to detect the contradiction between the automatically generated diagnosis and the actual system state. Seven participants were complacent in the sense of an incomplete verification before sending the repair order. Only two of those participants did not check the information that falsified the aid and thus were not able to detect the wrong diagnosis.

A significant difference emerged between participants who did and did not detect the automation failure when we additionally contrasted the time spent per accessed system parameter. This time was significantly shorter for participants committing a commission error, $F(1, 35) = 12.13$, $p < .01$. To see whether this difference was already present in the preceding blocks with reliable automation support, we contrasted the verification behavior of participants who committed a commission error with that of participants who did not. A 4 (Block: 2, 3, 4, False Diagnosis) x 2 (Commission Error: aid's wrong diagnosis detected vs. not detected) ANOVA revealed a significant Block effect, $F(3, 117) = 4.40$, $p < .01$, and a significant Block x Commission Error interaction, $F(3, 117) = 11.36$, $p < .01$, for time spent per parameter. Figure 3.10 shows this effect. With reliable automation, there was no difference between the two groups. However, when the automated decision aid provided a false diagnosis, participants who did not detect the false diagnosis spent the same amount of time per parameter as in normal operation trials whereas participants who detected the false diagnosis invested more time per parameter to inspect the system. Although both groups sampled the same number of parameters, they differed considerably with respect to the time spent dealing with the sampled information in case it contradicted the aid's

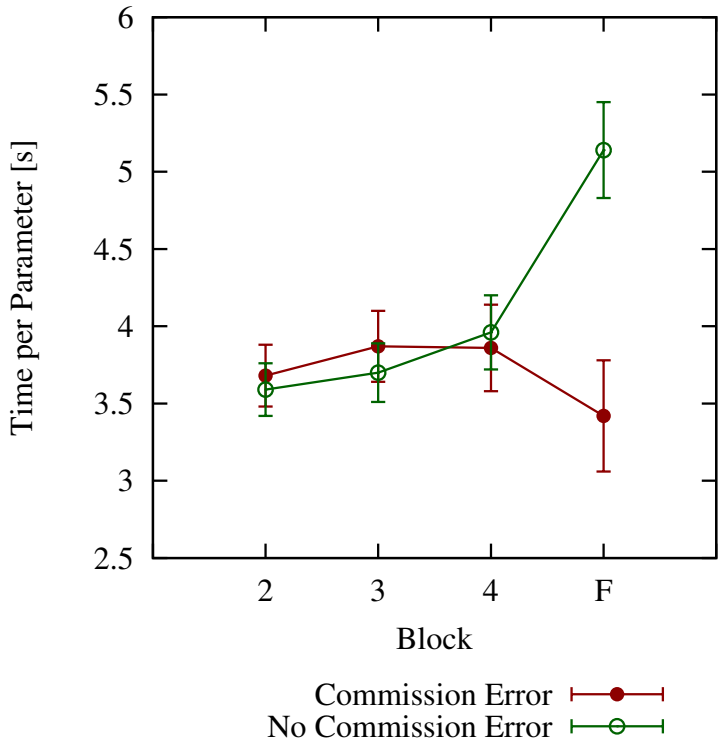

Figure 3.10: * Study I. Automation Bias: Time Spent Per Parameter for Blocks 2, 3, 4, and False Diagnosis

proposed diagnosis.

3.3 Discussion

3.3.1 Performance Benefits

Participants made use of the provided automation support for fault identification and management, and it clearly benefited routine performance. As expected, quality and speed of fault identification as well as fault management performance was better with the aid than without the aid. Improved primary task performance did

not come at the expense of secondary task performance, on the contrary, secondary task performance improved as well. Prospective memory performance was better in automation supported groups than in the manual group. Likewise, subjective workload decreased with the introduction of an automated aid. As expected, performance benefits depended on degree of automation.

3.3.2 Effects of Degree of Automation on Routine Performance

The most highly automated aid yielded the greatest benefits in most performance measures. Fault management performance was the best in the Action Implementation Support group, participants had shorter out-of-target times than the manual control group and the other two automation supported groups. This reflects the automated, optimal fault management that started as soon as the participants confirmed the proposed diagnosis. Participants in the AI Support group had short fault identification times so the optimal fault management was implemented only shortly after the onset of a system fault.

Fault identification times were the shortest for AI Support in the automation supported blocks. (Note that fault identification time and automation verification time are the same in case the automation support works reliably and the correctly proposed diagnosis is accepted.) This result is surprising since the support for diagnosing the system fault was the same for all automation supported groups. In all three groups the automation proposed a diagnosis that the participants were trained to verify before accepting it. Only the support for error management differed between the automation supported groups. However, the procedure for

sending an appropriate repair order differed slightly between the DOA groups. The AI Support group had to press a button when they agreed with the proposed diagnosis. This issued a repair order and started the fault management immediately. In contrast, the IA Support group as well as the AS Support group had to open the repair order catalogue, select the appropriate repair order and then press a button to send the repair order. The AI group had to do less steps to issue the repair order and could save some time with this easier confirmation procedure. Unfortunately, it was not possible to analyze the time saved because the log file data only contained information about when the repair order was sent (clicking the button "Send repair order" or "OK") but not when the repair menu was opened. However, this difference in the confirmation procedure cannot solely account for the eight second difference in fault identification time between the highest DOA group and the two lower DOA groups. So what else could have led to shorter automation verification times? Did the AI Support group sample less information to verify a given diagnosis?

Analysis of the automation verification behavior showed that the AI Support group carried out the same number of diagnostic clicks as the other groups, but they performed significantly fewer regulatory control actions. Regulatory control actions were needed for system stabilization. Participants of the IA and AS Support group probably started with regulatory control actions already before sending a repair order in order to stabilize the system. For the AI group this was not sensible to do because their aid stabilized the system automatically as soon as the repair order was issued. In addition, regulatory control actions were needed for the diagnosis of one system fault per block (defective oxygen sensor or oxygen valve stuck open). The AI Support group also omitted those time-consuming control

actions that were necessary to verify the proposed diagnosis. Note that the AI group is the only group that is supported with the fault management implementation and never has to implement control actions after sending a repair order. This seems to generalize to their verification behavior during the diagnostic phase before sending a repair order. This finding is in contrast with Lorenz et al. (2002). They found that operators in the higher automated condition sampled more information. However, in their study, the most highly automated aid presented a diagnosis and in addition a countdown; fault management would automatically start after the countdown if not vetoed by the operator. This setup did not afford rapid confirmation if the suggested diagnosis was found to be correct but offered more time to cross-check other system information. This was different in the present study. Fault management was not started until the proposed diagnosis was confirmed. So the earlier the proposed diagnosis was confirmed, the earlier fault management started.

Secondary task performance and subjective workload profited from providing an automated aid as well. Prospective memory task improvements emerged in all automation supported groups compared to manual performance. The group with the highest automated aid (AI Support) showed the most improvements, and improvements manifested right away when automation support was first provided. In contrast, the two lower automation groups (IA and AS Support) showed improvements only later in time. This probably reflects the higher memory load that those two groups had compared with the AI Support group. While in the AI Support group fault management was implemented automatically, the IA Support group had to remember the appropriate fault management procedure for each system fault, implement it manually and check if the system reacts as expected. The

AS Support group was provided with a list of appropriate actions but still had to remember how to implement them manually and check the system's reaction.

The highest automated aid was also associated with the highest decrease of subjective workload. On the downside, however, it also led to the sharpest increase of workload when returning to manual performance.

3.3.3 Effects of Degree of Automation on Failure Performance

Unwanted side effects of automation support often show when the automation fails and operators have to return to manual performance. Also in this study, we found some performance decrements when the automation failed and operators resumed manual performance.

Fault management performance decreased for the group with the highest DOA aid (AI Support) when comparing the first and the last block. The other two automation supported groups improved their fault management performance. Remember that the highest automated aid (AI Support) supported fault management, participants in this group never had to implement fault management manually. This lack of training shows in Block 5 when they have to go back to manual fault identification and implementation.

Fault identification was not negatively affected by use of automation support, in contrast, performance improved compared to the first manual block. Even with the diagnostic aid, participants were required to validate the proposed diagnosis, so they still had to access the system raw data and check if the automation's suggestion was correct. So even when they did not have to diagnose the underlying

system fault without assistance, they were still kept in the loop and practiced the diagnostic procedures. Similar results were found by Endsley and Kiris (1995) who investigated effects of level of automation on decision performance in an automobile navigation task. Decision accuracy remained at a very high level even after automation failure, but decision time increased immediately after automation breakdown. In the present study, fault identification was slower after automation failure than with reliable automation support but improved compared to manual performance in the first block. The continuing practice prevented the loss of skills needed for fault diagnosis. Even if diagnosing system faults was supported and the participants merely had to validate an automatically proposed diagnosis, this was enough to stay involved and keep up or even enhance system knowledge and diagnostic capabilities.

3.3.4 Automation Verification and Automation Bias

Automation verification time differed between automation groups. The group with the highest automation support (AI support) spent less time with verification. This results is consistent with results about the fault identification time. Note that fault identification time and automation verification time are the same in case the automation support works reliably and the correctly proposed diagnosis is accepted.

Information sampling decreased over time, independent of DOA. This supports earlier results by Bahner et al. (2008) who found a decrease in sampling of relevant information. However, Bahner et al. (2008) only considered relevant parameters, in constrast, in this study, we looked at both relevant and necessary

parameters. It is noteworthy that in this study this effect only shows when *relevant* parameters are considered, but it is not found for *necessary* parameters. Relevant parameters contain necessary parameters and in addition useful parameters. Useful parameters can make diagnosis easier, but diagnosis is possible without them. Only necessary parameters are absolutely essential for diagnosis, it is not possible to find a correct diagnosis or verify a proposed diagnosis without them. When considering the *necessary* parameters, which are only a subset of the relevant parameters, this effect of decreased sampling over time disappears. Instead, it shows that necessary parameters are sampled at a constantly high although not quite perfect level across all blocks. Reducing the sampling of optional, useful information while keeping the sampling of necessary information at a very high level can be seen as an optimization of the information sampling effort. Only those parameters are sampled that are absolutely essential and required for an unambiguous diagnosis. Additional information that can be useful but is not necessary for diagnosis is disregarded more and more over time. Participants learned to focus on the necessary parameters, thus they could reduce the effort needed to verify the aid's diagnosis. This can be seen as a strategy to reduce effort and workload while preserving the diagnostic performance.

The complexity of the verification process affected the verification of diagnostic advice. While verification was almost complete for low complexity errors, it decreased for medium complexity errors and high complexity errors. Most interestingly, verification for low and medium complexity errors did not differ between the three DOA groups. However, for high complexity errors, which required the implementation of control actions as part of the verification procedure, the AI Support group completed significantly less verification steps than the IA and AS

Support groups. When looking at which part of the verification is not completed, it is noticeable that only participants in the AI Support group omitted control actions. Remember, the AI Support group is the only group that is supported with the fault management implementation and never had to implement control actions after sending a repair order. This seemed to generalize to their verification behavior during the diagnostic phase before sending a repair order.

Despite sampling a high portion of necessary parameters, up to half of the participants made a commission error when the automated aid provided a false diagnosis, regardless of DOA. Surprisingly, even participants who checked all necessary information to verify the aid's diagnosis and thus were not complacent in the sense of incomplete information sampling, made a commission error and followed the aid's wrong advice.

Out of the 18 participants who committed a commission error and followed the aid's wrong advice, 11 participants did so despite a complete verification. They should have been able to detect the aid's failure. Complacency in the sense of incomplete verification was not the underlying cause of the commission error in these cases. Did those participants discount contradictory information, suggesting a decision bias, or did they not attentively process the accessed information, suggesting an attentional bias? Seven participants who followed the wrong recommendation had not completely verified the suggested diagnosis and would be classified as complacent in a sense of incomplete verification. The automatically proposed diagnosis required checking four parameters. Two of these parameters contained information that contradicted the proposed diagnosis. Out of the seven complacent participants, two had not checked the information that falsified the aid. Only those two participants were definitely not able to recognize that the aid

suggested a wrong diagnosis and find the true underlying system fault. The other five participants could have detected that the aid erred.

This resembles findings of Bahner and colleagues (Bahner, 2008; Bahner, Hüper, et al., 2008). In their study, participants did not sample all relevant information, but all the participants who made a commission error did sample the one piece of falsifying information. They could have detected the aid's false diagnosis but instead followed the wrong recommendation.

So in addition to incomplete verification there seems to be an active discounting or an inattentive processing of contradictory information that can lead to commission errors.

When comparing the sampling behavior of participants who committed a commission error and participants who did not, no difference was found for the number of parameters accessed. However, participants who did not commit a commission error spent more time per parameter they checked when the automation provided a false diagnosis. There was no difference in time spent per parameter in the preceding trials with reliable automation, so they did not spend more time examining parameters in general. This suggests that they realized that there was an inconsistency between the aid's advice and the system's raw data, hence they looked more closely and spent more time with the evaluation of the data. Similar results were found by Schriver, Morrow, Wickens, & Talleur (2008). In a simulated flight, expert pilots payed more attention to cues indicative of failures when a failure was present, and their decision accuracy and speed were better than that of novices. Schriver et al. (2008) suggest that differences in attentional strategies affect the decision outcome.

Participants who did not detect the false diagnosis, on the other hand, spent

the same amount of time per parameter they checked regardless of whether the diagnosis was correct or wrong, even if they sampled system information that contradicted the aid's proposed diagnosis. This points to an inattentive processing rather than discounting of contradictory information.

So in addition to an incomplete verification that can lead to commission errors, there seems to be a sort of "looking but not seeing effect" that can cause commission errors. Information that is looked at is not processed, analogous to "inattentional blindness", the "inability to perceive [...] caused by the fact that subjects were not attending to the stimulus but instead were attending to something else" (Mack & Rock, 1998, p.12). Even if fixated, information is not perceived if attention is on another object.

These findings are in line with results from Sarter, Mumaw, and Wickens (2007) who studied professional pilot's monitoring strategies. They found that even if pilots fixated flight mode annunciations (FMAs) they did not detect an inappropriate mode annunciation after an unexpected mode transition. Only one pilot detected the inappropriate mode annunciation; he fixated FMAs more often and longer.

To further explore this effect, a second study was conducted. In addition to underlying causes of commission errors we investigated the effect of system experience and failure experience on trust development and commission errors.

Chapter 4

Study II: The Impact of System Experience

Study I showed that commission errors cannot only result from incomplete automation verification but can also occur despite complete verification. Study II explores to what extent attentional processes play a role. As in study I, we analyzed the information that was accessed, and in addition, we analyzed to what extent participants were aware of the information they accessed. Right after the automation provided a wrong diagnosis and participants sent a repair order, the trial stopped and participants were asked to answer the Automation Verification Questionnaire. The questionnaire asked about which system parameters they accessed in order to validate a given diagnosis, and what the critical relations between the accessed parameters were. Based on this information, commission errors were attributed to three possible causes: incomplete automation verification, complete automation verification without awareness, or discounting of contradictory information.

Additionally, we explored the influence of system experience and especially

failure experience on trust development and risk of commission errors. Trust is calibrated according to the experience made with an automation (Lee & See, 2004; Merritt & Illgen, 2008; Seong & Bisantz, 2008), and automation failure has been shown to influence trust in automation (de Vries, Midden, & Bouwhuis, 2003; Dzindolet et al., 2003; Lee & Moray, 1992, 1994; Madhavan et al., 2006). Even a single automation failure can cause a decrease in trust, and trust recovers only slowly (Lee & Moray, 1992). In monitoring tasks, detection of early failure is better than detection of late failure, if additional tasks have to be performed concurrently (Molloy & Parasuraman, 1996). The experience of automation failures has also been shown to affect automation verification (Manzey et al., 2006; Bahner, Hüper et al., 2008; Bahner, Elepfand et al., 2008).

In the second study, we investigated how experience of reliable automation (positive experience) and automation failure (negative experience) affect trust development, automation verification, and the risk of commission errors. It was expected that trust would increase with increasing experience of reliable automation. Automation failure was expected to cause a decrease in trust, with following experience of reliable automation leading to a slow recovery in trust. Automation verification was expected to increase after automation failure experience, leading to a reduced risk of commission error when the automation would fail a second time. It is an open question if the effects of an early failure experience diminish after a long period of fault-free automation.

In Study II, a methodological improvement was made concerning the false diagnosis. The proposed diagnosis and the true underlying system fault in case of a false diagnosis were altered such that all necessary parameters had to be checked in order to detect the false diagnosis and identify the true system fault. Actual

system fault and proposed diagnosis required the same set of parameters to be checked. There was no subset of falsifying information that could have led the operator to detect the false diagnosis without complete automation verification.

4.1 Methodology

4.1.1 Participants

88 engineering students (65 male, 23 female; mean age 24.1 years) participated. Participants were paid 70 Euro for completing the study.

4.1.2 Apparatus: AutoCAMS 2.0

The same simulation of a supervisory process control task was used as in the first experiment. However, only the most highly automated decision aid (Action Implementation Support) was used for this study.

4.1.3 Design

The study involved four experimental groups that differed with respect to how long participants worked with the aid until an automation failure eventually occurred and whether this automation failure was the first or second failure the participants experienced. The study design is shown in Figure 4.1.

Participants of the first experimental group worked with the aid for one 35-minute block before a first automation failure occurred. During this time, AFIRA provided correct diagnoses for five system faults in a row before it eventually failed and provided a false diagnosis.

The second experimental group experienced a false diagnosis right in the beginning of the experimental trial. The remaining part of the trial was identical to Group 1 with five correct diagnosis and one false diagnosis at the end of the block. Thus, the automation failure at the end of the session was the second automation failure experience for Group 2.

A similar variation was realized for Experimental Groups 3 and 4. However, participants of these groups worked for a considerably longer period of time (four blocks of 35 minutes each, with a total of 20 system faults) with the system before the critical automation failure at the end of the session occurred. For Group 3 this was the first automation failure, for Group 4 it was the second automation failure after a first automation failure at the beginning of Block 1.

In order to keep the experience of system faults the same, Groups 1 and 3 were presented a correctly diagnosed malfunction when Groups 2 and 4 experienced a false diagnosis. In all groups, the diagnosis was Defective Oxygen Sensor. For Groups 1 and 3, this was the true system fault, for Groups 2 and 4, the true system fault was Oxygen Valve Stuck Open. The same verification procedure is required to disambiguate those two system faults. This first fault was not included in the further analysis. Also, the proposed diagnosis for the first fault (Defective Oxygen Sensor) never occurred again during the experiment.

Analyses of the relative impact of negative and positive experience on trust and automation verification behavior over time were based on Groups 3 and 4. The analysis of time-related and experience-related effects on automation bias involved all four groups.

Experienced reliability defined as number of correct diagnosis divided by total number of diagnosis is depicted in Figure 4.2. Distribution and timing of systems

fault are shown in Table 4.1.

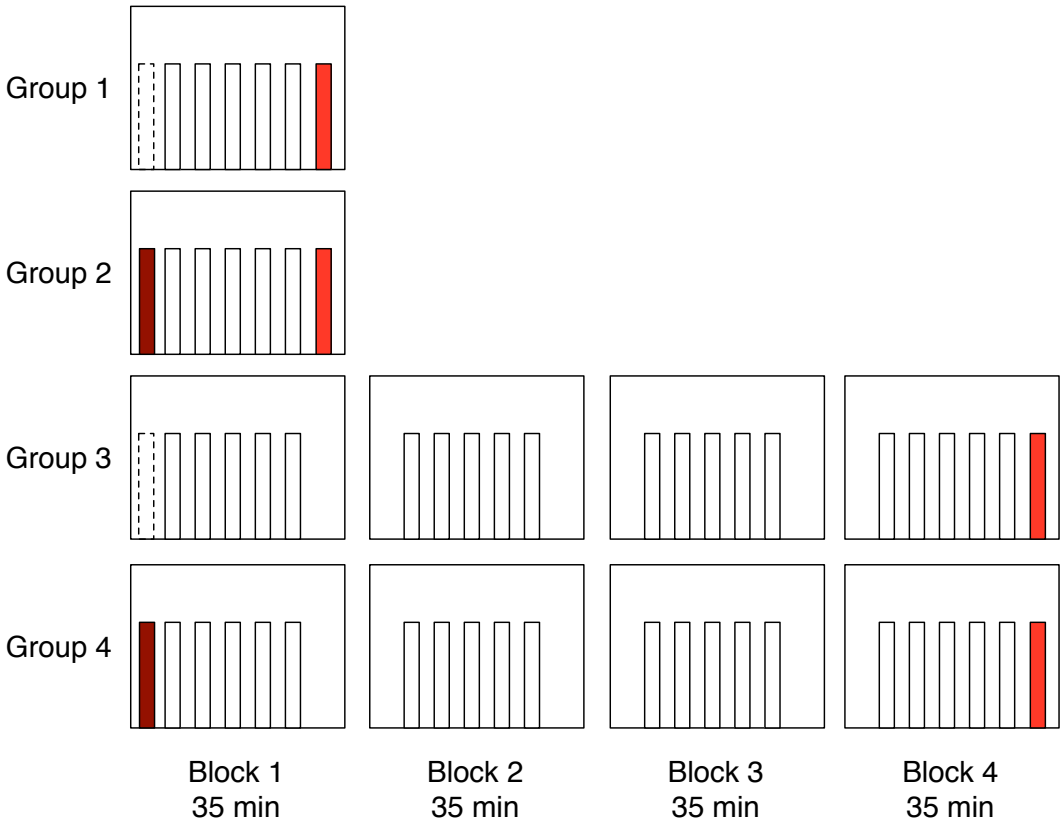

Figure 4.1: Study II. Experimental Design. The figure shows the distribution of system faults and automation failures across blocks for the four experimental groups. Each column represents one system fault. As part of the experimental treatment, the first system fault was either correctly diagnosed by AFIRA, represented by a dotted column, or AFIRA provided a wrong diagnosis, represented by a dark red column. The light red column represents the critical automation failure at the end of the session.

Table 4.1: Study II. Distribution and Timing of System Faults Across Blocks

Block	Fault X	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6
Block 1	Defective Sensor O_2 156s	Valve Blockage N_2 550s	Valve Leak O_2 800s	Valve Blockage N_2 / O_2 1071s	Valve Stuck Open O_2 1353s	Valve Leak N_2 1629s	Valve Blockage Mixer 1967s
Block 2		Valve Leak O_2 230s	Valve Blockage N_2 662s	Valve Stuck Open O_2 971s	Valve Blockage O_2 1373s	Valve Leak N_2 1829s	
Block 3		Valve Blockage N_2 252s	Valve Leak N_2 682s	Valve Blockage N_2 931s	Valve Leak O_2 1473s	Valve Stuck Open O_2 1888s	
Block 4		Valve Blockage N_2 190s	Valve Leak O_2 462s	Valve Blockage O_2 789s	Valve Stuck Open O_2 1153s	Valve Leak N_2 1409s	Valve Blockage Mixer 1707s

Note. For Groups 2 and 4, Fault X actually was an O_2 Valve Stuck Open but was falsely diagnosed by AFIRA as Defective O_2 Sensor, so all groups faced the same diagnosis. Fault 6 in Block 1 or 4, respectively, (Mixer Valve Blockage) was falsely diagnosed by AFIRA as O_2 Valve Blockage for all groups.
If Fault 6 occurred in Block 4 (Groups 3 and 4), Fault 3 in Block 1 was an N_2 valve blockage. If Fault 6 occurred in Block 1 (Groups 1 and 2), Fault 3 in Block 1 was an O_2 valve blockage. Thus, the 5 faults before the late automation failure (Fault 6 in Block 1 for Groups 1 and 2, Fault 6 in Block 4 for Groups 3 and 4) were identical for all groups.

4.1.4 Procedure

The experiment consisted of two familiarization and practice sessions and one experimental session distributed across three days. The practice session on the first day lasted approximately four hours and included familiarization and practice with the AutoCAMS system. Participants were trained to identify and manage all possible system faults manually, without automation support.

On the second day, all participants had to perform a 45-minutes test trial which served to test their acquired skills according to a predefined criterion. Only those participants who passed this test were accepted for the experiment.

The experimental session on the third day started with an introduction to AFIRA. This familiarization included a description of the aid's function as well as a short practice trial. During this practice trial AFIRA always provided correct diagnoses and recommendations. However, participants were informed that the aid's reliability is high but not perfect. They were cautioned to always cross-check the proposed diagnoses before confirming it.

After this introduction the experimental run started. Participants were randomly assigned to one of the four experimental groups. Independent of the specific experimental group and the length of the experimental session (one 35-min block for Groups 1 and 2, four 35-min blocks for Groups 3 and 4), all participants were instructed that the whole experiment would include a total of five 35-minutes blocks. This instruction was given to assure that all participants worked with the same attitude and expectation and were not able to anticipate the real ending of the experiment. For Groups 3 and 4, the blocks were separated by short breaks of about 3 minutes.

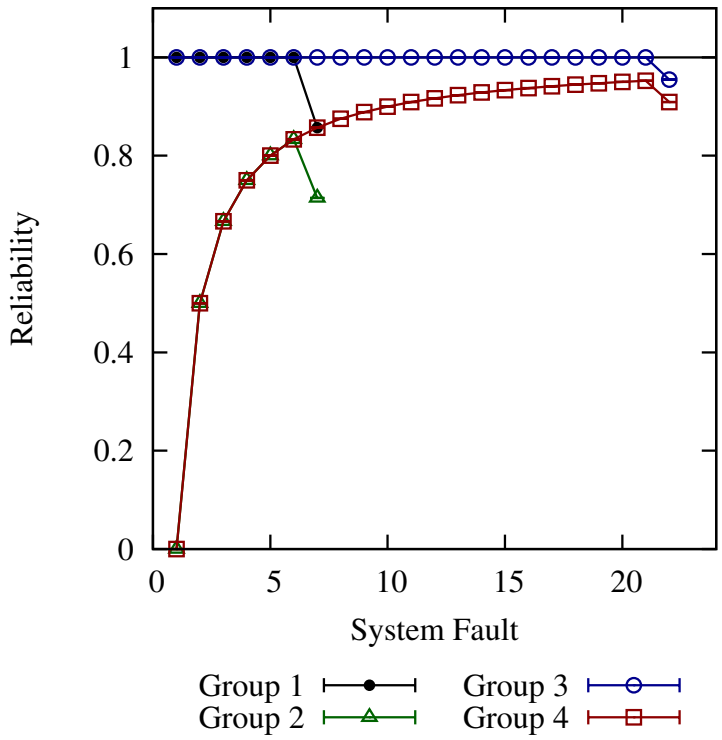

Figure 4.2: Study II. Experienced Reliability. This figure shows the reliabilty an operator experiences over time in interaction with the automation. The reliability for the four groups results from the distribution of system faults and the aid's correct diagnoses and failures as shown in figure 4.1

After the automation failure at the end of the session (first failure for Groups 1 and 3, second failure for Groups 2 and 4), the simulation stopped as soon as the participant decided to either follow the aid's advice or disagreed with AFIRA's diagnosis. Participants were then asked about their approach of automation verification by means of the Automation Verification Questionnaire (for details see dependent measures below). Ratings of subjective trust in the components of the AutoCAMS 2.0 system (e.g., oxygen, nitrogen, carbon dioxide subsystems) and AFIRA as well as reliability ratings were collected after each block. Trust ratings

were additionally collected before the first block.

4.1.5 Dependent Measures

Dependent measures were derived from questionnaires and from data that were logged during the experiment, including participants' mouse-clicks and AutoCAMS system dynamics.

Subjective trust in the diagnostic function of AFIRA was assessed by asking the participants how trustworthy they thought AFIRA was (How much did you trust the assistance system AFIRA?). Respondents answered on a 10-point Likert-type scale ranging from not at all to absolutely.

Subjective reliability rating of the diagnostic function of AFIRA was assessed by asking the participants to rate reliability ranging from 0 to 100 % reliability (Please rate the reliability of the individual subsystems: AFIRA diagnosis).

To avoid any demand characteristics, the specific questions relevant for the study were part of a larger questionnaire consisting of 18 questions that asked for subjective ratings of confidence in own performance and performance estimates, trust and estimated reliabilities not only for AFIRA but for all subsystems of AutoCAMS (e.g., oxygen and nitrogen subsystems). The questionnaire can be found in Appendix D.3.

Measures used to assess the effort invested in **automation verification** included

(a) *Automation Verification Time* (AVT), defined as the time interval (in seconds) from the appearance of the master alarm until sending a first repair order, regardless of whether this repair order was correct or wrong.

(b)*Automation Verification Information Sampling of Necessary System Parameters* (AVIS-N), defined as the proportion of all system parameters accessed that were necessary to verify a given diagnosis unambiguously.

Performance consequences of a possible **automation bias** in terms of *commission error* were analyzed by the proportion of participants who followed the aid's advice when the automation failed and provided a wrong diagnosis at the end of the experiment (first failure for Groups 1 and 3, second failure for Groups 2 and 4).

In addition, the underlying *determinants of commission errors* were analyzed. For this purpose, the simulation was stopped as soon as a participant had decided to either follow the aid's wrong advice or disagree with it, and participants were then asked questions about their approach of automation verification by means of a standardized Automation Verification Questionnaire (AVQ). Specifically, they had to provide information about (a) which diagnosis was proposed by AFIRA, (b) which parameters they had sampled to verify the aid's diagnosis, and (c) what the critical relations were between the parameters accessed (the relation between parameters provides the critical information needed to disambiguate similar system faults). This questioning was done in order to check to what extent the participants were aware of the steps they had performed and the system information they had accessed. Based on the AVQ results, we assessed how many participants committing a commission error made this error because of

(a) an *incomplete automation verification*, that is, the proportion of all system parameters accessed that were necessary to cross-check a given diagnosis unambiguously (AVIS-N) was less than 100 percent

(b) a *complete automation verification without awareness*, that is, participants

looked at all information needed to verify the aid's diagnosis (AVIS-N = 100%) but were not able to report what they had seen

(c) a *discounting of contradictory information*, that is, participants looked at all necessary parameters (AVIS-N = 100%) and were able to report the contradictory information but nevertheless followed the aid's wrong diagnosis.

4.2 Results

4.2.1 Perceived Reliability and Subjective Trust in Automation

Effects of positive and negative experience with AFIRA on perceived reliability and subjective trust were explored on the basis of data from Experimental Groups 3 and 4.

A 2 (Group) x 5 (Block) ANOVA revealed significant main effects of Group, $F(1, 41) = 4.62$, $p < .04$, and Block, $F(4, 164) = 10.43$, $p < .001$, as well as a significant Group x Block interaction, $F(4, 164) = 5.56$, $p < .001$. Trust development is shown in Figure 4.3. As expected, trust development depended on the experience the participants made with the aid. After familiarization with AFIRA during training, both groups started with a relatively high level of initial trust. Trust further increased for participants of Group 3 who experienced a reliable automation. At the end of Block 4, they experienced the first automation failure. Trust dropped down to a level lower than the initial trust. Participants of Group 4 experienced a first automation failure already in the beginning of Block 1. A sharp decrease of trust can be seen at the end of Block 1, even though the automation

failure was followed by five correct diagnosis before the trust rating after Block 1. Trust increased again in the following blocks with correct automation advice, but it never reached the trust level of Group 3. Trust decreased again after the second failure in Block 4, yet less steep than after the first failure.

Trust development parallels development of subjective and objective reliability. As becomes evident from Figure 4.4, trust development reflects the development of the perceived reliability and the objective, experienced reliability. Note that participants were asked for reliability ratings only after each block, not before the first block, so Figure 4.4 shows no pre-test ratings as included for trust in Figure 4.3.

Reliability ratings are consistently underestimating objective reliability but reflect the trend of objective reliability. Subjective reliability was higher for Group 3 than for Group 4. It increased with increasing objective reliability and decreased when objective reliability decreased. A 2 (Group) x 4 (Block) ANOVA showed a significant effect of Group, $F(1, 41) = 14.71$, $p < .001$, Block, $F(3, 123) = 11.05$, $p < .001$, and a significant Group x Block interaction, $F(3, 123) = 5.87$, $p < .001$, for subjective reliability ratings of AFIRA's diagnostic performance.

4.2.2 Automation Verification During Reliable Automation Support

To explore whether the effects seen in subjective trust ratings would also be reflected in differences in automation verification behavior, we compared to what extent participants of Groups 3 and 4 sampled necessary parameters before confirming the proposed diagnosis. Only system faults for which AFIRA provided

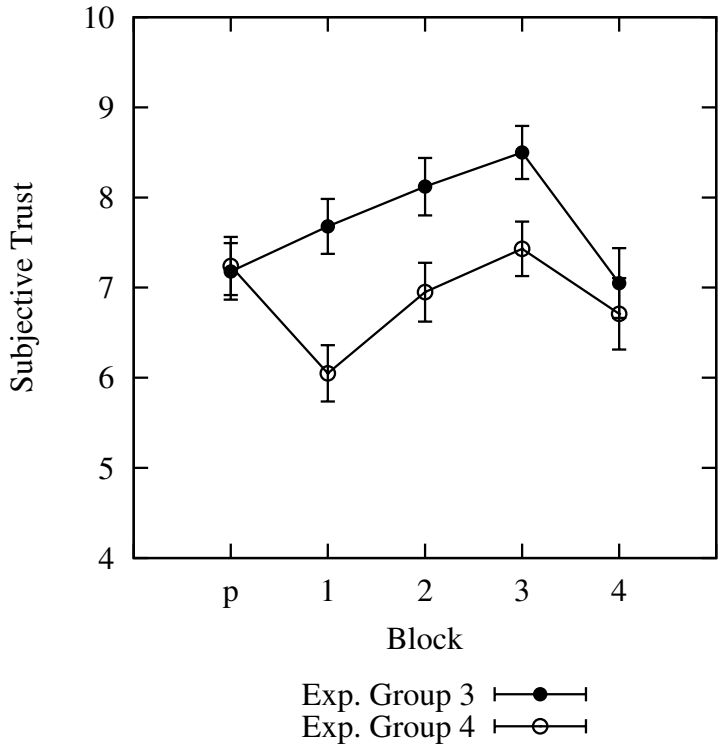

Figure 4.3: * Study II. Subjective Trust in Automation. Trust ratings after familiarization with AFIRA during training and after Blocks 1-4.

a correct diagnosis were considered for this analysis. A 2 (Group) x 4 (Block) ANOVA revealed a significant Group effect, $F(1, 42) = 6.82$, $p < .02$. Neither the Block effect, $F(3, 126) = 1.11$, nor the Group x Block interaction, $F(3, 126) < 1$, were significant. The effects are shown in Figure 4.5. The experience of an automation failure early in the experiment (Group 4) led to an increase in automation verification. Participants with early failure experience sampled more necessary parameters ($M = 97.4\%$) than participants without early failure experience ($M = 92.0\%$). This difference persisted throughout the experiment.

There was no difference in automation verification time between groups, $F <$

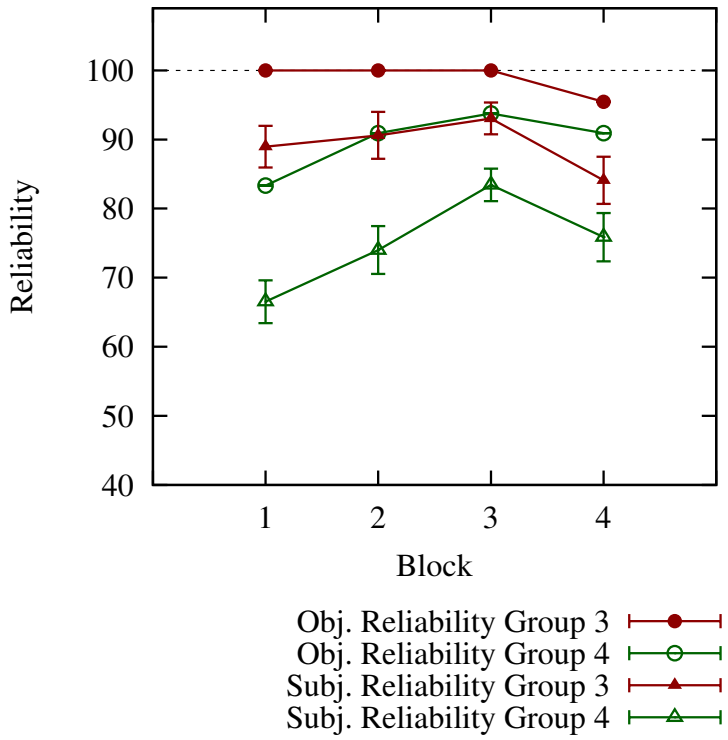

Figure 4.4: Study II. Objective Reliability and Subjective Reliability Rating

1.

4.2.3 Automation Bias and Automation Verification in Case of Automation Failure

The false diagnosis at the end of the experimental session was the first false diagnosis for half of the participants, for the other half it was the second false diagnosis they experienced. A significant difference was found between participants who had experienced a false diagnosis before, and participants with no prior failure experience, $\chi^2(1) = 5.10, p < .03$. The risk of a commission error was higher for

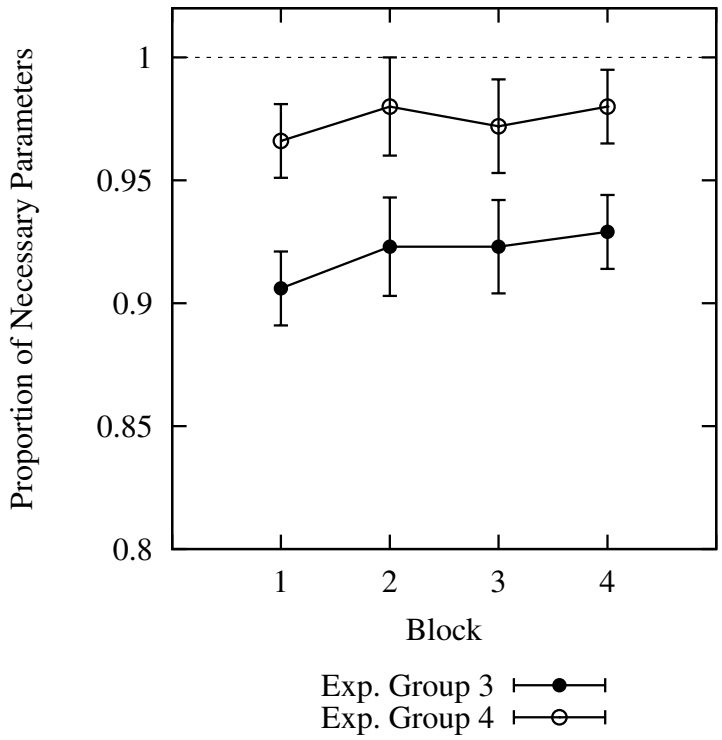

Figure 4.5: * Study II. Automation Verification During Reliable Automation Support

participants with no prior failure experience. Whereas 9 out of 44 (20.5%) participants followed AFIRA's diagnosis even though it was wrong when it was the first false diagnosis, only 2 (4.5%) participants followed AFIRA's false diagnosis when it was the second false diagnosis they experienced.

The number of correct diagnoses prior to the automation failure did not have any significant effects on risk of commission errors, $\chi^2 < 1$. Table 4.2 shows the number of participants who committed a commission error when AFIRA provided a false diagnosis at the end of the experimental session.

Table 4.2: Study II. Participants Committing a Commission Error by Following the False Diagnosis at the End of the Experiment

	No of correct diagnosis prior to false diagnosis		
Prior failure experience	5	20	Total
No	6	3	9
Yes	0	2	2
Total	6	5	11

4.2.4 Microanalysis of Commission Errors

For the early automation failure, the proposed diagnosis was a defective oxygen sensor while the true system fault was a stuck-open oxygen valve. A defective oxygen sensor did not occur again during the trials.

For the late automation failure, the proposed diagnosis was an oxygen blockage while the true system fault was a defective mixer valve. To verify the aid's diagnosis, it was necessary to cross-check four parameters (oxygen tank level, oxygen flow, nitrogen flow, standard flow rates).

Out of the eleven participants who followed the wrong automation advice at the end of the experiment, only six would be classified as complacent as they did not check all the information that was necessary to verify the aid's diagnosis.

The other five participants followed the wrong automation advice despite having checked all parameters that were necessary to realize that the automatically generated diagnosis was wrong. Four of these participants seemed to have conducted these cross-checkings without or with less attention. This finding was revealed by the results of the AVQ questionnaire that was administered after they

had confirmed the aid's false diagnosis. Although all five participants in fact had checked all necessary system information to verify the aid's diagnosis, four of them were not able to recall correctly what they had seen. Three of these participants stated that the nitrogen flow was on standard level - which is an indicator for the system fault that was wrongly proposed by the aid - although it was actually much lower, replicating the phantom memory phenomenon reported by Mosier et al. (1998, 2001). Another participant could only report the relation between two parameters that is necessary to exclude one similar system fault but failed to report the other relation that is necessary to exclude a second similar system fault and thus arrive at an unambiguous diagnosis.

Only one of the eleven participants committed a commission error despite being aware of all the contradictory system information. He checked all the parameters that were necessary to verify the diagnosis and stated all necessary relations correctly. In addition, he noted that the automated aid was helpful but to make sure he always checked the system information. Despite this, he followed the aid's wrong advice.

In contrast, out of the 77 participants who had correctly identified the aid's wrong diagnosis, only four were not able to recall all necessary parameters that they had sampled before.

4.3 Discussion

The experience of an early automation failure led to decreased trust and increased automation verification information sampling. It could reduce the risk of commission errors but did not prevent it completely.

Participants were rather conservative in their reliability ratings, they consistently underestimated the objective reliability, Group 4 with early failure experience even more so. However, subjective reliability ratings did reflect the trend of objective reliability. Participants were able to calibrate their trust pretty good in accordance with the objective reliability. Development of subjective trust paralleled the development of objective reliability. Early failure experience caused a sharp decline in trust that recovered only slowly over time. In fact, the trust level of Group 4, who experienced an automation failure early on, never reached the trust level of Group 3, who experienced 100 % reliability of the aid before the aid failed at the end of Block 4. Trust declined again after a second failure experience, but this decline was less sharp. This reflects the development of objective reliability. Objective reliability (defined as number of correct diagnosis / total number of diagnosis) was lower for Group 4 than Group 3 throughout the experiment. Thus, it comes at no surprise that subjective trust ratings of Group 4 stayed at a lower level than trust ratings of Group 3 participants. For Group 3 the automated aid provided over 95% correct diagnosis. A single automation failure after more than two hours experience of a reliably working aid caused a sharp decrease in trust. These results are in accordance with findings about trust development reported by Lee & Moray (1992).

Trust decrease after the first failure in Group 3 at the end of the session was even sharper than trust decrease after the first failure in Group 4 in the beginning of the experiment. This is surprising since objective reliability drops from 100 % (100 % correct diagnosis during training session with AFIRA) down to 83% (5/6 correct diagnosis by the end of the first block, training experience excluded) for Group 4, while for Group 3 objective reliability drops from 100% down to 95%

(21/22 correct diagnosis by the end of Block 4). The even sharper decline of trust in Group 3 compared to the first trust decrease in Group 4 might be attributed to the timely proximity of failure experience and trust rating. Group 3 experienced the failure right before the trust rating while Group 4 experienced the failure followed by five correct diagnosis before the trust rating.

Which influence does timing of automation failure have on trust and automation use? Of course, automation failure cannot be predicted or even timed in real-world scenarios. But timing of automation failures can be relevant in training design. It has been shown in other research and in this study that failure experience leads to decreased trust and increased automation verification (Bahner et al., 2008; Smith, 2012), and can reduce the risk of commission errors. This study showed that trust decrease was stronger after a late first failure compared with trust decrease after an early first failure. Since our experiment ended after the late failure, we have no information about the further development of automation verification. It would be interesting to see if not only trust decreases relatively more but if automation verification would also increase relatively more. This would suggest that failure experience should be integrated in training after a longer period of working with reliable automation support. This remains an open question and needs further research.

With the distribution of correct diagnosis and automation failures in this study, we could not realize identical experienced overall reliability of the automation. It would be interesting to see if participants who have experienced automation with the identical objective reliability but a different timing of automation failures show comparable trust and automation bias. Figures 4.6 and 4.7 illustrate the idea of identical experienced reliability with different timing of automation failures.

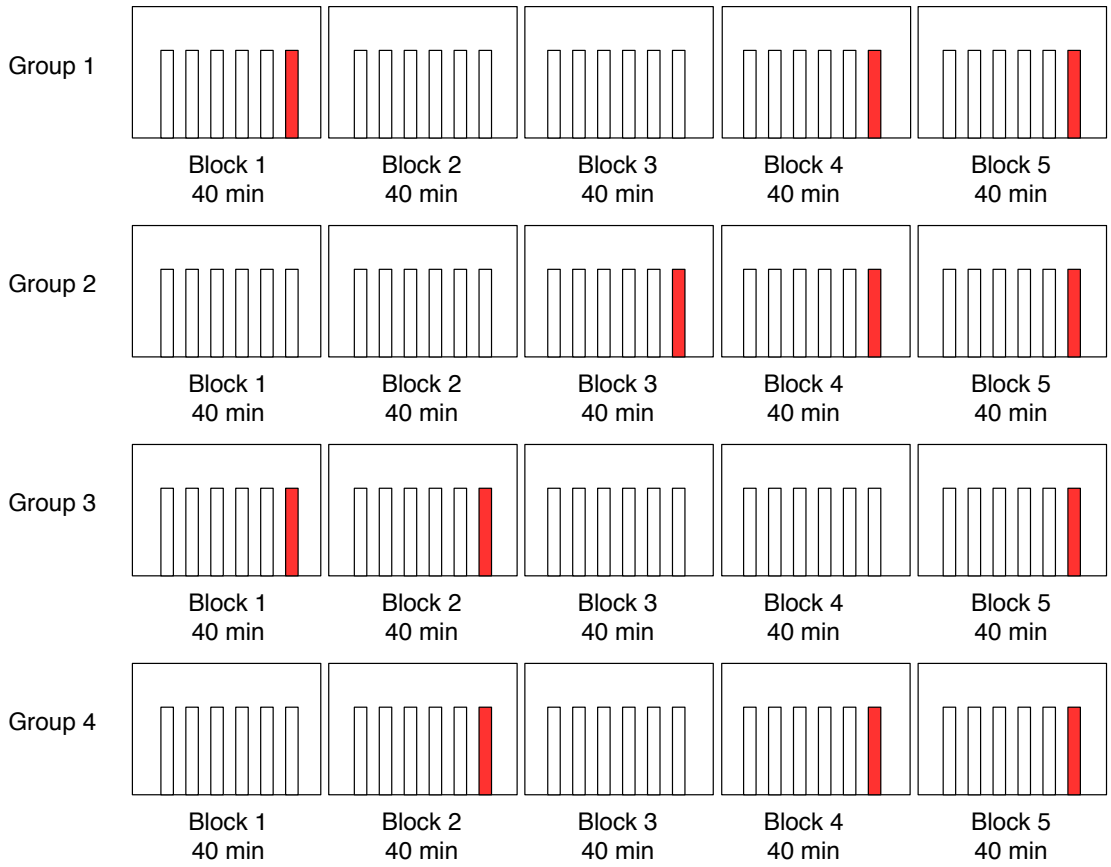

Figure 4.6: Possible failure distribution to realize identical experienced reliabilities at different points in time.

With a low number of events, as in this experiment, one automation failure has a dramatic effect on objective reliability. It is not clear if the same drastic effects in trust and perceived reliability would be found if more events could be studied such that one failure causes only a slight decrease in reliability (e.g., 1 false diagnosis in 1000 events, i.e., 99.9% reliability). It is not feasible to study such a high number of events resulting in very high reliability rates using AutoCAMS. With a similar timing of system faults (one fault in about 6 minutes), it would take 100 hours to experience 1000 system faults and AFIRA diagnoses. This is not only a problem

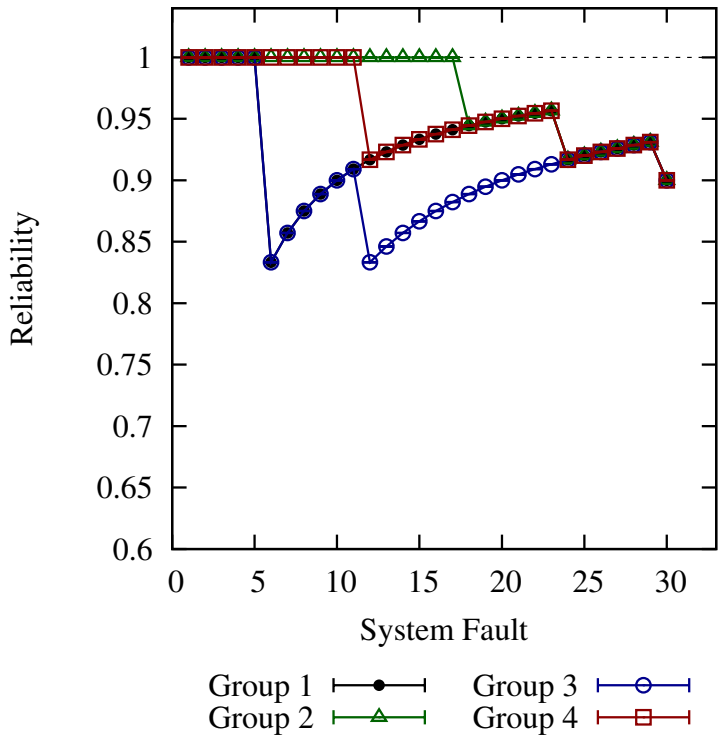

Figure 4.7: Experienced reliability for the failure distribution shown in Figure 4.6

with the current study. Due to time constraints in a laboratory experiment, in a lot of studies reliability rates are rather low compared to reliability rates that are acceptable in industry, medicine, aviation, or other areas.

Early failure experience did not only reduce trust but also led to a higher sampling rate. In contrast to trust, which recovered slowly over time, the sampling rate stayed at a very high level. This replicates results from Bahner, Hüper, et al. (2008). Early failure experience was also associated with reduced risk of commission errors. However, it could not prevent it completely. Similar results are reported by Bahner, Elepfand, et al. (2008) who found a decreased risk of omission errors after failure experience.

The duration of failure-free automation experience (half an hour with 5 correctly diagnosed system faults versus more than two hours with 20 correctly diagnosed system faults) did not affect automation bias in terms of commission errors. This contrasts results found by Molloy and Parasuraman (1996). In a simple visual discrimination task as well as in a multitask scenario with system monitoring, tracking, and fuel management, they found higher detection rates for early failures than for late failures. However, in their study early failures occurred within the first ten minutes, late failures in the last ten minutes of a 30-minutes trial. In the present study, the automation failure occurred after 30 minutes or two hours.

In the present study, long positive experience with an automation did not diminish the effects of an early failure experience. However, the total study duration was only about two hours. It is unclear if automation verification would decrease and risk of commission errors would increase again after a longer period of positive experience with the automation.

The results of this study support the proposed feedback loops. Failure experience reduced trust and led to more complete verification. This in turn reduced the risk of automation bias and commission errors. The results further indicate that the negative feedback loop in case of an automation failure has a stronger effect than the positive feedback loop in case of reliable automation. The effect of a single automation failure on trust could hardly be compensated by 20 correct diagnosis of the aid. Trust was offset so much that it never reached the level of trust that participants developed who had no prior failure experience. The effect of a second failure on trust was not as grave but still reduced trust. This development of trust reflects the development of the objective reliability, suggesting that participants were able to calibrate their trust according to the automation's reliability.

The effect was even more persistent on the behavioral level. Automation verification stayed on a very high level over the entire time of the experiment even when trust was regained.

Eleven out of 88 participants (12.5%) followed the wrong recommendation of the automated aid at the end of the experimental session. This is a rather low number compared to the high commission error rates of up to 50% in study I. In study I, all participants experienced only one automation failure after 18 correct diagnosis (almost 95% correct diagnosis). In study II, half of the participants experienced an automation failure right at the beginning of the experimental session. When we consider only those participants without prior automation failure experience (Groups 1 and 3), the commission error rate is 20.5% (9 out of 44). This is still lower that in study I, but it is comparable to the numbers reported by Bahner, Hüper, et al. (2008).

After the participants either approved the aid's diagnosis or sent an alternative repair order, the simulation stopped and participants were asked to answer the Automation Verification Questionnaire. Participants were asked to provide information about the last diagnosis, the parameters they sampled to verify the diagnosis, and the critical relations between the checked parameters. They did not have to give exact values of flow meters or tank levels but only had to indicate if the flow was on standard value, lower or higher, and if it matched the outflow from the tanks. The exact values were not necessary to detect that the aid provided a wrong diagnosis, the relations were sufficient.

With the information from the AVQ in combination with the logfile data about information sampling we identified three different causes for committing a commission error: incomplete automation verification, complete verification without

awareness, and discounting of contradictory information. Six participants sampled only a part of the necessary information. They committed the commission error due to an incomplete automation verification. Five participants committed the error despite sampling all necessary information. Four of them could not recall what they had just seen. One participant could not recall the critical relation between two parameters he checked. Three participants stated that nitrogen flow was on standard when it actually was much lower. A standard nitrogen flow was to be expected given the aid's diagnosis. This indicates that participants remembered what they expected to see given the aid's proposed diagnosis. Mosier et al. (1998, 2001) called this the phantom memory phenomenon.

One participant was aware of the contradictory system information but decided to follow the aid's recommendation anyway. He could not give a clear reason why he followed the wrong advice. In this case, we attribute the commission error to an active discounting of contradictory information.

In contrast, only 5% of the participants who had correctly identified the aid's wrong diagnosis were not able to recall all necessary parameters that they had sampled before.

This finding supports the looking-but-not seeing hypothesis derived from the results of the first experiment.

Chapter 5

Study III: The Impact of Operator Functional State

In the third experiment we studied the influence of operator functional state and degree of automation on operator performance in interaction with a decision aid. Participants worked with AutoCAMS during the day and after prolonged wakefulness during the night.

Following the compensatory control model (Hockey, 1997; Hockey, 2003b) we expected that primary task performance can be protected during normal operation even after prolonged wakefulness, i.e., diagnostic performance and fault management performance were expected to be stable. This would come at the cost of secondary task performance decrements and increased effort and workload. Return to manual performance was expected to suffer.

Automation support has been shown to reduce negative effects of stressors, and operators preferred higher DOA support when under stress (Sauer et al., 2011; 2013). We expected that higher DOA support can help protect performance

of sleep-deprived operators better than lower DOA support.

Effects on complacency and automation bias are difficult to predict based on previous research. In the context of supervisory control tasks, the use of simplified strategies and a shift towards less monitoring after one night of sleep deprivation are reported (Hockey et al., 1998; Sauer et al., 2003). Accordingly, we would expect participants to reduce verification effort and sample less information to validate the automated aid's diagnoses. As a consequence, the risk of commission errors would be higher during the night after prolonged wakefulness.

Other research suggests that the propensity to take risks is reduced after sleep deprivation (Chaumet et al., 2009; Killgore, 2007). However, other studies showed the opposite, risk taking behavior was increased after sleep deprivation (Killgore, Balkin, & Wesensten, 2006; for a review see Womack, Hook, Reyna, & Ramos, 2013). Sleep-deprived participants might be more careful in the interaction with the automated aid and invest more effort in automation verification in order to prevent overlooking failures. In this case, we would expect higher information sampling and lower risk of commission errors.

5.1 Methodology

5.1.1 Participants

32 engineering students (25 male, 7 female) ranging in age from 19 to 32 years (M = 24.9) participated in the study. Based on results in a Morningness-Eveningness Questionnaire (Griefahn, Kuenemund, Broede, & Mehnert, 2001) extreme evening types were excluded from the experiment. Participants were paid 150 Euro for

completing the study.

5.1.2 Apparatus: AutoCAMS 2.0

The same simulation of a supervisory process control task, AutoCAMS 2.0, was used as in the first and second experiment. The automated aid provided Information Analysis Support and Action Implementation Support.

5.1.3 Design

The study used a 2 (Time of Day: Day vs. Night) x 5 (Block) x 2 (DOA: IA Support vs. AI Support) design with DOA defined as between-subjects factor, and Time of Day and Block defined as within-subject factors. Participants were randomly assigned to one of the two DOA groups. Half of the participants ($n = 16$) worked with IA Support, the other half with AI Support. The sequence of day and night session was balanced within each experimental group. The study design is illustrated in Figure 5.1.

During Blocks 1 and 5, participants had to perform fault identification and management manually, during Blocks 2, 3, and 4 they were supported by the automated aid. During the first session ("Day" for half of the participants, "Night" for the other half), six system faults occurred in each block which were all correctly indicated and diagnosed by the automated aid. The second session was identical to the first session, with one exception. During the second session an additional 7th fault occurred at the end of Block 4 for which AFIRA provided a wrong diagnosis (AFIRA proposed an Oxygen Valve Block when the actual system malfunction was a Mixer Valve Block). This failure of AFIRA was implemented to simulate a

“first automation failure effect”. Because the failure always occurred during the second experimental session, those participants who started the experiment with the day session experienced this failure at night. The other half of the participants experienced it during the day session.

Faults in all blocks were matched with respect to type and complexity. Thus, it was ensured that the fault identification and management procedures were equivalent for all blocks. Both groups worked with the same set and distribution of faults (see Table 5.1).

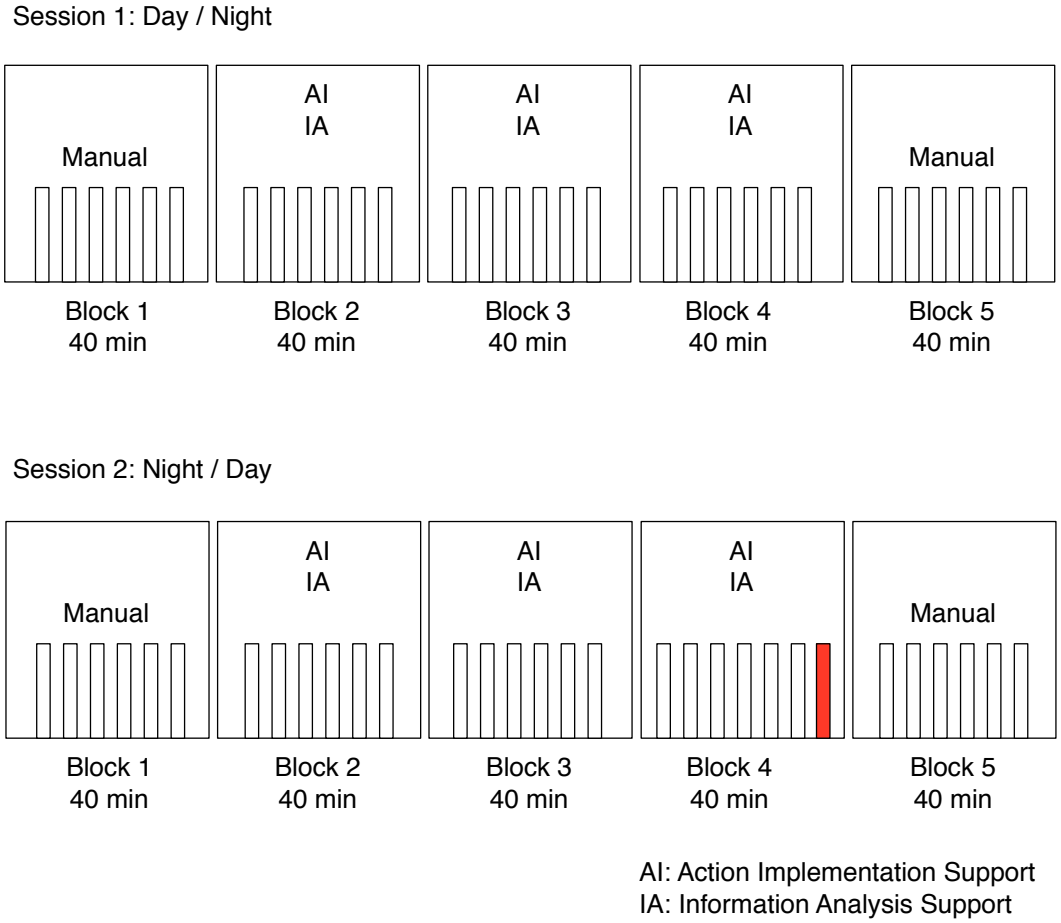

Figure 5.1: Study III. Experimental Design. The figure shows the distribution of system faults and automation failures across blocks, and the available automation support for each block. Each column represents one system fault. The red column represents the critical automation failure at the end of Block 4 in the second session of the experiment.

Table 5.1: Study III. Distribution and Timing of System Faults Across Blocks

Block	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6	Fault 7
Session 1 Block 1	Valve Blockage N_2 142s	Valve Leak O_2 462s	Valve Stuck Open O_2 841s	Valve Leak N_2 1253s	Valve Blockage O_2 1629s	Defective Sensor O_2 2014s	
Session 1 Block 2	Defective Sensor O_2 89s	Valve Blockage O_2 486s	Valve Leak O_2 858s	Valve Blockage N_2 1238s	Valve Stuck Open O_2 1624s	Valve Leak N_2 2052s	
Session 1 Block 3	Valve Leak N_2 124s	Valve Stuck Open O_2 479s	Valve Leak O_2 863s	Valve Blockage O_2 1224s	Defective Sensor O_2 1610s	Valve Blockage N_2 2049s	
Session 1 Block 4	Valve Leak N_2 102s	Valve Blockage O_2 459s	Defective Sensor O_2 828s	Valve Blockage N_2 1227s	Valve Leak O_2 1574s	Valve Stuck Open O_2 2060s	
Session 1 Block 5	Valve Blockage N_2 142s	Valve Leak O_2 462s	Valve Stuck Open O_2 841s	Valve Leak N_2 1253s	Valve Blockage O_2 1629s	Defective Sensor O_2 2014s	
Session 2 Block 4	Valve Leak N_2 102s	Valve Blockage O_2 414s	Defective Sensor O_2 742s	Valve Blockage N_2 1087s	Valve Leak O_2 1474s	Valve Stuck Open O_2 1820s	Valve Blockage Mixer 2191s

Note. Blocks 1, 2, 3, and 5 were identical in Session 1 and 2. Fault distribution in Block 4 was identical in Session 1 and 2, but timing differed due to the additional Fault 7. Fault 7 in Block 4 in Session 2 (Mixer Valve Blockage) was falsely diagnosed by AFIRA as O_2 Valve Blockage.

5.1.4 Procedure

The experiment consisted of two practice sessions and two experimental sessions distributed across four days. The first practice session lasted four hours and included familiarization with the AutoCAMS system. Participants were introduced to the different subsystems and trained to manually identify and manage all possible system faults. On the second day, all participants had to perform a 45-minutes test trial which served to test their acquired performance skills according to a predefined criterion. All participants passed this test successfully.

On the third day, the first experimental session took place. For half of the participants, this session was scheduled during the day (10:00 a.m. - 2:00 p.m.), for the other half it was scheduled after 20+ hours of continuous wakefulness, during the night at the nadir of the circadian system (4:00 a.m. - 8:00 a.m.). Before the session started, participants were assigned to one of the two experimental groups (Information Analysis Support or Action Implementation Support) and familiarized with their aid. During this trial all recommendations provided by AFIRA were correct. However, participants were informed that the aid's reliability is high but not perfect, and cautioned to validate the proposed diagnoses before initiating a repair. After this introduction, the first session of the experiment started, consisting of five blocks of 40 minutes each. During Blocks 1 and 5 all participants worked manually without the assistance of AFIRA. During Blocks 2, 3, and 4 they were supported by AFIRA.

The second experimental session took place one week later. The second session was identical to the first session, with the only difference that a first automation failure occurred at the end of Block 4.

Participants were instructed to get up at 8:00 a.m. on experimental days which was controlled by an actiwatch. During both day and night session, sleepiness and workload ratings were collected. The Psychomotor Vigilance Task (PVT; Wilkinson & Houghton, 1982; Dinges & Powell, 1985; Mueller, 2007) was administered before the first block started. Subjective sleepiness was assessed before the first block and after each single block. Subjective workload was assessed after each block.

5.1.5 Dependent Measures

Dependent measures were derived from questionnaires and from data that were logged during the experiment, including participants' mouse-clicks and AutoCAMS system dynamics.

Three **primary task performance** measures were calculated for each block:

(a) *Percentage of correct diagnoses* was the percentage of the six faults occurring per block for which the first repair order sent was correct, a measure of quality of fault identification performance.

(b) *Fault identification time* was defined as time (in seconds) from appearance of the master alarm until the correct repair order was issued. This measure was used to assess speed of fault identification performance.

(c) *Out-of-target error* was defined as the time (in seconds) the most critical system parameter (oxygen) was out of target range when a system fault was present, a measure of quality of fault management performance.

Secondary task performance was assessed by two measures:

(a) *Simple Reaction Time* defined as mean response time (in milliseconds) to

the appearance of the “communication link” icon (”connection check” task) and

(b) *Prospective Memory Performance* defined as proportion of entries of carbon dioxide records that were provided within the correct time interval (at every full minute with a tolerance of 5 seconds).

Only performance during periods when a participant had to deal with a system fault was considered for secondary task performance.

Subjective workload was assessed by the *NASA Task Load Index* (NASA-TLX; Hart & Staveland, 1988) and was defined as the mean of the ratings provided for the six subscales.

Sleepiness measures included subjective measures and performance measures:

(a) *Subjective sleepiness* was assessed by the Stanford Sleepiness Scale (SSS; Hoddes, Dement, & Zacone, 1972).

(b) *Performance indicators of sleepiness* were derived from the Psychomotor Vigilance Task (PVT; Wilkinson & Houghton, 1982; Dinges & Powell, 1985). PVT is a short-term simple visual reaction time task which was developed to evaluate effects of arousal-related stress on performance. In this study we used the PEBL PVT (Mueller, 2007). Measures included the overall mean of reaction times and the mean of the 10% slowest reaction times.

Measures used to assess the effort invested in **automation verification** included

(a) *Automation Verification Time* (AVT), defined as the time interval (in seconds) from the appearance of the master alarm until sending a first repair order, regardless of whether this repair order was correct or wrong.

(b)*Automation Verification Information Sampling of Necessary System Parameters* (AVIS-N), defined as the proportion of all system parameters accessed that

were necessary to verify a given diagnosis unambiguously.

Automation bias was analyzed by the proportion of participants committing a *commission error*, defined as percentage of participants who followed the diagnosis of the automated decision aid for Fault 7 in Block 4 during the second experimental session although it was wrong.

5.2 Results

5.2.1 Sleepiness

Effects on subjective sleepiness were analyzed by a 2 (Time of Day) x 6 (Block) x 2(DOA) ANOVA. A significant main effect of Time of Day, $F(1, 30) = 184.21$, $p < .01$, a significant Block effect, $F(5, 150) = 10.65$, $p < .01$, and a Time of Day x Block interaction, $F(5, 150) = 2.61$, $p < .03$, were found. On a scale ranging from 1 ("feeling active and vital, alert, wide awake") to 7 ("almost in reverie, sleep onset soon, lost struggle to remain awake"), subjective sleepiness was rated higher during the night ($M = 4.67$) than during the day ($M = 2.44$). During the night session sleepiness increased considerably across blocks whereas during the day session it stayed on a low level. Subjective sleepiness is illustrated in Figure 5.2.

The PVT results were analyzed by a t-test. Mean response times were longer after extended wakefulness ($M = 309$ ms) than during the day session ($M = 293$ ms), $t(31) = 4.22$, $p < .01$. A similar effect emerged for the 10% slowest reactions (Ms = 446 ms and 417 ms), $t(31) = 2.15$, $p < .05$.

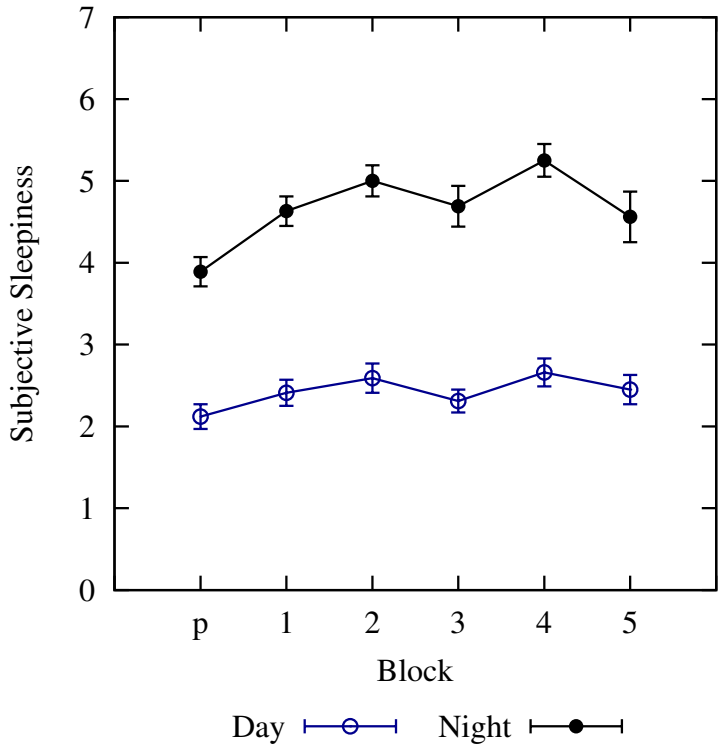

Figure 5.2: Study III. Subjective Sleepiness

5.2.2 Primary Task Performance

Primary task performance measures were analyzed by a 2 (Time of Day) x 5 (Blocks) x 2 (DOA) ANOVA. Because of the operational relevance of performance decrements in impaired functional states, the following report of effects involving the Time of Day factor will not be limited to significant effects ($p < .05$) but also consider effects which approach the conventional level of significance ($.05 < p < .10$).

Percentage of correct fault identifications varied across blocks, $F(4, 120) = 12.87$, $p < .01$. On average 91.9% and 96.1% of faults were correctly identified in the two manual blocks (Block 1 and 5). With automation support in Blocks 2 - 4

this already high level of performance increased to an almost perfect performance, more than 98% of all faults were correctly identified in Blocks 2 - 4. No effects of Time of Day or DOA were found but a complex three-way interaction, $F(4, 120) = 2.70$, $p < .04$, which, however, did not reveal any meaningful pattern of effects.

Fault identification time also profited considerably from automation support in Blocks 2 - 4 compared to manual performance in Blocks 1 and 5, $F(4, 120) = 43.59$, $p < .01$. Fault identification times tended to be shorter during the day than during the night session, $F(1, 30) = 3.87$, $p < .06$. No main effect of DOA emerged, $F(1, 30) = 1.20$, This pattern of effects was moderated by a significant Time of Day x Block interaction, $F(4, 120) = 2.97$, $p < .03$, and a Time of Day x Block x DOA interaction, $F(4, 120) = 2.41$, $p = .05$. This is illustrated in Figure 5.3.

Participants working with the less automated aid (IA Support) initially were less able than the AI Support group to protect performance at night compared to daytime performance but improved over time. They did not show any greater difficulties of return-to-manual performance during the night than during the day. In contrast, participants working with AI Support were able to perfectly maintain performance even in a state of sleep loss if automation support was available (Blocks 2 - 4). However, they showed considerably greater difficulties of return-to-manual performance during the night than during the day, reflected in the day-to-night performance difference in Block 5.

A similar pattern of effects was found for fault management performance. Quality of fault management performance as reflected in out-of-target error improved when participants were supported by the automated aid. In blocks with automation support, participants had shorter out-of-target times compared to manual

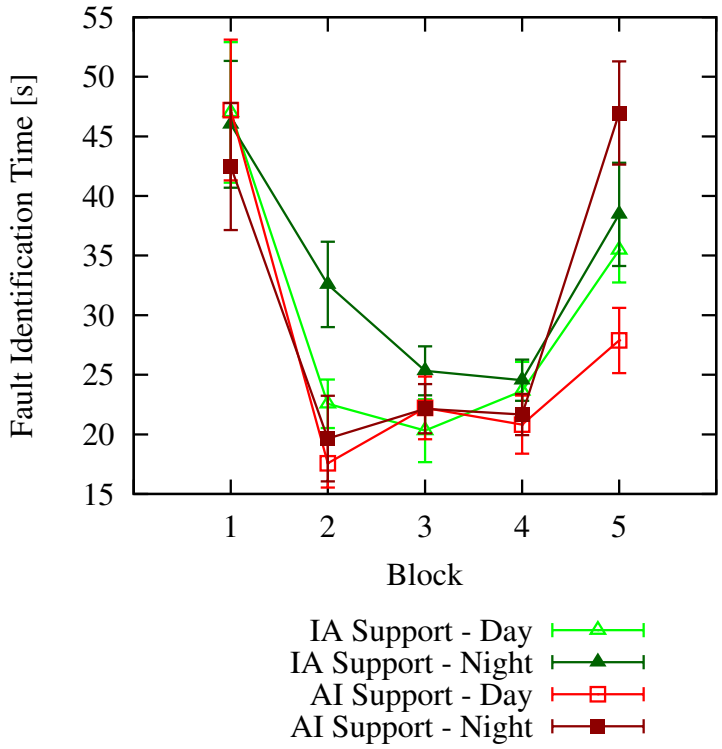

Figure 5.3: * Study III. Primary Task Performance: Fault Identification Time

performance, $F(4, 120) = 13.47$, $p < .01$. As expected, participants with AI Support showed better fault management performance than participants with IA Support, $F(1, 30) = 6.22$, $p < .02$. Performance was better during the day than during the night, $F(1, 30) = 8.55$, $p < .01$. Interactions approached significance, Block x DOA, $F(4, 120) = 2.31$, $p = .06$, Time of Day x Block x DOA, $F(4, 120) = 2.13$, $p = .08$. This is illustrated in Figure 5.4.

5.2.3 Secondary Task Performance

Performance of both secondary tasks was analyzed by a 2 (Time of Day) x 5 (Block) x 2 (DOA) ANOVA. A significant main effect of Time of Day, $F(1, 30) =$

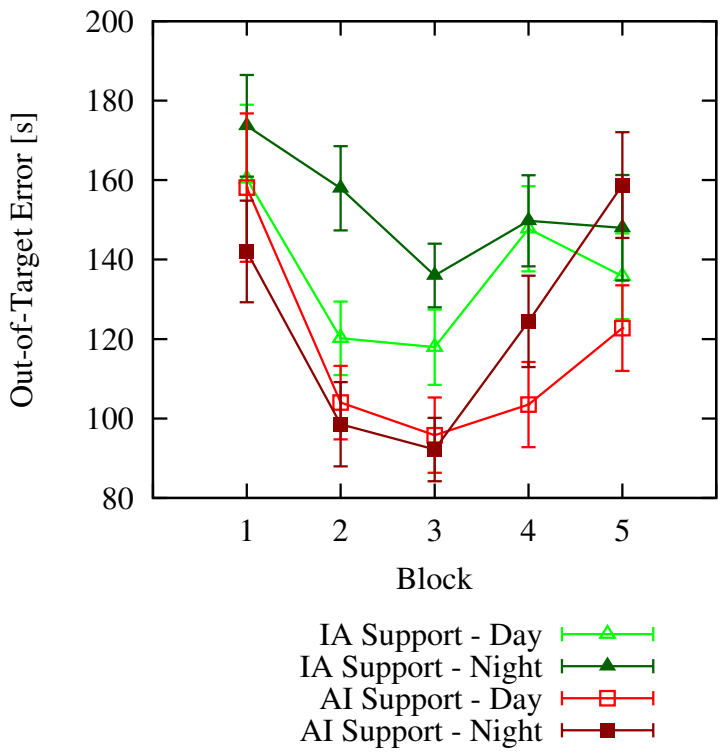

Figure 5.4: * Study III. Primary Task Performance: Out-of-Target Error

8.45, p < .01, and a significant Time of Day x Block interaction, $F(4, 120) = 3.14$, p < .02, were found for reaction times in the connection check task. No effect of DOA emerged for this measure. See Figure 5.5 for illustration. During the day, an automation benefit was observed with faster response times during automation supported blocks than during manual blocks. At night, secondary task response times increased over time, even in the automation supported blocks.

For the prospective memory task (entry of CO_2 levels), a significant Block effect was found, $F(4, 120) = 21.52$, p < .01. Participants were better able to make timely entries in blocks with automation support compared to manual performance. No other effects were found.

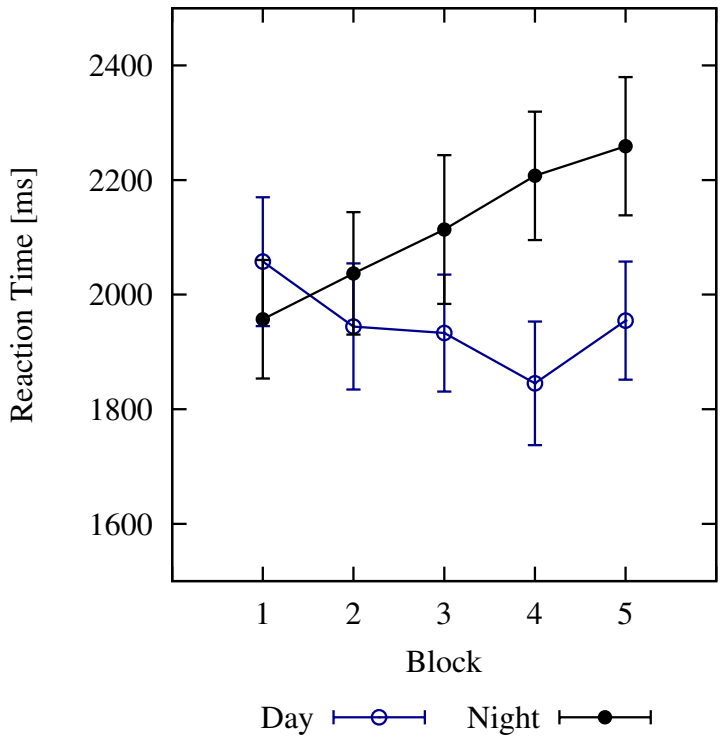

Figure 5.5: * Study III. Secondary Task Performance: Simple Reaction Time

5.2.4 Subjective Workload

Subjective workload was analyzed by a 2 (Time of Day) x 5 (Block) x 2 (DOA) ANOVA. Results are depicted in Figure 5.6 and 5.7.

Analysis of subjective workload revealed a significant main effect of Time of Day, $F(1, 30) = 8.45$, $p < .01$, and a significant Block effect, $F(4, 120) = 3.17$, $p < .02$. Participants showed higher workload ratings during the night than during the day. In addition, subjective workload was higher in the first block of each session compared to all other blocks. Neither the main effect of DOA nor any of the interaction effects were significant.

A separate analysis for the Effort scale of the NASA TLX revealed a main effect

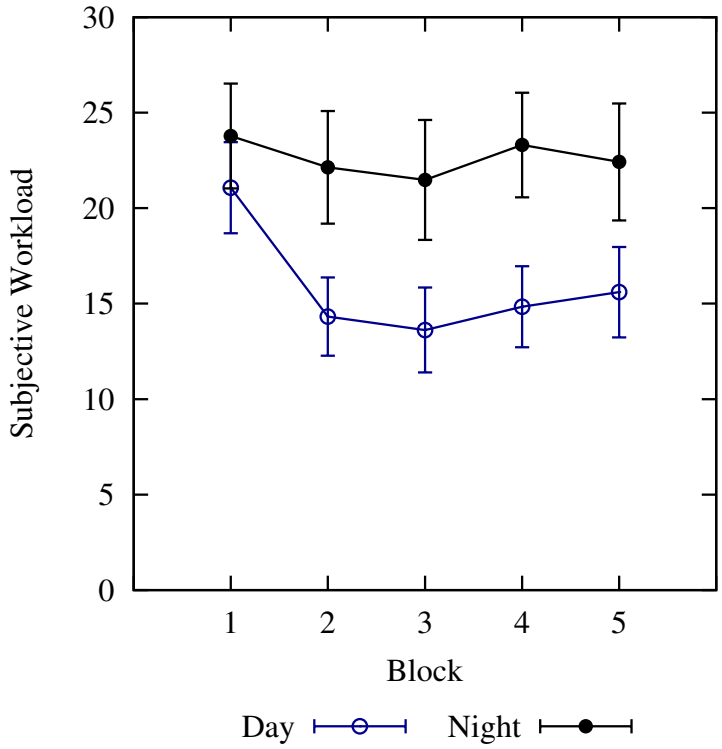

Figure 5.6: Study III. Subjective Workload

of Time of Day, $F(1, 30) = 20.04$, $p < .01$, and a Time of Day x Block interaction, $F(4, 120) = 2.92$, $p < .03$. As expected, effort was rated higher during the night than during the day. Also, during the night a sharp increase in subjective effort ratings was observed after Block 4 in which the automation failure occurred.

5.2.5 Automation Verification During Reliable Automation Support

Automation verification behavior was analyzed by a 2 (Time of Day) x 3 (Block) x 2 (DOA) ANOVA.

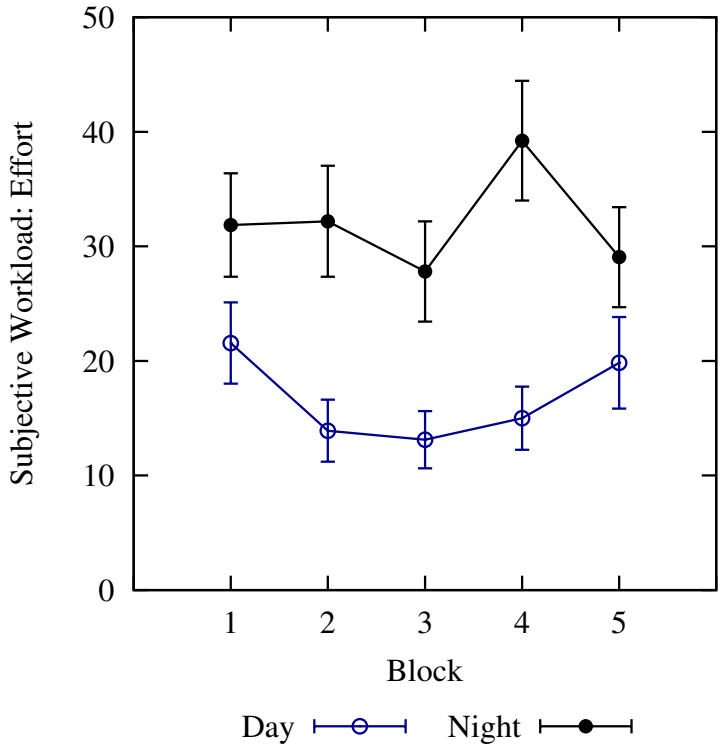

Figure 5.7: Study III. Subjective Workload: Effort

For automation verification time, a significant Time of Day effect emerged, $F(1, 30) = 6.37$, $p < .02$. Participants spent significantly more time with automation verification during the night ($M = 22.7$ s) than during the day ($M = 19.6$ s). For information sampling, the main effect of Time of Day was significant, $F(1, 30) = 4.34$, $p < .05$. Participants sampled a higher portion of necessary parameters to verify a given diagnosis during the night ($M = 97\%$) than during the day ($M = 92\%$). There was no effect of DOA or Block for either AVT or AVIS-N.

The effect of complexity of fault diagnosis on automation verification was analyzed by a 2 (Time of Day) x 3 (Block) x 2 (DOA) x 3 (Complexity) ANOVA. It was necessary to check two parameters for verification of low complexity errors,

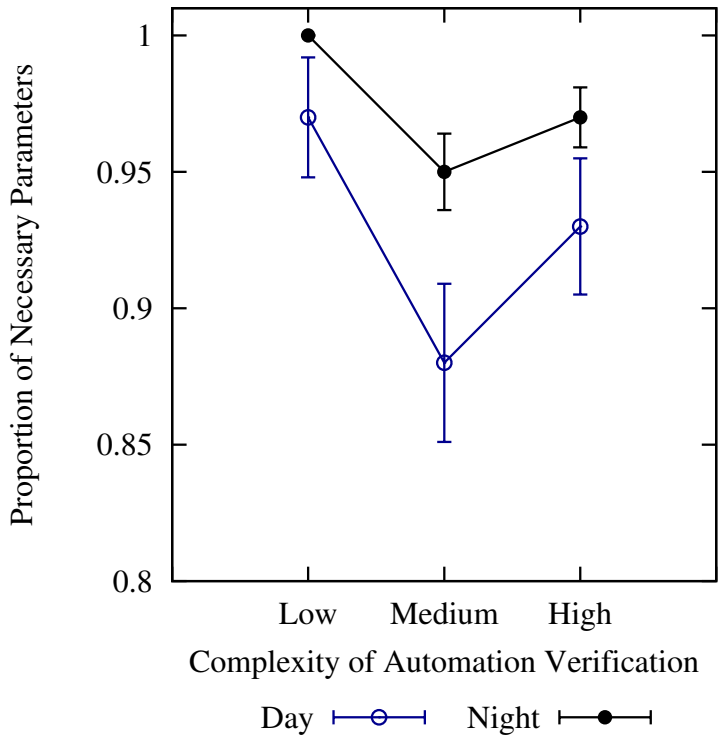

Figure 5.8: Study III. Automation Verification During Reliable Automation Support: Effect of Complexity

four parameters were necessary for medium complexity errors, and four steps were necessary for verification of high complexity errors, two of them were regulatory control actions the participants had to implement to be able to disambiguate two possible diagnoses. A marginally significant effect of Time of Day, $F(1, 30) = 3.98$, $p = .055$ and a significant effect of Complexity $F(2, 60) = 12.20$, $p < .01$ emerged. The effects are depicted in Figure 5.8. More parameters were sampled during the night. While low complexity errors were checked almost completely ($M = 98.6\%$), only 91.9% of the necessary parameters of medium complexity errors were checked. For high complexity errors 95.1% of the necessary parameters were checked.

5.2.6 Automation Bias and Automation Verification in Case of Automation Failure

Clear evidence for automation bias leading to a commission error was found by analyzing the fault identification performance for Fault 7 in Block 4 of the second experimental session when the automated aid failed for the first time and provided a wrong diagnosis. The strength of this effect was moderated by the operator functional state. Whereas seven of the 16 participants (43.8%) who experienced the automation failure during the day followed the diagnosis even though it was wrong, only one of the 16 participants (6.3%) who experienced the automation failure during the night committed a commission error. A 2 (Time of Day) x 2 (DOA) ANOVA revealed a significant Time of Day effect, $F(1, 28) = 6.63$, $p < .02$. No effect of DOA was found.

In order to investigate whether the information sampling behavior of participants who committed a commission error differed from the verification behavior of participants who detected the automation failure, an additional 2 (Commission Error) x 4 (Block) ANOVA was run, contrasting the extent to which both groups of participants verified the diagnoses of the automated aid for the 18 system faults in Blocks 2 - 4 for which AFIRA provided a correct diagnosis as well as the critical Fault 7 in Block 4 for which the diagnosis was wrong. Only those 16 participants who were confronted with the false diagnosis during the day session were considered in this analysis.

For automation verification time, a significant Block effect, $F(3, 42) = 11.05$, $p < .001$, Commission Error effect $F(1, 14) = 7.01$, $p < .02$, and Commission Error x Block interaction, $F(3, 42) = 12.46$, $p < .001$, were found. Automation verifica-

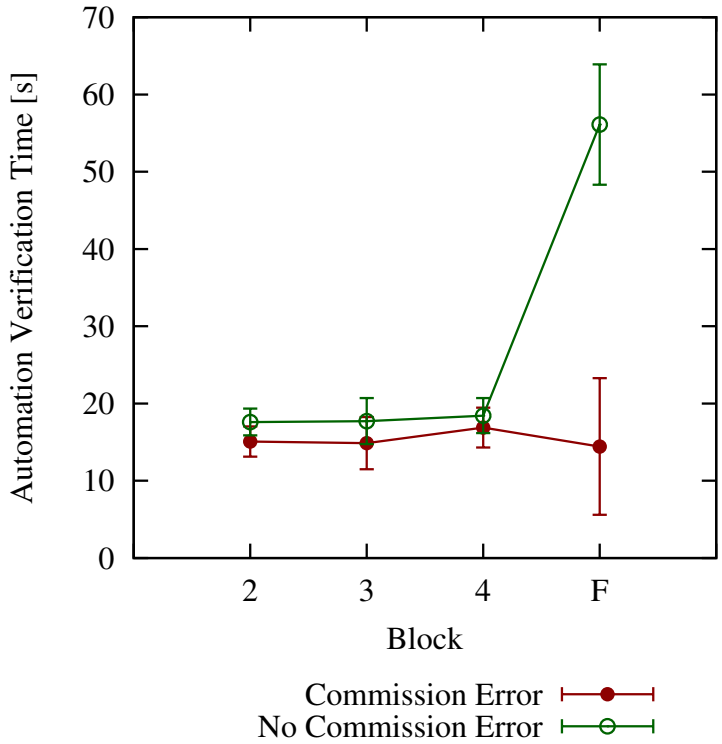

Figure 5.9: Study III. Automation Bias: Automation Verification Time for Blocks 2, 3, and 4 and the False Diagnosis.

tion time was slightly higher for participants not committing a commission error when the automated aid provided correct diagnoses compared to participants who committed a commission error. When AFIRA provided a false diagnosis, automation verification time stayed at the same level for participants who committed a commission error while it sharply increased for participants who detected that the diagnosis was wrong. The effects are depicted in figure 5.9.

For AVIS-N, a main effect of Commission Error, $F(1, 14) = 5.36$, $p < .04$, and a Commission Error x Block interaction, $F(3, 42) = 3.22$, $p < .04$, were found. The participants who committed a commission error during the day checked less

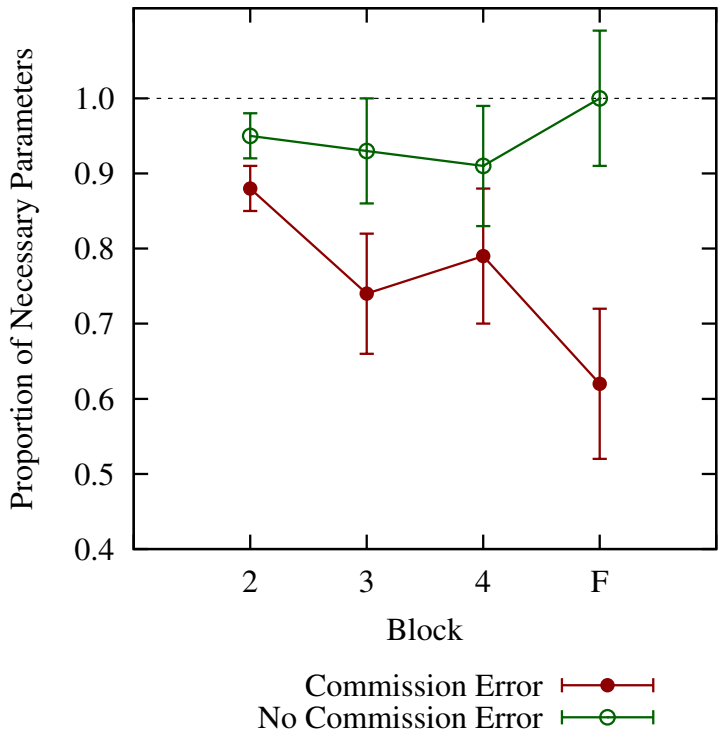

Figure 5.10: * Study III. Automation Bias: Automation Verification Information Sampling of Necessary System Parameters for Blocks 2, 3, and 4 and the False Diagnosis.

necessary parameters ($M = 75.5\%$) than the participants who did not commit a commission error ($M = 94.9\%$). Furthermore, participants committing a commission error reduced their automation verification effort over time while the sampling rate stayed at a very high level for the other participants. For the critical system Fault 7 in Block 4 participants committing a commission error on average cross-checked only 61.5% of the necessary parameters, whereas participants who detected the wrong diagnosis fully ($M = 100\%$) verified the aid's diagnosis. These effects are illustrated in Figure 5.10.

A detailed analysis of information sampling behavior of participants who com-

mitted a commission error during the day session revealed that out of the seven participants who followed the aid's false diagnosis, three did so despite having accessed all necessary parameters. The other four participants accessed only a part ($n = 3$) or none ($n = 1$) of the necessary parameters.

To check the aid's diagnosis (Oxygen Valve Block), it was necessary to sample four parameters (Oxygen Tank Level, Oxygen Flow, Nitrogen Flow, Standard Flow Rates). The Oxygen Tank Level and Oxygen Flow readings confirmed the proposed diagnosis. These two parameters were accesses by all participants except the one that sampled no parameters at all. It was necessary to additionally access Nitrogen Flow and Standard Flow Rates to falsify the proposed diagnosis and identify the true system malfunction (Mixer Valve Block). All four participants who did not sample all necessary information omitted Standard Flow Rates, three of them additionally omitted Nitrogen Flow.

The participant who committed a commission error during the night also failed to check Nitrogen Flow and Standard Flow Rates. Fault 7 in Block 4 was the first time that he did not check all necessary parameters when the aid proposed an Oxygen Valve Block. In all preceding blocks, he sampled 100% of the necessary parameters when the aid proposed an Oxygen Valve Block, which occurred once per block. There were three cases in which he did not sample 100% of the necessary parameters, all of them during his second experimental session at night: oxygen sensor in Block 2, nitrogen valve block in Block 3, and nitrogen valve block in Block 4.

5.3 Discussion

5.3.1 Effects of Operator State and DOA on Routine and Failure Performance

In study III, performance after prolonged wakefulness was studied, simulating a first night shift. The control variable Subjective Sleepiness confirmed that participants were sleepier during the night session than during the day session. This was further substantiated by the Psychomotor Vigilance Task performance which was better during the day than during the night, as reflected in the mean response time as well as the 10% slowest reactions.

Providing automation support benefitted routine performance and could help operators protect their performance even after prolonged wakefulness. Accuracy of diagnostic performance, which was already high in manual blocks, further increased when automation support was provided. Speed of diagnostic performance also improved with the introduction of an automated aid, moderated by DOA and Time of Day. Participants working with the higher DOA aid (AI Support) were better able to protect performance during the night compared to day time performance when the aid worked reliably. Participants working with the lower DOA aid (IA Support) were less able to protect performance in the beginning but improved over time. On the other hand, participant of the IA Support group showed no greater performance decrements when returning to manual performance during the night than during the day, while the AI Support group showed rather poor manual performance during the night compared to daytime performance.

A similar pattern of effects was found for fault management performance. Also

for fault management, performance improved when automation support was provided. Performance was better during the night than during the day. The AI Support group profited even more than the IA Support group, replicating results of study I. The automated aid started the fault management as soon as participants approved, so the fault management performance of the AI Support group shows the optimal fault management implemented by the automation. However, just as for speed of diagnostic performance, when returning to manual performance, fault management performance suffered more during the night in the higher DOA group (AI Support) than the lower DOA group (IA Support).

During the day session, secondary task performance improved when automation support was provided. However, during the night session, secondary task performance could not be protected, response times increased over time even with automation support.

As expected, subjective workload was higher during the night than during the day. Effort was not only higher during the night than during the day. It also showed a sharp increase after Block 4 in which the automation failure occurred.

These findings suggests that participants tried to protect their primary task performance (diagnosis and fault management), and in large parts they were able to do so even after prolonged wakefulness. However, protecting primary task performance was only possible at the expense of secondary task performance and increased subjective workload and effort. These results are in line with earlier results from Hockey et al. (1998) and support Hockey's compensatory control model (Hockey, 1997; Hockey, 2003b).

Providing higher automation support could better help participants keep a high level of performance even when sleep-deprived. However, higher automation

support seemed to amplify performance decrements when returning to manual performance. So while high automation support can help protect performance after sleep deprivation when the aid works reliably, it can have detrimental effects on performance when the aid fails and the operator has to return to manual performance. This supports earlier findings by Endsley & Kiris (1995) suggesting that risks of out-of-the-loop unfamiliarity and related issues of return-to-manual performance increase with higher levels of automation.

Lower automation support was less able to protect performance in sleep-deprived operators. However, decrements when returning to manual performance were far less severe than with higher DOA support.

When decisions about providing automation are to be made, this has to be kept in mind. When the likelihood of an automation failure and the costs associated with an automation failure are low, higher DOA support should be preferred as it is better able to protect routine performance even in sleep-deprived operators. However, if automation failure is likely or associated costs are high, and operators are expected to return to manual performance also during night shifts, lower DOA support might be the better choice. Even if routine performance after prolonged wakefulness is weaker with lower DOA support, manual performance after automation break down suffers far less with lower DOA than higher DOA support.

5.3.2 Automation Verification and Automation Bias

Interestingly, we did not find evidence that the risk of complacency and automation bias might be elevated after sleep deprivation. In this study, we found participants

to spend more time verifying the aid's proposed diagnosis and sample more necessary parameters before confirming a diagnosis during the night than during the day.

This finding is in contrast with earlier findings by Hockey et al. (1998). Although they found protection of primary task performance at the expense of secondary task performance and effort, as was found in this study, they found a decrease in system monitoring. In contrast, in the current study, participants increased verification time and sampling rate during the night. This might have been an attempt to keep themselves involved and awake. The task did not offer a lot to do, so sampling information might have served as a means to get themselves involved and fight sleepiness. However, participants reduced sampling of *relevant* information even at night, so fighting off sleepiness might not be the sole reason for a higher sampling rate of *necessary* information at night.

Participants might have been aware of the elevated risk of missing an automation failure when they are sleepy, and as a countermeasure, they sampled more necessary information to make sure the aid's proposed diagnosis was correct. This is in accordance with results from Chaumet et al. (2009) and Killgore (2007) who showed a reduced propensity to take risks after sleep deprivation. Sleepy operators might be more careful and attentive in interaction with an automation and invest more effort in verifying its recommendation before accepting it, thus reducing the risk of missing automation failures. This would also fit results from a simulator study with anesthesiologists which showed that monitoring performance degraded after sleep deprivation, however, monitoring efficiency for short periods in alarm situations did not degrade (Beatty, Ahern, & Katz, 1977).

The higher information sampling rate during the night led to lower commis-

sion error rates. During the night session, only one participant followed the aid's wrong recommendation, whereas during the day session, seven participants made a commission error. This effect was directly related to the automation verification behavior. During the night session, automation verification time and automation verification information sampling was high in all blocks. However, participants who committed a commission error during the day session sampled less necessary parameters in preceding blocks and reduced sampling over time. This replicates earlier findings by Manzey and colleagues (Manzey et al., 2006; Bahner, Hüper et al., 2008). When the automation failed, verification time was at the same level as in the preceding blocks, and information sampling was incomplete. In contrast, for participants who realized that the aid's advice was wrong, information sampling was complete and automation verification time increased drastically when the automation provided a wrong diagnosis. Participants realized that there was an inconsistency between the aid's diagnosis and the system's parameters, so they invested more time in checking the system raw data before they finally did *not* accept the proposed diagnosis and instead sent an alternative repair order. This replicates results from study I.

Out of the eight participants who followed the aid's wrong advice, three did so despite complete information sampling. One participant accessed no information at all. The other four participants sampled a part of the necessary information. Interestingly, they accessed those parameters that confirmed the proposed diagnosis but omitted the falsifying information that would have enabled them to identify the true system malfunction. This replicates results from study I and II and shows yet again that incomplete automation verification is but one possible cause of commission errors.

Chapter 6

General Discussion

The present work intended to examine performance benefits and performance decrements in interaction with an automated decision aid. In three laboratory experiments, we studied the effects of degree of automation, system experience, and operator functional state on routine performance and failure performance as well as complacency and automation bias.

A simulated process control task served as an experimental task. Fault diagnosis and management was either performed manually or with the support of an automated aid. Analyzing information sampling in relation to a predefined optimal sampling allowed us to study complacency independent of possible performance consequences such as commission errors.

Results show that routine performance benefits from providing automation support. Accuracy and speed of diagnostic performance increased, fault management performance and secondary task performance improved, subjective workload decreased. On the downside, we found complacency and automation bias, skill degradation, and increased workload when returning to manual performance. De-

gree of automation, system experience, and operator functional state affected routine performance and failure performance as well as complacency and automation bias.

6.1 Effects of Degree of Automation

In study I, the best diagnostic and fault management performance and the lowest workload was found for the highest DOA support. Also secondary task performance benefited the most from the highest DOA. However, the highest DOA was also associated with increased workload and performance decrements when returning to manual performance: fault management performance decreased compared to manual performance before automation support. However, performance losses are not seen in the diagnostic performance. During training, participants were instructed to always check the automatically proposed diagnoses. Even with automation support, participants went through the diagnostic process to validate the proposed diagnosis. However, implementation of fault management steps was done automatically in the highest DOA condition, so there was no further practice of fault management. This resulted in a loss of skills needed to perform fault management manually. Freeing operators in the highest DOA condition from manual fault management led to an unexpected effect during the diagnostic process. Operators also omitted control actions that were needed in the diagnostic process. For low and medium complexity errors that did not need control actions for verification, there was no difference between DOA groups. In study III, this effect was not found. Verification did not differ between low and high DOA. Complexity of verification affected information sampling rate.

Risk of commission errors was independent of DOA, as results from study I and III showed.

6.2 Effects of Operator Functional State

Results from study III confirmed that providing automation could help protect performance even after prolonged wakefulness. High DOA support proved to be especially helpful during reliable automation support. However, performance decrements were manifest when returning to manual performance during the night. Lower DOA support could reduce the performance costs when returning to manual performance during the night. On the other hand, lower DOA was less able to help protect routine performance after sleep-deprivation.

These results suggest that desicions about automation should take into account the risk of automation failure and the cost of automation failure. While higher DOA can better support routine performance, even in sleep-deprived operators, detrimental effects on failure performance are increased. When automation is very unlikely to fail, or failure costs are tolerable, higher DOA can improve total human-automation performance more than lower DOA. However, if automation failure is likely, or costs are unacceptable, then lower DOA is preferable over higher DOA, as failure performance costs are less severe.

Protecting primary task performance was only possible at the expense of secondary task performance and increased subjective workload and especially effort, supporting the compensatory control model (Hockey, 1997; 2003b). In contrast to earlier findings (Hockey et al., 1998), operators did not adopt simpler monitoring strategies, instead automation verification was increased during the night.

This might have served as a strategy to keep involved and awake. However, total information sampling was not increased but specifically sampling of necessary information. Participants seemed to be aware of the elevated risks at night and invested more effort in information sampling in order to not miss an automation failure. This more attentive sampling payed off when the automation failed, only one participant followed a wrong automation advice at night, whereas during the day, commission error rate was rather high.

6.3 Effects of System Experience

Results of study II showed that early failure experience did not only reduce trust but also led to a higher information sampling rate and reduced the risk of commission errors. This suggests that automation failure experience can reduce automation bias effects. However, it can not prevent them completely.

Trust recovered slowly when the automation worked reliably again. Trust decrease after a second automation failure was not as sharp as after a first automation failure. This shows that participants were able to calibrate their trust according to the automation's reliability. However, experiencing an automation failure after a longer time of working with a perfectly reliable automation lead to a sharper trust decrease than early failure experience, which would not be expected based on the objective reliability.

The effect of failure experience on information sampling behavior was more enduring. Unlike trust which increased again with increasing reliability, information sampling stayed at a very high level throughout the experiment. This lead to a reduced risk of commission errors but did not prevent them completely.

It would be of great value for training design to know if late failure experience also brings about even stronger effects in information sampling just as it did in trust ratings. Since study II ended right after the late false diagnosis, we have no data about the further development of sampling behavior, so this remains an open question.

The number of correct diagnosis that were experienced prior to an automation failure did not affect the risk of commission errors. This is positive as is shows that the risk of commission errors did not increase even after a longer time of working with the aid. However, the total time of working with the aid and the number of diagnosis during this time was low compared to real-world systems. Based on this data it is difficult to predict how the risk might change if an operator works with an aid for a much longer time.

6.4 Complacency and Automation Bias

Sampling of relevant information decreased over time while sampling of necessary information stayed on a high level. This suggests that automation verification was optimized, saving resources while still controlling the automation. However, the more complex the validation procedure was, the less parameters were checked.

Despite high automation verification, a high portion of the participants committed a commission error, following a wrong automation advice. Results of study I and III show that participants committing a commission error spent about the same amount of time checking system information in preceding trial with reliable automation as participants who overrode the automation. However, when the automation's suggested diagnosis was wrong, verification time sharply increased

for participants who detected the false diagnosis. In contrast to participants who followed the wrong diagnosis, they not only sampled the information but realized that the system information did not match the suggested diagnosis and invested more time in checking the system's raw data.

A high portion of commission errors were made despite complete automation verification in all three studies. Information sampling data and questionnaire data from study II suggest that there are three possible causes that can lead to commission errors: a) Incomplete automation verification. Attention is withdrawn from the automated task, and automation is not sufficiently verified any more. b) Complete automation verification without attentive processing of contradictory information, analogous to a "looking-but-not-seeing" effect. System information is reported that is to be expected given the diagnosis, replicating the "phantom memory phenomenon" reported by Mosier et al. (1998, 2001). c) Discounting of contradictory information. Only one participant in study II was aware of contradictory system information and followed the automation advice anyway.

6.5 Attention Allocation Strategies & Overall System Performance

Given that most participants in our studies fall into group a) or b), commission errors seem to be more of a problem of attention bias than decision bias. Operators focus on unsupported tasks and pay less attention to the automated tasks. This way they can improve performance in unsupported tasks. With limited mental resources it can be an effective coping strategy to allocate attention to

non-automated tasks and rely on the automation to deal with the tasks allocated to the automation. Considering the demands of concurrent tasks, it can lead to short-term advantages in the overall human-automation system performance to be complacent. As long as the automation works reliably, overall performance benefits.

Issues can arise when the automation fails and automation advice is wrong, or when the automation support breaks down completely and the operator has to resume manual performance. After experiencing an automation failure, participants focus more on automation verification. This reduces the risk of commission errors but comes at the expense of secondary task performance decrements. Depending on the frequency of automation failure and the costs associated with missing or delayed detection of automation failures, it can still be beneficial for the combined human-automation system performance to invest more in unsupported tasks and shift attention away from automation verification.

Inattentive processing of system information seems to be a greater problem than incomplete verification. Incomplete verification could be detected even in real-time, given that optimal information sampling can be predefined. Actual operator sampling behavior can be compared with the optimal sampling, and if sampling is incomplete, operators can be reminded to check all necessary information before accepting automation advice. However, a high portion of commission errors happened despite complete verification. Sampling all necessary system information does not guarantee that this information is attentively processed. Thus, monitoring the operator's information sampling behavior could only prevent a small portion of commission errors.

Even if the accessed parameters were not attentively processed, which was

shown by the AVQ in study II, it still seems to help to do the necessary diagnostic procedure in automation supported blocks. After automation break down and returning to manual performance, diagnosis of system faults was fast and correct. Fault identification times improved compared to the first manual block (study I and III). Even if diagnosing system faults was supported by the aid and participants only had to validate the proposed diagnosis, this was enough to stay involved in the diagnostic process and keep up or even enhance system knowledge and diagnostic capabilities. In study III, fault identification times were lower already in Block 1 compared to performance in study I. In study III, there was an additional 45-minutes test session after the 4-hours training session, that was used to examine the acquired performance skills. So participants in study III had more practice in manual diagnosis than study I participants.

6.6 Practical Conclusions

High degrees of automation can highly benefit performance when the automation works reliably. However, as an automation can fail, countermeasures should be taken in order to prevent possible negative effects which were especially noticeable with high DOA support. Lower DOA support did not bring about the same performance advantages, but performance decrements in case of automation break down were also less severe. Practicing cognitive or manual skills that are not needed with automation support on a regular basis can help prevent skill degradation. This can be done in separate training sessions in a safe surrounding like a simulator. Another way could be adaptive automation that gives tasks back to the operator when workload is low and supports the operator when workload is high.

A problem here might be determining workload in real-time.

When automation fails, failure should be salient. When the operator has other tasks to work on, he will take attention away from the automated tasks and concentrate more on non-assisted tasks. When the automation is not sure about a decision because values the decision is based on are close to a decision criterion, or system information is unclear or ambiguous, or environmental information is cluttered, it should be made clear to the operator that the automated decision should be cross-checked as failure is possible or even highly probable. However, when the automation is sure based on the information that is available for a decision, validation should not be expected. Automatic procedures that check if the operator has fulfilled his trained validation procedures will necessarily fail as operators can check information without attentively processing it.

Experience of automation failure can decrease the risk of complacency and automation bias. This should be considered in training of operators. However, it is important to also keep in mind that experiencing automation failure in training cannot fully prevent complacency and automation bias.

Automation should be built with the limitations of operators in mind, and some operators will be complacent. Operators will rely on the automation when automation is in place, and they will not always fully and attentively cross-check the automation. As complacency and automation bias cannot fully be prevented, countermeasures should be implemented in automated systems that can fail. This is especially important if automation failure is associated with high costs.

If an automation can inform the operator about how sure it is about a decision, the operator should have access to this information. In addition to a diagnosis, a level of certainty could be indicated. This way, an operator could concentrate

on automation verification and check system information to back up the automation's advice in case the automation is not certain about its decision. In case an automation decision is clear, there is no need to seek backup information, instead the operator can continue working on unsupported tasks. In the context of automated driving this has beed shown to be a promising approach. Drivers who received information about system uncertainty took over manual control faster when needed, and could perform tasks not related to driving without compromising safety (Helldin, Falkman, Riveiro, & Davidsson, 2013). Situation awareness and automation acceptance also benefit from providing uncertainty information (Beller, Heesen, & Vollrath, 2013). Also in the context of aviation, presenting levels of confidence could protect performance in case of false recommendations (McGuirl & Sarter, 2006).

When system failure is highly unlikely and failure costs are low, it may be beneficial to be complacent even if that can lead to omission and commission errors. For example, in a traffic situation during rush hour in town it may be safer to pay more attention to the traffic and blindly follow the recommendations of an navigational aid even if the aid might suggest a suboptimal route. However, if failure costs are high, as is the case in safety critical domains like aviation or medicine, complacency and automation bias and the associated errors in case of automation failure can be a serious threat to safety. This needs to be taken into account when designing automation.

6.7 Limitations

In all three studies, automation reliability was high, but compared to automation reliabilities found in real systems, it was still rather low. With the low number of events that can be realized in an experiment using a simulation such as AutoCAMS, one automation failure has a dramatic effect on reliability. It would need a lot more events to realize reliability rates that would be acceptable in industry, aviation, or medicine.

Although the simulated process control task used in the experiments is a rather complex experimental task, it is still simple compared to real-world process control tasks. Furthermore, the experiments lasted only a few hours. Operators use automation for years. While we already found effects in these short-term human-automation interaction scenarios, some performance consequences might only show in the long run, or might be more or less severe than we found in these studies.

In our studies, we had no access to real process control systems or operators of such systems. Inviting only students with a background in engineering was the best compromise. Students with no technical background or understanding proved to be unsuitable already in the pilot studies. Psychology students were often not able to diagnose system faults manually after four hours of training.

References

Akerstedt, T., Fredlund, P., Gillberg, M., & Jansson, B. (2002). A prospective study of fatal occupational accidents - relationship to sleeping difficulties and occupational factors. *Journal of Sleep Research, 11,* 69-71.

Alberdi, E., Povyakalo, A. A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology, 11,* 909-918.

Alberdi, E., Povyakalo, A. A., Strigini, L., & Ayton, P. (2009). Computer Aided Detection: Risks and benefits for radiologists' decisions. In: E. Samei & E. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques.* (pp. 320-332). Cambridge, UK: Cambridge University Press.

Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P., & Given-Wilson, R. (2008). CAD in mammography: Lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer-Assisted Radiology and Surgery, 3,* 115-122.

Aust, F., Moehlenbrink, C., & Jipp, M. (2011). Operationalization of learned carelessness. An experimental approach. In *Proceedings of the Human Factors and Ergonomics Society, 55,* 1735-1739. Santa Monica, CA: Hu-

man Factors and Ergonomics Society.

Bainbridge, L. (1983). *Ironies of automation.* Automatica, 19, 775-779.

Bagheri, N., & Jamieson, G. A. (2004a). Considering subjective trust and monitoring behavior in assessing automation-induced complacency. In *Proceedings of Human Performance, Situation Awareness and Automation Conference*, Marietta, GA.

Bagheri, N., & Jamieson, G. A. (2004b). The impact of context-related reliability on automation failure detection and scanning behavior. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics 1,* 212-217.

Bahner, J. E. (2008). *Uebersteigertes Vertrauen in Automation: Der Einfluss von Fehlererfahrungen auf Complacency und Automation Bias.* Dissertation, Technische Universitaet Berlin.

Bahner, J. E. & Manzey, D. (2004). Complacency - Begriffsklärung, Stand der Forschung und Implikationen für die Verlässlichkeit der Mensch-Maschine Interaktion. In M. Grandt (Hrsg.), *Verlässlichkeit der Mensch-Maschine Interaktion (DGLR-Bericht 2004-03)* (S. 35-48). Bonn: Deutsche Gesellschaft für Luft- und Raumfahrt e.V.

Bahner, J. E., Elepfandt, M., & Manzey, D. (2008). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52,* 1330-1334. Santa Monica, CA: Human Factors and Ergonomics Society.

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

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

Beatty, J., Ahern, S. K., & Katz, R. (1977). Sleep deprivation and the vigilance of anesthesiologists during simulated surgery. In R. R. Mackie (Ed.), *Vigilance. Theory, Operational Performance, and Physiological Correlates* (pp. 511-527). Springer US.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction. An approach using automation uncertainty. *Human Factors, 55,* 1130-1141.

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach.* Mahwah, NJ: Lawrence Erlbaum.

Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G., & Huff, E. M. (1976). *Aviation safety reporting system.* Technical Report TM-X-3445. Moffett Field, CA: NASA Ames Research Center.

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms.* Mahwah, NJ: Erlbaum.

Carmody, M. A., & Gluckman, J. P. (1993). Task-specific effects of automation and automation failure on performance, workload and situational awareness. In R.S. Jensen & D. Neumeister (Eds.), *Proceedings of the 7th International Symposium on Aviation Psychology* (pp. 167-171). Columbus, OH: Ohio State University, Department of Aviation.

Chaumet, G., Taillard, J., Sagaspe, P., Pagani, M., Dinges, D. F., Pavy-Le-Traon, A., Bareille, M.-P., Rascol, O., & Philip, P. (2009). Confinement and sleep deprivation effects on propensity to take risks. *Aviation, Space and Environmental Medicine, 80,* 73-80.

De Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human- Computer Studies, 58,* 719-735.

De Waard, D., van der Hulst, M., Hoedemaeker, M., & Brookhuis, K. A. (1999). Driver behavior in an emergency situation in the automated highway system. *Transportation Human Factors, 1,* 67-82.

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of sleep research, 4(s2)*, 4-14.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analysis of performance on a portable, simple visual RT task sustained operations. *Behavioral Research Methods, Instrumentation, and Computers, 17,* 652-655.

Domeinski, J., Wagner, R., Schoebel, M., & Manzey, D. (2007). Human redundancy in automation monitoring: Effects of social loafing and social compensation. In *Proceedings of the Human Factors and Ergonomics Society*

Annual Meeting, 51, 587-591. Santa Monica, CA: Human Factors and Ergonomics Society.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of *Human-Computer Studies, 58,* 697-718.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society Annual Meeting, 32,* 97-101. Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32-64.

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in dynamic control task. *Ergonomics, 42,* 462-492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37,* 381-394.

Fiske, S.T. & Taylor, S.E. (1991). *Social cognition (2nd ed.).* New York: McGraw-Hill.

Folkard, S., Lombardi, D. A., & Tucker, P. T. (2005). Shiftwork: safety, sleepiness and sleep. *Industrial Health, 43,* 20-23.

Funk, K., Lyall, B., Wilson, J., Vint, R., Niemczyk, M., Suroteguh, C., & Owen, G. (1999). Flight deck automation issues. *International Journal of Aviation Psychology, 9,* 109-123.

Galletta, D. F., Durcikova, A., Everard, A., & Jones, B. M. (2005) Does spell-checking software need a warning label? *Communications of the ACM, 48,* 82-86.

Gigerenzer, G., & Todd, P. A. (1999). *Simple heuristics that make us smart.* London, England: Oxford University Press.

Griefahn, B., Kuenemund, C., Broede, P., & Mehnert, P. (2001). Zur Validitaet der deutschen Uebersetzung des Morningness-Eveningness-Questionnaires von Horne und Oestberg. *Somnologie, 5,* 71-80.

Härmä, M., Sallinen, M., Ranta, R., Mutanen, P., & Müller, K. (2002). The effect of an irregular shift system on sleepiness at work in train drivers and railway traffic controllers. *Journal of sleep research, 11,* 141-151.

Harrison,Y. & Horne, J.A. (2000). The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied, 6,* 236-249.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp.139-183). Amsterdam, Netherlands: Elsevier.

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting sys-

tem uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 210-217). ACM.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology, 45,* 73-93.

Hockey, G. R. J. (2003a). Introduction: Operator functional state in the analysis of complex performance. In G.R.J. Hockey, A.W.K. Gaillard & O. Burov (Eds.), *Operator functional state: The assessment and prediction of human performance degradation in complex tasks* (pp. 3-7). Amsterdam: IOS.

Hockey, G. R. J. (2003b). Operator functional state as a framework for the assessment of performance degradation. In G.R.J. Hockey, A.W.K. Gaillard & O. Burov (Eds.), *Operator functional state: The assessment and prediction of human performance degradation in complex tasks* (pp. 8-23). Amsterdam: IOS.

Hockey, G. R. J., Wastell, D. G., & Sauer, J. (1998). Effects of sleep deprivation and user interface on complex performance: A multilevel analysis of compensatory control. *Human Factors, 40,* 233-253.

Hoddes, E., Dement, W. C. & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiology, 10,* 431-436.

Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: a review.

Occupational and environmental medicine, 56, 289-294.

Kaber, D. B., & Endsley, M. R. (2004).The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues of Ergonomics Science, 5,* 113-153.

Kahnemann, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* New York, NY: Cambridge University Press.

Kerstholt, J. H., Passenier, P. O., Houttuin, K., & Schuffel, H. (1996). The effect of a priori probability and complexity on decision making in a supervisory control task. *Human Factors, 38,* 65-80.

Killgore, W. D. S. (2007). Effects of sleep deprivation and morningness-eveningness traits on risk-taking. *Psychological Report, 100,* 613-626.

Killgore, W. D., Balkin, T. J., & Wesensten, N. J. (2006). Impaired decision making following 49 h of sleep deprivation. *Journal of Sleep Research, 15,* 7-13.

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

Lee, J., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40,* 153-184.

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

Lorenz, B., Di Nocera, F., Roettger, S., & Parasuraman, R. (2002). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine, 73,* 886-897.

Mack, A., & Rock, I. (1998). *Inattentional blindness.* Cambridge, MA: MIT Press.

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

Manzey, D. & Bahner, J. E. (2005). Vertrauen in Automation als Aspekt der Verlässlichkeit von Mensch-Maschine-Systemen. In K. Karrer, B. Gauss & C. Steffens (Hrsg.), *Beiträge zur Mensch-Maschine-Systemtechnik aus Forschung und Praxis - Festschrift für Klaus-Peter Timpe* (S. 93-109). Düsseldorf: Symposion.

Manzey, D., Bahner, J. E., & Hüper, A. D. (2006). Misuse of automated aids in process control: complacency, automation bias and possible training interventions. In *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 220-224). San Francisco, CA

Manzey, D., Bleil, M., Bahner-Heyne, J. E., Klostermann, A., Onnasch, L., Reichenbach, J., & Roettger, S. (2008). *AutoCAMS 2.0. Manual.* Available from www.aio.tu-berlin.de/?id=30492 [February 19, 2009].

Manzey, D., Reichenbach, J., & Onnasch, L. (2008). Performance consequences of automated aids in supervisory control: The impact of function allocation. In *Proceedings of the Human Factors and Ergonomics Society*

Annual Meeting, 52, 297-301. Santa Monica, CA: Human Factors and Ergonomics Society.

Manzey, D., Reichenbach, J., & Onnasch, L. (2009). Human performance consequences of automated decision aids in states of fatigue. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53,* 329-333. Santa Monica, CA: Human Factors and Ergonomics Society.

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

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48,* 656-665.

Merritt, S. M., & Illgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50,* 194-210.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors, 47,* 35-49.

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

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behavior. *International Journal of Industrial Ergonomics, 31,* 175-178.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science, 1,* 354-365.

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,* 44-58.

Mosier, K. L., Palmer, E. A., & Degani, A. (1992). Electronic checklists: Implications for decision making. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 7-11). Santa Monica, CA: Human Factors and Ergonomics Society.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman, & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 201-220). Mahwah, NJ: Lawrence Erlbaum Associates.

Mosier, K. L., Skitka, L. J., Dunbar, M., & McDonnell, L. (2001). Aircrews and automation bias: The advantages of teamwork? *International Journal of Aviation Psychology, 11,* 1-14.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision-making and performance in high-tech cockpits. *International Journal of Aviation Psychology, 8,* 47-63.

Mueller, S.T. (2007). The PEBL Manual. Available from http://pebl.sourceforge.net

Navon, D., & Gopher, D. (1979). On the economy of the human information processing system. *Psychological Review, 86,* 214-255.

Onnasch, L., Wickens, C. D., Li, H. , & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors, 56,* 476-488

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics, 43,* 931-951.

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

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

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced complacency. *International Journal of Aviation Psychology, 2,* 1-23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230-259.

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

Philip, P., & Akerstedt, T. (2006). Transport and industrial safety, how are they affected by sleepiness and sleep restriction? *Sleep medicine reviews, 10,*

347-356.

Reichenbach, J., Onnasch, L., & Manzey, D. (2010). Misuse of automation: The impact of system experience on complacency and automation bias in interaction with automated aids. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54,* 374-378. Santa Monica, CA: Human Factors and Ergonomics Society.

Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human performance consequences of automated decision aids in states of sleep loss. *Human Factors, 53,* 717-728.

Rogers, A. E., Hwang, W. T., Scott, L. D., Aiken, L. H., & Dinges, D. F. (2004). The working hours of hospital staff nurses and patient safety. *Health affairs, 23,* 202-212.

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

Samel, A., Wegmann, H. M., & Vejvoda, M. (1995). Jet lag and sleepiness in aircrew. *Journal of Sleep Research, 4(s2),* 30-36.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007) Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors, 49,* 347-357.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing.

Human Factors, 43, 573-583.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2.ed., pp. 1926-1943). New York, NY: Wiley.

Sauer, J., Hockey, G. R. J., & Wastell, D. G. (2000). Effects of training on short- and long-term skill retention in a complex multiple-task environment. *Ergonomics, 43,* 2043-2064.

Sauer, J., Kao, C. S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics, 55,* 840-853.

Sauer, J., Kao, C. S., Wastell, D., & Nickel, P. (2011). Explicit control of adaptive automation under different levels of environmental stress. *Ergonomics, 54,* 755-766.

Sauer, J., Nickel P., & Wastell, D. (2013). Designing automation for complex work environments under different levels of stress. *Applied Ergonomics, 44,* 119-127.

Sauer, J., Wastell, D., Hockey, G. R. J., & Earle, F. (2003). Performance in a complex multiple-task environment during laboratory-based simulation of occasional night work. *Human Factors, 45,* 657-669.

Schriver, A. T., Morrow, D. G., Wickens, C. D., & Talleur, D. A. (2008): Expertise differences in attentional strategies related to pilot decision making. *Human Factors, 50,* 864-878.

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics, 38,* 608-625.

Sheridan, T. B. (1988). Trustworthiness of command and control systems. In *Proceedings of the International Federation of Automatic Control Symposium on Man-Machine Systems* (pp. 427-431). Elmsford, NY: Pergamon.

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments, 1,* 120-126.

Sheridan, T. B., & Parasuraman, R. (2000). Human vs. automation in responding to failures: An expected-value analysis. *Human Factors, 42,* 403-407.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Man-Machine Systems Lab Report). Cambridge: Massachusetts Institute of Technology.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993a). Automation-induced "complacency": Development of the complacency potential rating scale. *International Journal of Aviation Psychology, 3,* 111-122.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993b). Individual differences in monitoring failures of automation. *Journal of General Psychology, 120,* 357-373.

Singh, I. L., Molloy, R., & Parasuraman, R. (1997). Automation-related monitoring inefficiency: The role of display location. *International Journal of Human-Computer Studies, 46,* 17-30.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51,* 991-1006.

Skitka, L. J., Mosier, K. L. & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52,* 701-717.

Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: Are crews better than individuals? *International Journal of Aviation Psychology, 10,* 85-97.

Smith, A. G. (2012). *Level of automation effects on situation awareness and functional specificity in automation reliance* (Master Thesis). University of Toronto.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The big five factors of personality. *Journal of Experimental Psychology: Applied, 17,* 71-96.

Taplin, S. H., Rutter, C. M., & Lehman, C. D. (2006) Testing the effect of computer-assisted detection on interpretive performance in screening mammography. *American Journal of Roentgenology, 187,* 1475-1482.

Thackray, R. I., & Touchstone, R. M. (1989). Detection efficiency on an air traffic control monitoring task with and without computer aiding. *Aviation, Space, and Environmental Medicine, 60,* 744-748.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Wandke, H. (2005). Assistance in human-machine interaction: a conceptual framework and a proposal for a taxonomy. *Theoretical Issues in Ergonomics Science, 6,* 129-155.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54,* 389-393. Santa Monica, CA: Human Factors and Ergonomics Society.

Wiener, E. L. (1981). Complacency: Is the term useful for air safety? In *Proceedings of the 26th Corporate Aviation Safety Seminar* (pp. 116-125). Denver, CO: Flight Safety Foundation, Inc.

Wilkinson, R. T., & Houghton, D. (1982). Field test of arousal: A portable reaction timer with data storage. *Human Factors, 24,* 487-493.

Womack, S. D., Hook, J. N., Reyna, S. H., & Ramos, M. (2013). Sleep loss and risk-taking behavior: A review of the literature. *Behavioral sleep medicine, 11,* 343-359.

Zheng, B., Swensson, R. G., Golla, S., Hakim, C. M., Shah, R., Wallace, L., & Gur, D. (2004) Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments. *Academic radiology, 11,* 398-406.

Appendix A

The 10-Level Model of Human-Automation Interaction

Table A.1: The 10-Level Model of Human-Automation Interaction, adapted from Sheridan & Verplank, 1978, pp. 8-17 – 8-19 (Levels of Automation in Man-Computer Decision-Making)

1	human does the whole job up to the point of turning over to the computer to implement
2	computer helps by determining the options
3	computer helps determine options and suggests one, which human need not follow
4	computer selects action and human may or may not do it
5	computer selects action and implements it if human approves
6	computer selects action, informs human in plenty of time to stop it
7	computer does the whole job and necessarily tell human what it did
8	computer does whole job and tells human what it did only if human explicitly asks
9	computer does whole job and tells human what it did and it, the computer, decides he should be told
10	computer does whole job if it decides it should be done, and if so tell human, it it decides he should be told

Appendix B

Models of Complacency and Automation Bias

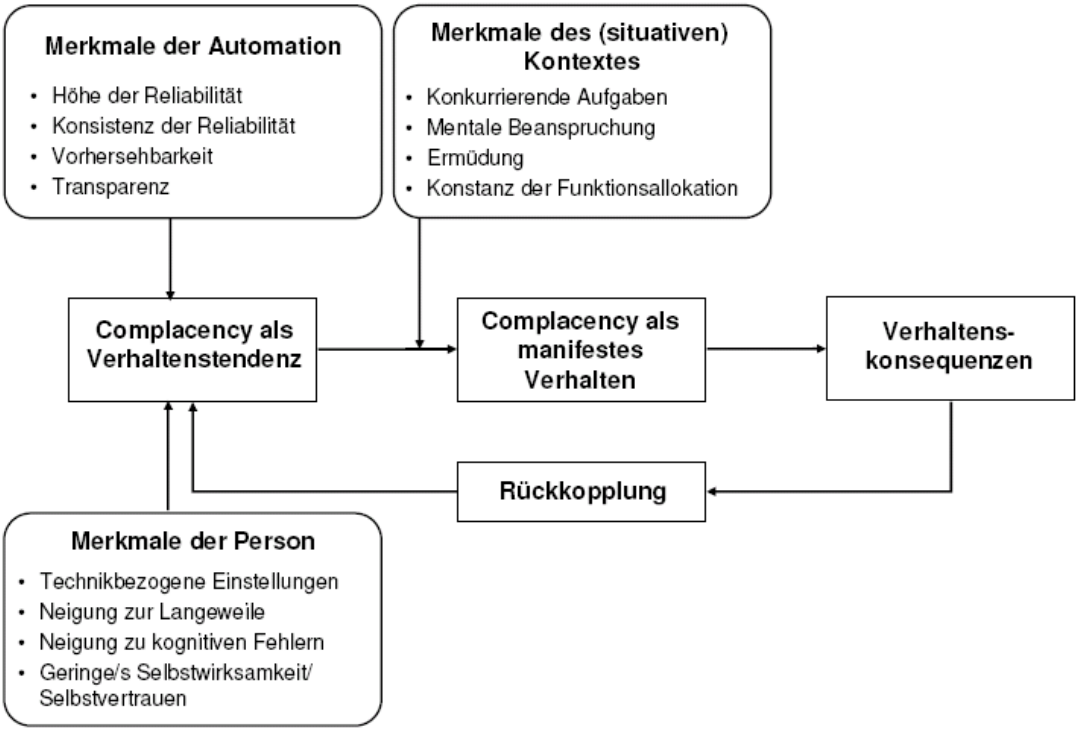

Figure B.1: Model of Complacency (Manzey & Bahner, 2005, p. 103), reproduced from Bahner (2008, p. 36)

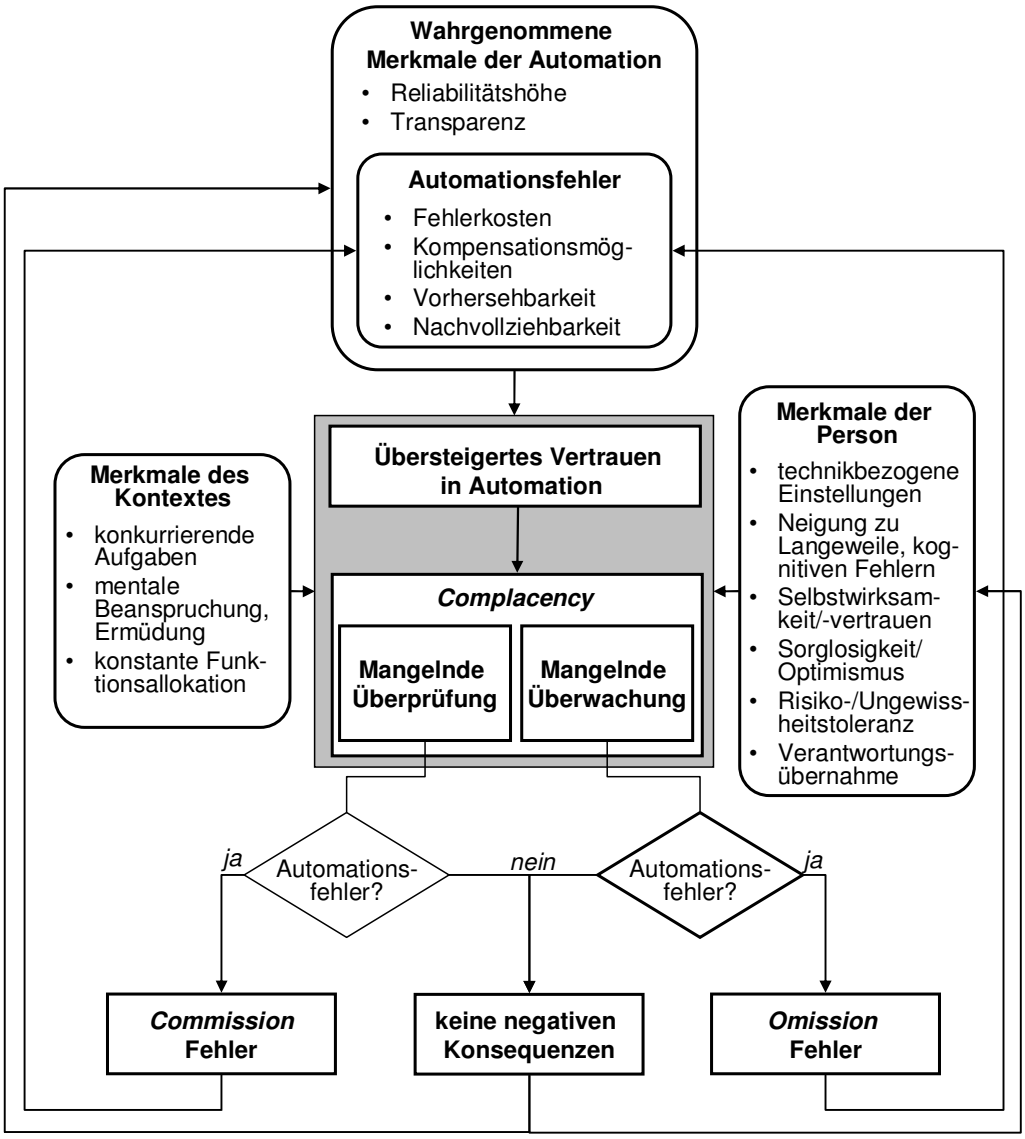

Figure B.2: Integration of Complacency and Automation Bias (Bahner, 2008, p. 52)

Appendix C

Material Study I

The material can be found on the enclosed CD.

C.1 Training Material Day 1

Training Day 1 (power point presentation)

Handout Necessary Parameters

Handout Flow Chart and Control Panel

Questions for Oral Repetition of Diagnosis

Group Proficiency Test

Individual Proficiency Test

C.2 Training Material Day 2

Training Day 2 for Manual Group (power point presentation)

Training Day 2 for Information Analysis Support Group (power point presentation)

Training Day 2 for Action Selection Support Group (power point presentation)

Training Day 2 for Action Implementation Support Group (power point presentation)

Questions for Oral Repetition of Diagnosis

C.3 Distribution and Timing of System Faults During Training

Table C.1: Training Study I. Distribution and Timing of System Faults

Block	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6
Training Day 1	Defective Sensor O_2 98s	Valve Blockage O_2 466s	Defective Sensor N_2 845s	Valve Stuck Open O_2 1198s	Valve Leak N_2 1602s	Valve Blockage Mixer 1988s

Appendix D

Material Study II

The material can be found on the enclosed CD.

D.1 Training Material Day 1

Training Day 1 (power point presentation)

Handout Necessary Parameters

Handout Flow Chart and Control Panel

Questions for Oral Repetition of Diagnosis

Group Proficiency Test

Individual Proficiency Test

D.2 Proficiency Test Day 2

Questions for Oral Repetition of Diagnosis

Proficiency Test AutoCAMS (power point presentation)

Proficiency Test Decision Tree

D.3 Material Day 3

Training Day 3 for Groups 1 - 4 (power point presentation)

Questions for Oral Repetition of Diagnosis

Automation Verification Questionnaire

Reliability Questionnaire

D.4 Distribution and Timing of System Faults During Training and Proficiency Test

Table D.1: Training Study I. Distribution and Timing of System Faults

Block	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6	Fault 7
Training Day 1	Valve Stuck Open O_2 104s	Valve Blockage N_2 476s	Valve Leak O_2 878s	Valve Blockage O_2 1221s	Valve Leak N_2 1614s	Defective Sensor O_2 1999s	Valve Blockage Mixer 2398s
Training Day 2: Proficiency Test	Valve Leak O_2 98s	Defective Sensor O_2 466s	Valve Blockage N_2 845s	Valve Stuck Open O_2 1198s	Valve Blockage Mixer 1602s	Valve Leak N_2 1988s	Valve Blockage O_2 2378s

Appendix E

Material Study III

The material can be found on the enclosed CD.

E.1 Training Material Day 1

Training Day 1 (power point presentation)

Handout Necessary Parameters

Handout Flow Chart and Control Panel

Questions for Oral Repetition of Diagnosis

Group Proficiency Test

Individual Proficiency Test

E.2 Proficiency Test Day 2

Questions for Oral Repetition of Diagnosis

Proficiency Test AutoCAMS (power point presentation)

Proficiency Test Decision Tree

E.3 Material Day 3

Training Day 3 for Information Analysis Support Group (power point presentation)

Training Day 3 for Action Implementation Support Group (power point presentation)

Questions for Oral Repetition of Diagnosis

E.4 Material Day 4

Training Day 4 for Information Analysis Support Group (power point presentation)

Training Day 4 for Action Implementation Support Group (power point presentation)

Questions for Oral Repetition of Diagnosis

E.5 Distribution and Timing of System Faults During Training and Proficiency Test

Table E.1: Training Study I. Distribution and Timing of System Faults

Block	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6	Fault 7
Training Day 1	Valve Stuck Open O_2 104s	Valve Blockage N_2 476s	Valve Leak O_2 878s	Valve Blockage O_2 1221s	Valve Leak N_2 1614s	Defective Sensor O_2 1999s	Valve Blockage Mixer 2398s
Training Day 2: Proficiency Test	Valve Leak O_2 98s	Defective Sensor O_2 466s	Valve Blockage N_2 845s	Valve Stuck Open O_2 1198s	Valve Blockage Mixer 1602s	Valve Leak N_2 1988s	Valve Blockage O_2 2378s