SPECIAL ISSUE: Innovations in Cognitive Engineering and Decision Making, Part I

Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience

Dietrich Manzey
Juliane Reichenbach
and Linda Onnasch
Berlin Institute of Technology

Abstract: Two experiments are reported that investigate to what extent performance consequences of automated aids are dependent on the distribution of functions between human and automation and on the experience an operator has with an aid. In the first experiment, performance consequences of three automated aids for the support of a supervisory control task were compared. Aids differed in degree of automation (DOA). Compared with a manual control condition, primary and secondary task performance improved and subjective workload decreased with automation support, with effects dependent on DOA. Performance costs include return-to-manual performance issues that emerged for the most highly automated aid and effects of complacency and automation bias, respectively, which emerged independent of DOA. The second experiment specifically addresses how automation bias develops over time and how this development is affected by prior experience with the system. Results show that automation failures entail stronger effects than positive experience (reliably working aid). Furthermore, results suggest that commission errors in interaction with automated aids can depend on three sorts of automation bias effects: (a) withdrawal of attention in terms of incomplete cross-checking of information, (b) active discounting of contradictory system information, and (c) inattentive processing of contradictory information analog to a "looking-but-not-seeing" effect.

Keywords: automation, automation bias, level of automation

ADDRESS CORRESPONDENCE TO: Dietrich Manzey, Berlin Institute of Technology, Department of Psychology and Ergonomics, Chair Work, Engineering and Organizational Psychology, Marchstrasse 12, Sekr. F 7, D-10587 Berlin, Germany, dietrich.manzey@tu-berlin.de

Journal of Cognitive Engineering and Decision Making, Volume 6, Number 1, March 2012, pp. 57-87.
DOI: 10.1177/1555343411433844.

Introduction

In many complex work environments, human operators are supported by an increasing number of automated systems. In particular, systems that support decision making, such as navigation systems in cars, image-based assistance systems for surgery, or diagnostic aids for process control, have gained much attention lately. In providing human operators with such aids, system designers usually intend to improve the overall system's reliability and performance as well as to reduce the workload of the user while he or she is performing the supported task. However, the benefits of automation may not always be realized but can be offset by some unwanted performance consequences resulting from an inappropriate use of the systems. These performance consequences include overreliance on automation, loss of situation awareness, and possible loss of skills needed to perform the automated functions manually in case of automation failure (Endsley & Kiris, 1995; Parasuraman, Sheridan & Wickens, 2000).

Framework Models of Human-Automation Interaction

To guide the research on automation-induced performance consequences, different framework models have been proposed that allow for a standardized characterization of automated systems with respect to how functions are distributed between humans and machine (e.g., Endsley & Kaber, 1999; Endsley & Kiris, 1995; Parasuraman et al., 2000; Sheridan, 2000). Common to all of these models is the basic assumption that automation is not an all-or-none phenomenon and that the performance consequences in terms of benefits and costs of automation will directly depend on which and how many functions are automated. One of the currently most recognized models in this respect is the types and levels taxonomy of automation proposed by Parasuraman et al. (2000). This model distinguishes automated systems with respect to two aspects. The first aspect involves the stages of human information processing that are supported by a given automated system. Four successive stages are distinguished, which are referred to as information acquisition, information analysis, decision making and response selection, and action execution. The second aspect regards to what extent information-processing functions within each of these stages are supported, that is, how much the human is still kept involved in the given function. Closely related to this taxonomy is the distinction of different degrees of automation (DOA) introduced by Wickens, Li, Santamaria, Sebok, and Sarter (2010). According to this concept, "higher degrees of automation can be accomplished by both higher levels within a stage and by including later stages" (Wickens et al., 2010, p. 389). For example, an alarm system informing human operators about the presence of a specific critical situation would be considered to represent a higher DOA than would a simple master alarm (both supporting information analysis functions on different levels). However, the same system would represent a lesser DOA compared with a system that also provides advice for proper responses (support of decision making and response selection) in addition to the alerting function. With respect to human performance consequences,

it is assumed that intended benefits of automation in terms of better performance and workload reduction directly increase with DOA. In contrast, medium DOA, which keeps the human in the loop at least to some extent, has been assumed to provide the best choice for realizing benefits from automation and, at the same time, preventing what has been referred to as out-of-the-loop unfamiliarity, namely, a loss of situation awareness and a loss of manual skills, which may lead to performance problems if the operator has to resume manual control after an automation breakdown (Endsley & Kiris, 1995). Thus far, only few studies have addressed the impact of DOA on human performance systematically, with somewhat mixed results. Whereas some studies have provided evidence for the benefits of medium DOA for maintaining situation awareness and skills (e.g., Endsley & Kaber, 1999; Endsley & Kiris, 1995; Kaber & Endsley, 2004), other findings suggest that higher DOA may provide benefits in this respect (Lorenz, Di Nocera, Roettger, & Parasuraman, 2002).

Misuse of Automation: Complacency and Automation Bias

One particular factor that has not been addressed in the studies already referred to includes misuse of automation, that is, an uncritical reliance on the proper functioning of an automated system without recognizing its limitations and the possibilities of automation failures (Parasuraman & Riley, 1997). In typical supervisory control tasks, misuse of automation is reflected in an insufficient monitoring and checking of automated functions, a phenomenon that also has been referred to as "automation-induced complacency" (Moray & Inagaki; 2000, Parasuraman, Molloy, & Singh, 1993). Possible performance consequences of complacency are suggested to include a loss of situation awareness, difficulties in returning to manual performance in case of automation failures, and an elevated risk that operators fail to detect and manage automation failures in due time (Endsley & Kiris, 1995; Parasuraman et al., 1993).

A distinct but related aspect of misuse of automation has been described for human interactions with automated decision aids. According to Mosier and Skitka (1996), the availability of automated aids can lead the user to make decisions that are not based on a thorough analysis of all available information but that are strongly biased by the automatically generated advice, a phenomenon that they have referred to as "automation bias." One possible performance consequence of this bias involves commission errors, which occur when operators follow a recommendation of an automated aid even though it is wrong. As has been described by Skitka, Mosier, and Burdick (1999), these errors "can be the result of not seeking out confirmatory or disconfirmatory information, or discounting other sources of information in the presence of computer-generated cues" (p. 993). This description suggests that at least two sorts of automation bias should be distinguished. The first one reflects a decision bias in a strict sense. Having obtained contradictory information from different sources, the operator decides for some reason to trust the provisions of the automated aid. However, the second one, that is, following the aid's recommendation without verification, seems to reflect a kind of decision bias that directly corresponds to complacency

in automation monitoring. Similar to complacent operators who do not monitor an automated process sufficiently, users of an automated aid misuse the automation to the extent that they directly follow the automatically generated advice without cross-checking its validity against other available and accessible information.

Direct empirical evidence for this supposed theoretical link between complacency and automation bias has been provided by a recent series of studies (Bahner, Elepfandt, & Manzey, 2008; Bahner, Hueper, & Manzey, 2008). In these studies, participants were required to monitor an autonomously running life-support system and to intervene whenever this system failed. This fault identification and management task was supported by an automated aid, which alerted the participants of a system failure and provided an automatically generated diagnosis together with suggestions for appropriate steps of error management. To explore the supposed link between complacency and automation bias, it was assessed whether participants fully verified the automatically generated diagnosis before accepting it and how this would relate to the occurrence of commission errors in case the aid provided an incorrect recommendation. Between 21% (Bahner, Hueper, et al., 2008) and 75% (Bahner, Elepfandt, et al., 2008) of the participants committed a commission error when the aid provided a diagnosis that in fact was wrong. More detailed analyses showed that these operators generally tended to rely on the diagnoses of the aid without verifying it appropriately with other available information. Moreover, this tendency was stronger for participants who never had the practical experience before (e.g., during training) that the aid could fail. In contrast, only a minority of participants were found to commit this error if they had checked all necessary information to verify the aid's recommendation before. This finding suggests that the majority of commission errors observed in these studies reflected not the effect of a classical decision bias—that is, an inappropriate weighting of different information—but the effect of a bias of information processing characterized by an automation-induced withdrawal or reallocation of attention, which resembles the complacency effect in automation monitoring (cf. Parasuraman & Manzey, 2010).

Most interesting in the current context are recent studies that suggest that automation bias effects in interaction with a decision aid are affected by the DOA of an aid (Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001). For example, Sarter and Schroeder (2001) examined the performance of pilots interacting with automated decision aids that supported decision making in case of in-flight icing events. Two types of decision aids were compared. The first involved an aid that provided information about the specific icing condition (i.e., wing icing vs. tailplane icing) but left the selection of a proper action with the pilots. Thus, the support remained limited to the stage of information analysis (Parasuraman et al., 2000). The second aid provided not only information about the icing condition but also recommendations for decision making and the selection of appropriate responses. Compared with a baseline condition in which pilots had to manage in-flight icing encounters without any automation support, the availability of the aids increased the number of correct decisions and responses

considerably. However, this performance benefit was observed only when the aids provided correct recommendations. In case of inaccurate information, the aids resulted in performance decrements compared with baseline conditions. This impairment of performance was mainly related to the pilots' inadvertently following the aids' recommendation even though the available kinesthetic cues contradicted it. Moreover, a significant interaction effect was found between this indication of automation bias and the type and accuracy of the decision aid. Whereas both aids led to worse performance when the information provided was inaccurate compared with an accurate condition, this effect was stronger for the more highly automated aid. Similar results were also reported by Rovira et al. (2007). They explored to what extent automated aids differing in DOA (information automation vs. three levels of decision automation) affected the speed and quality of command and control decisions. As expected, the availability of all kinds of automated aids improved performance when they provided accurate advice. However, in case of inaccurate recommendations, clear performance costs in terms of decreased decision accuracy were identified compared with an unsupported (manual) control condition. Furthermore, evidence was found that the DOA moderated these effects. Decrements of decision accuracy in case of inaccurate automation advice were most pronounced if the aid provided a high level of support of decision-making functions (i.e., provided a specific recommendation for an optimum decision).

However, other results suggest that higher DOA actually may also lead to less reliant behavior in interaction with automated aids. Lorenz et al. (2002) investigated DOA effects of automated aids providing support for fault identification and management in a simulated process control task. More specifically, they contrasted an automated aid that provided automatically generated diagnoses for a given fault combined with recommendations for appropriate actions with a system that additionally performed all necessary actions autonomously if not vetoed by the operator. The latter kind of aid was associated with better return-to-manual performance after a complete failure of the automation. This effect seemed to be related to the fact that participants working with the most highly automated aid spent more time cross-checking the proposed diagnoses, that is, were less reliant on the automated processes, to maintain their own assessment of the situation. However, obvious differences in time pressure between the different DOA conditions, given the provision for a fixed time interval for a veto with the highest DOA, as well as strong practice effects, make a clear-cut interpretation of these results difficult.

Current Research

The experiments presented in this article extend this line of research. The first experiment addressed an evaluation of human performance consequences of automated decision aids dependent on their DOA. With the use of essentially the same task as in the studies by Lorenz et al. (2002) and Bahner, Hueper, et al. (2008), the model used for this research included a simulated supervisory control task that was supported by different automated aids for fault identification

and management. Three kinds of aids were compared that differed in how much the human operator was kept actively involved in the overall process of fault management by providing support for different stages of information processing. The first aid (information analysis [IA] support) provided an automatically generated diagnosis for a given system fault but left it to the operator to plan and implement all necessary actions. The second aid (action selection [AS] support) provided additional recommendations for necessary actions, which had then to be implemented manually by the operator. The third aid (action implementation [AI] support) performed the whole process of fault management autonomously if the operator confirmed the proposed diagnosis and plan of interventions. In addition, participants in a manual control condition had to perform the entire fault identification and management without any automation support. The evaluation of DOA effects on performance included an evaluation of the intended positive effects on primary task performance and workload as well as an evaluation of possible negative performance consequences in the event of an automation failure. With respect to the latter, a particular focus was laid on automation bias effects. However, also issues of return-to-manual performance in case of automation failure were assessed. It was expected (a) that providing automated decision support would lead to performance benefits compared with manual performance and (2) that more highly automated aids would show greater performance improvements than less automated aids. With respect to negative performance consequences, it was supposed (c) that higher compared to lower DOAs would lead to stronger automation bias effects in terms of less careful automation monitoring and a resulting higher number of commission errors as well as increased difficulties of return-to-manual performance. Furthermore it was supposed (d) that the strength of complacency-like automation bias effects would depend on the effort needed to invest for automation verification.

The second experiment capitalized on the results of the first one and addressed issues of automation bias in interaction with the aids in more detail. Specifically, the dynamic interactive development of trust and automation bias effects in interaction with a decision aid was analyzed. Furthermore, the impact of system experience, that is, whether the operator had the practical experience that the aid may fail, and the specific origins of automation bias were investigated. On the basis of earlier research (e.g., Bahner, Hueper, et al., 2008; de Vries, Midden, & Bouwhuis, 2003; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & Moray, 1992, 1994; Merritt & Illgen, 2008), we expected that trust and automation bias in human-automation interaction represent adaptive phenomena that develop dynamically over time and are determined by feedback loops that are driven by the practical experience made with a certain aid.

Experiment 1

Method

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. Participants were paid €70 for completing the study.

Apparatus: AutoCAMS 2.0. A revised version of AutoCAMS (Hockey, Wastell, & Sauer, 1998; Lorenz et al., 2002) was used for the experiment (AutoCAMS 2.0; Manzey et al., 2008). This task was developed as a small-scale simulation of a typical supervisory control task of control room operators. Specifically, it simulates an autonomously running life-support system consisting of five subsystems that are critical to maintain atmospheric conditions in a remote space capsule (i.e., oxygen, nitrogen [cabin pressure], carbon dioxide, temperature, humidity). During nominal operation, all parameters are automatically kept within a target range. However, because of 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 block of a valve, a leak of a valve, a stuck-open valve, a sensor defect, or a defect of the mixer valve.

The user interface of AutoCAMS 2.0 is shown in Figure 1. 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 1G), and a time counter starts displaying how much time has elapsed since the occurrence of the fault. 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. The repair itself takes 60 s. 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 1F). If the repair order sent is 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.

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 one of three kinds of an automated aid (Automated Fault Identification and Recovery Agent [AFIRA]; Figure 1H). In case of IA support, the master alarm is accompanied by a message providing a specific diagnosis for the given fault. However, action planning and implementation is left to the operator. In case of AS support, the diagnosis is complemented by a list of appropriate actions, which the operator has to implement. In case of the most comprehensive AI support, AFIRA does not only display a diagnosis and a listing of necessary actions but also implements all steps autonomously if confirmed by the participant.

To identify faults in the manual control condition or verify proposed diagnoses in conditions with automation support, operators have independent access to all important parameters (Figures 1A through 1C). These include relevant system parameters and a history graph for each of the five subsystems. However, this information is not always visible but has to be activated for a 10-s view by mouse click on the tank, flow meter, or history graph, respectively. Every system

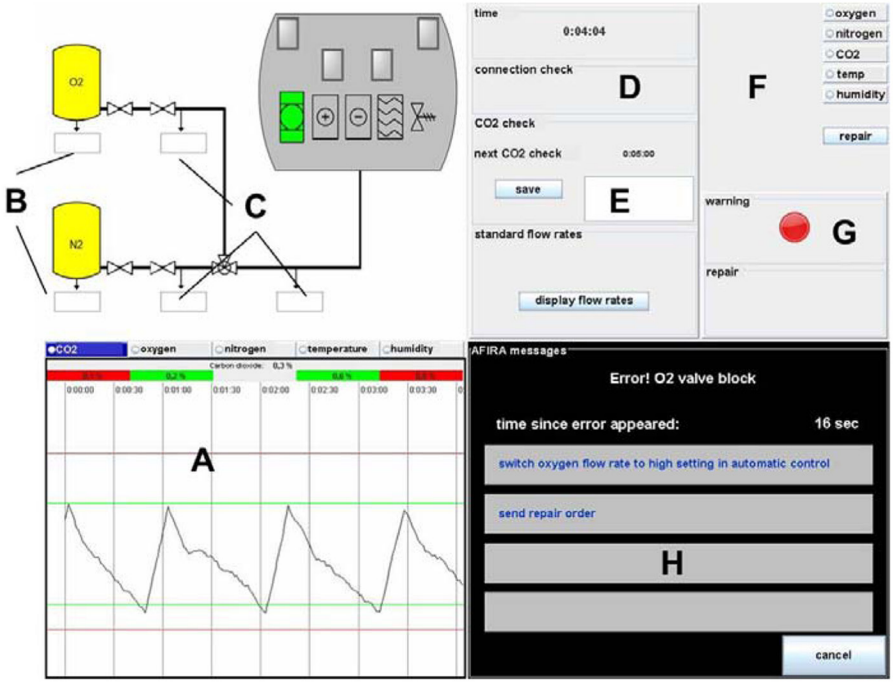

Figure 1. User interface of AutoCAMS 2.0. The figure shows the system with active action selection support. (A) history graphs, (B) tank level readings, (C) valve flow readings, (D) field where the "connection check" icon appears (secondary task), (E) field for entry of carbon dioxide readings (secondary task), (F) menu for manual control and repair orders, (G) master alarm, (H) assistance system (Automated Fault Identification and Recovery Agent).

malfunction has specific symptoms in such a way that it is possible for the operator to identify most malfunctions or to verify the diagnosis provided by AFIRA by accessing two to four specific parameters, depending on the complexity of the fault. However, identifying most complex faults unambiguously additionally requires interventions in the system.

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 s in a specific data entry field, which is provided in the AutoCAMS interface (Figure 1E). 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 1E) as fast as possible. This icon appears in random intervals roughly twice per minute.

Although AutoCAMS 2.0 represents a laboratory task, its main cognitive and multitask demands have been designed to resemble those of real supervisory

control tasks. Furthermore, performing the task requires complex system knowledge and skills, which makes the task inherently motivating for participants who have acquired these competencies during training.

Design. The study used a 4 (DOA) × 5 (block) design with DOA (manual control, IA support, AS support, AI support) defined as between-subjects factor and block defined as within-subjects factor. The five blocks per session differed with respect to whether automation support was available. During the first block, all participants worked manually, that is, 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 automated support. In Block 5, participants of all experimental groups had to return to manual performance, that is, diagnose and manage all system faults manually again without automation support. In each block, six kinds of 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. In the AFIRA groups, the six faults in each block 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. This failure of AFIRA was implemented to simulate a "first automation failure effect."

Dependent measures. Dependent measures were derived from questionnaires and from mouse-click data, which were logged during the experiment.

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 (FIT) 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 (OTE) 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 the fault management.

Secondary task performance was assessed by two measures: (a) mean response time (in milliseconds) to the appearance of the "communication link" icon and (b) prospective memory performance, that is, proportion of entries of carbon dioxide records that were provided within the correct time interval (i.e., full minute ±5 s). Only performance during periods when a participant had to deal with a system fault was considered.

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), (b) automation verification information sampling of relevant system parameters (AVIS-R), and (c) automation verification

information sampling of necessary system parameters (AVIS-N). AVT was defined as the time interval (in seconds) from the appearance of the master warning until sending of a first repair order, regardless of whether this order was correct. AVIS-R was defined as the proportion of all system parameters accessed that, in principal, were considered useful (relevant) to cross-check the automatically generated diagnosis for a given malfunction. AVIS-N was defined as the proportion of all system parameters accessed immediately necessary to cross-check a given diagnosis unambiguously. Note that necessary parameters represent 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). The number of necessary parameters that were immediately needed to verify a given diagnosis unambiguously varied as a function of the complexity of a given system failure and included two parameters (lowest complexity), three to four parameters (medium complexity), or two parameters combined with two additional active interventions in the system needed to disambiguate two possible diagnoses (highest complexity).The number of parameters actually accessed (plus interventions correctly performed when needed) was then related to this normative model. Only parameters accessed between the occurrence of the master warning and the sending of 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, Hueper, et al. (2008).

Performance consequences of a possible automation bias in terms of commission error were analyzed by the 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 with that in Block 5 on the basis of primary task performance measures as defined previously.

Procedure. The study comprised two 4.5-hr sessions, which took place 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 identify and manage all possible system faults that could occur either in the oxygen or nitrogen subsystem. The training was concluded by a test (questionnaire) probing the knowledge of participants about procedures for the identification and management of the different faults. All participants passed this test successfully and were accepted for the experiment. Before starting the experiment on the 2nd day, participants were randomly assigned to one of the four experimental groups. Participants of the automation-support 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 would be high but not perfect and were warned explicitly to check the proposed diagnoses before initiating a repair order.

Participants of the manual control group performed the same practice trials with AutoCAMS 2.0 but without any automation support. Thereafter, the experiment started. Each of the five blocks lasted 40 min. Blocks were separated by short breaks, during which subjective ratings of workload were collected.

Results

Primary task performance. Primary task performance measures were analyzed by a 4 (DOA) × 5 (block) ANOVA. Percentage of correct fault identifications varied across blocks, $F(4, 208) = 25.09$, $p < .01$, and this effect was further moderated by a significant DOA × Block interaction, $F(12, 208) = 1.95$, $p < .03$. General level of performance was already relatively high for all experimental groups in Block 1 (manual control, 87%; IA support, 86%; AS support, 81%; AI support, 86%). As expected, using automated aids in Blocks 2 through 4 improved performance to about 100% correct diagnoses in all automation-supported groups, whereas the manual control group showed only slight improvement across these blocks ($M = 91\%$).

Effects for FIT and OTE are displayed in Figure 2. For FIT, the DOA effect, $F(3, 52) = 3.45$, $p < .03$; the block effect, $F(4, 208) = 48.25$, $p < .01$; and the DOA × Block interaction, $F(12, 208) = 2.68$, $p < .01$, became significant. As becomes evident from Figure 2 (left panel), FIT profited considerably from the use of automated aids in Blocks 2 through 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 FIT varied across the three automation groups. The latter effect was confirmed by a separate 3 (DOA) × 3 (block) ANOVA comparing FIT for the three automation-supported groups across Blocks 2 through 4, which showed a significant main effect of DOA, $F(2, 39) = 4.98$, $p < .02$. Post hoc contrasts (Bonferroni) revealed that mean FIT was significantly shorter for the group working with the most highly automated aid (AI support, 20.9 s) than the other two automation-supported groups (IA support, 28.3 s; AS support, 28.5 s), both contrasts $p < .05$.

Essentially the same pattern of results also was revealed for OTE (see Figure 2, right panel). Because of technical problems, only data from 13 and 11 out of 14 participants in the IA and AS groups, respectively, could be included in the analysis. Participants of all experimental groups were more able to stabilize the system during states of faults in Blocks 2 through 4 than in Blocks 1 and 5, $F(4, 192) = 62.27$, $p < .01$. Similar to the FIT data reported previously, the strength of this effect was dependent on DOA, $F(3, 48) = 3.60$, $p < .05$, and showed a different trend across blocks for the different DOA conditions, reflected in a significant DOA × Block interaction, $F(12, 192) = 2.48$, $p < .01$. The smallest improvements emerged in the manual control group, probably reflecting some kind of practice effect. Among the three automation-supported groups, performance improvements developed more quickly (see right panel of Figure 2) and were higher with AI support than with one of the less automated aids. The latter effect was also indicated by a separate 3 (DOA) × 3 (block) ANOVA, which yielded a main effect of DOA, $F(2, 37) = 4.37$, $p < .03$. According to post hoc contrasts (Bonferroni),

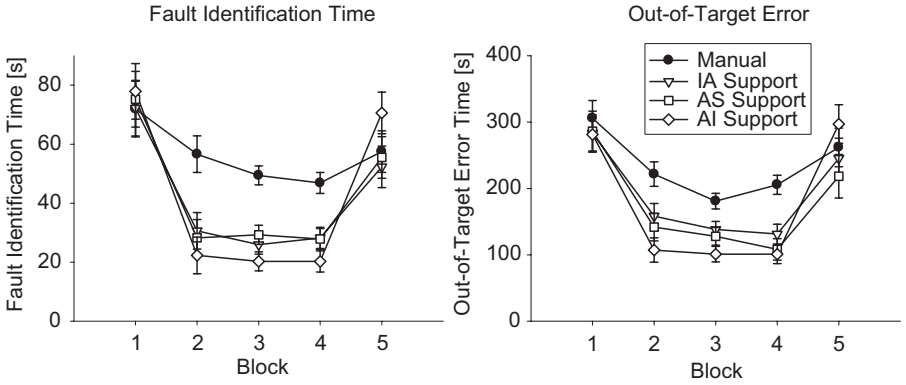

Figure 2. Primary task performance of the experimental groups across blocks. Left graph: Means and standard errors of fault identification time. Right graph: Means and standard errors of out-of-target error.

participants working with AI support were significantly better in keeping the oxygen level within the target range (mean OTE = 103.10 s) than was the IA group (142.51 s), $p < .03$, and a similar trend also emerged for the contrast with the AS group (137.59 s), $p < .07$. Taking into account that the AI aid was the only aid that also provided automation support for the control of the affected subsystem, one could expect these differences between the aids.

Secondary task performance. Performance in both secondary tasks was analyzed by a 4 (DOA) × 5 (block) ANOVA. No significant effects emerged for probe reaction times in the connection check task. However, a significant main effect of block, $F(4, 208) = 21.17$, $p < .01$, and a DOA × Block interaction, $F(12, 208) = 2.94$, $p < .01$, were found for prospective memory performance. The sources of the interaction can be derived from Figure 3 (left panel). As becomes evident, prospective memory performance improved immediately with the introduction of automation support (Blocks 2 through 4) for participants working with the most highly automated aid (AI support). In contrast, participants working with IA and AS support also improved across blocks but at a slower pace. Essentially no performance changes across blocks were found for the manual control group.

Subjective workload. Analysis of subjective workload was based on a 4 (DOA) × 5 (block) ANOVA. A significant block effect, $F(4, 208) = 24.99$, $p < .01$, was found, moderated by a significant DOA × Block interaction, $F(12, 208) = 2.33$, $p < .01$. This pattern of effects is shown in Figure 3 (right panel). All groups started at about the same level in Block 1. Although in Blocks 2 through 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 for the AI group.

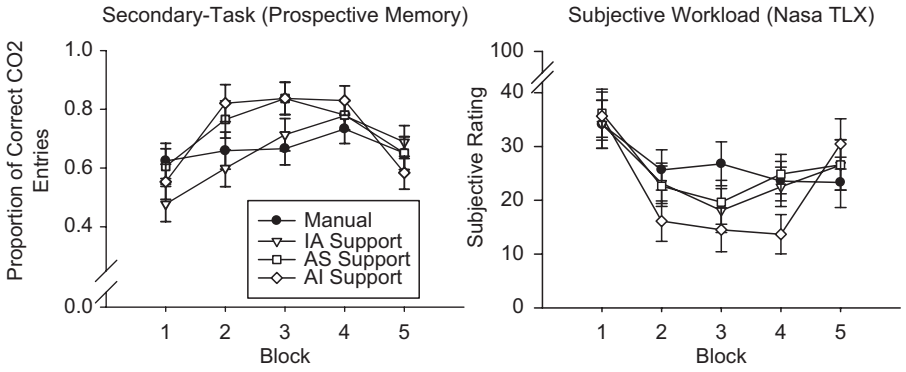

Figure 3. Secondary-task performance and perceived workload of the experimental groups across blocks. Left: Means and standard errors for the prospective memory task. Right: Means and standard errors of subjective workload ratings as assessed by NASA Task Load Index.

Return-to-manual performance. Assessment of return-to-manual performance for groups supported by the automated aid was based on a contrast of performance in Blocks 1 and 5 by a 3 (DOA) × 2 (block) ANOVA. Whereas no significant effects were found for percentage of correct diagnoses, FIT improved across blocks, $F(1, 39) = 6.24$, $p < .02$, probably reflecting effects of practice. However, some, albeit weak, indications of DOA effects on return-to-manual performance emerged for the OTE, reflecting fault management performance. Whereas 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 group supported by the highest-DOA aid. This effect was evaluated by aggregating the data of the IA and AS groups, both of which did not get any automation support for fault management actions, and contrasting it with the AI group. The means for this contrast across blocks are illustrated in Figure 4. A 2 (DOA) × 2 (block) ANOVA revealed a significant DOA × Block effect, $F(1, 38) = 4.46$, $p < .05$.

Automation verification during reliable automation support. Automation verification behavior was analyzed by a 3 (DOA) × 3 (block) ANOVA for the automation-supported groups in Blocks 2, 3, and 4. A first analysis of the time spent to verify the recommendation of the aid (AVT) revealed a significant effect of DOA, $F(2, 39) = 4.32$, $p < .02$. A post hoc analysis revealed that the group supported by the most highly automated 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 effect for block nor the interaction became significant. Further analyses focused on the extent of verification and was based on the different measures of information sampling. Sampling of system parameters that were considered to be relevant to verify a given

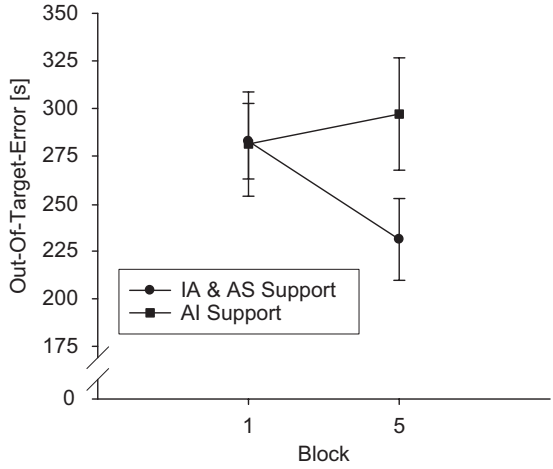

Figure 4. Effect of degree of automation on out-of-target error after return to manual performance (Block 5) compared with baseline performance in Block 1. Shown are means and standard errors for the condition with highest automation (i.e., action implementation [AI]) support in contrast to the two other automation support groups (information analysis [IA] and action selection [AS] support).

diagnosis (AVIS-R) did not vary dependent on DOA but decreased significantly across blocks, $F(2, 78) = 9.43$, $p < .01$. In contrast, sampling of system parameters immediately needed for verifying the automatically generated diagnoses of the aid (AVIS-N) remained stable on a comparatively high, albeit not perfect, level ($M = 93.5\%$) across blocks in all DOA groups. Neither the main effect of DOA nor the main effect of block or the DOA × Block interaction became significant for this measure. Obviously, participants reduced their effort of automation verification over time only by neglecting sampling of additional system parameters that were not immediately needed for a cross-check of the recommendations provided by the aid.

An additional 3 (DOA) × 3 (block) × 3 (complexity) ANOVA of the AVIS-N measure was run to explore the hypothesis that the effort spent for automation verification was affected by the complexity of the cross-check needed. The analysis revealed a significant main effect of complexity, $F(2, 78) = 7.50$, $p < .01$. Although verification was almost complete ($M = 97.6\%$) for low-complexity errors, that is, errors that could be verified by sampling just two system parameters, it decreased for medium-complexity errors (three to four parameters, $M = 93.1\%$) and even more for high-complexity errors, which also required interventions to verify the diagnosis ($M = 91.1\%$). Most interestingly, this effect was moderated by a significant Complexity × DOA interaction, $F(4, 78) = 3.92$, $p < .01$. This interaction is illustrated in Figure 5. As becomes evident, amount of verification did not differ dependent on DOA for low- and medium-complexity errors.

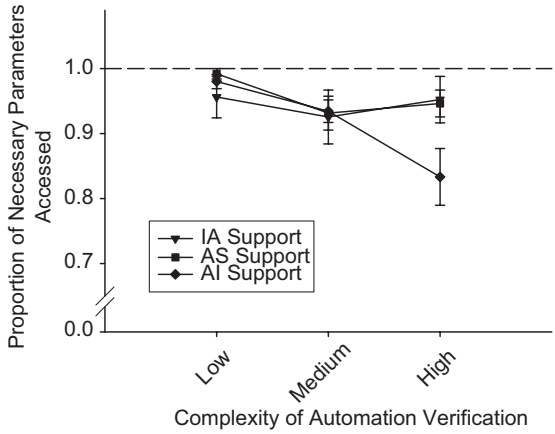

Figure 5. Extent of automation verification information sampling of necessary system parameters dependent on the complexity of automation verification procedures and degree of automation.

However, for the most complex failures, the group working with the most highly automated aid (AI support) completed significantly fewer verification steps than did the IA or the AS group. This finding was substantiated by additional one-way ANOVAs contrasting the automation verification behavior of the three DOA groups separately for the three levels of fault complexity. This analysis revealed a significant main effect only for high-complexity faults, $F(2, 39) = 3.74$, $p < .04$. When we looked at which part of the verification procedure was neglected, it became evident that most participants in the AI group sampled the necessary system parameters but tended to omit the additional control actions needed for unambiguous automation verification.

Commission errors and automation verification in case of automation failure. We found clear evidence for automation bias leading to a commission error in all automation-supported groups by analyzing fault identification performance for Fault 7 in Block 4. Up to half of the participants in the automation-supported groups followed the automatically generated diagnosis for this fault even though it was incorrect. However, no significant difference was found for the different kinds of 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 sent a correct repair order.

To investigate whether this effect was attributable 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) × 2 (wrong diagnosis detected vs. not detected) ANOVA revealed no significant effects for AVIS-N. A more detailed analysis of information sampling

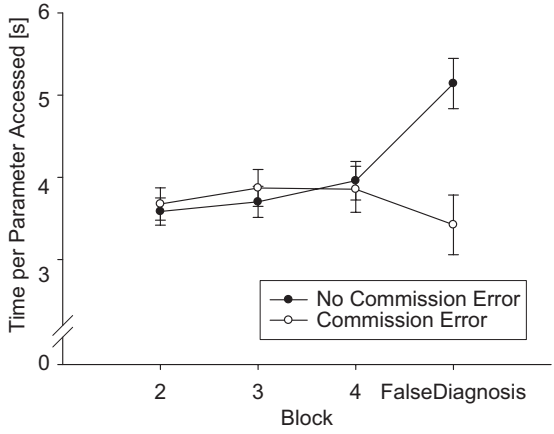

Figure 6. Time spent per system parameter accessed for automation verification. Shown are means and standard errors for the three automation support groups across the three blocks in which the Automated Fault Identification and Recovery Agent provided correct diagnoses and the first incident of an automation failure at the end of Block 4.

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, that is, they had accessed all system parameters needed to detect the contradiction between the automatically generated diagnosis and the actual system state. Only 7 participants showed a complacency-like automation bias effect reflected in an incomplete cross-check before sending the repair order. However, 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, Fault 7) × 2 (commission error: wrong diagnosis detected vs. not detected) ANOVA revealed a significant effect of block, $F(3, 117) = 4.40$, $p < .01$, moderated by a Block × Commission Error interaction, $F(3, 117) = 11.36$, $p < .01$, for time spent per parameter. Figure 6 shows this effect. With reliable automation, there was no difference between the two groups. However, when the automated decision aid provided a wrong diagnosis, participants who did not detect the wrong diagnosis spent the same amount of time per parameter as in normal operation trials, whereas participants who detected the wrong diagnosis invested more time per parameter to inspect the system. Obviously, both groups cross-checked the aid's diagnosis to a comparable degree but differed considerably with respect to the time spent for dealing with the sampled information in case it contradicted what the aid had proposed.

Discussion

Providing automation support for fault identification and management yielded the intended performance benefits. Compared with the performance in the manual control group, support by an automated aid led to reduced FIT as well as better fault management performance. In addition, operators supported by an aid reported less workload associated with their supervisory control task than did operators who did not get this kind of automation support. As expected, all of these performance benefits were found to be dependent on the level of automation. Compared with manual performance, the most pronounced and directly observable performance benefits were found with the highest-DOA aid. Participants working with AI support were faster in diagnosing a given fault and better able to keep the oxygen level within the target range, compared with the other two experimental groups. In addition, they also showed better secondary-task performance. This improvement was specifically reflected in prospective memory performance, which immediately improved for participants supported by the most highly automated aid, compared with manual performance. Comparable benefits were observed with IA and AS support but took more time to develop across blocks in which automation support was available. This latter effect might reflect the higher memory load associated with the latter kinds of support, which provided a diagnosis for a given fault but still required the user to retrieve the appropriate fault management actions from memory (IA support) or to remember how to implement the actions proposed by the aid (AS support).

At first sight, the finding that higher DOA led to greater benefits in primary- and secondary-task performance might not be very surprising. However, it is remarkable that the aid representing the highest DOA also shortened the FIT. Given the fact that the automated support for the subtask of fault identification was the same for all aids, this effect could not be expected. Obviously, participants working with the most highly automated aid accepted the suggestion of the automated aid more quickly than did those working with less automated versions. As discussed later, this effect seems to be related to the fact that operators working with the most highly automated aid appeared to spend less time in automation verification and invested less effort in cross-checking the aid's diagnosis if this cross-check required time-consuming interventions in the system. This result contrasts with findings of Lorenz et al. (2002), who, using essentially the same task, did not find effects of DOA on FIT. However, in their study, the most highly automated aid was associated with a veto function, which always provided a constant time for operators to intervene before a repair order was sent, that is, did not provide the opportunity to influence the speed of fault management by rapid confirmation.

With respect to return-to-manual performance, we found some indications for an automation-induced decline of skills when comparing manual fault identification and management performance before and after the participants had worked with automation support (Block 1 vs. Block 5). It is striking that they emerged specifically for fault management performance (OTE) in the group supported by the most highly automated aid. Whereas OTE performance of

participants working with IA and AS support improved in Block 5 compared with Block 1, participants supported by AI support did not show a comparable improvement. Similar results suggesting that return-to-manual issues increase with higher DOA have been reported by Endsley and Kiris (1995), who investigated the impact of automated decision support on a planning task. They found that decision time increased when their participants had to return to manual performance after using an automated decision support system for some time. The greatest performance loss was found for a group that was supported by a "consensual AI expert system," which resembled the AI support aid in the present study. However, in the current experiment, the skill loss observed became statistically significant only for manual fault management performance but not for the cognitive skills needed to diagnose a system fault. Obviously, the automated support for the cognitive skills involved did not lessen the benefits of practice, which were comparable to those in the manual control group. This finding suggests that participants were well able to maintain and further develop their system knowledge even when using the diagnostic aid. In contrast, automated support of system stabilization as provided by the AI aid tended to adversely affect the development of appropriate skills to control and stabilize the system manually. This finding constitutes an important difference to both the IA and the AS support, which left the planning and implementation (IA) or at least the implementation of actions (AS) to the human. Both of these kinds of support led to practice effects in manual system control across blocks that were comparable to those in the manual control group. This finding provides an interesting correspondence to results from Lorenz et al. (2002), who, using essentially the same task, reported comparable effects for return-to-manual performance dependent on DOA. This finding suggests that medium levels of automation specifically provide advantages for return-to-manual performance if manual skills need to be maintained. A similar conclusion was also drawn from Endsley and Kaber (1999), who found return-to-manual issues particularly for automated aids that supported the implementation of actions in a simulated dynamic control task.

However, the main focus of the present research with respect to possible performance costs of automation was on automation bias effects. On first sight, automation bias effects in terms of insufficient automation verification found in the present study were weaker than expected. Although participants of all experimental groups reduced the extent of automation verification over time, supporting earlier results of Bahner, Hueper, et al. (2008), this effect remained limited to the checking of what has been referred to as "relevant" system parameters. When looking at the stricter variable, that is, system parameters that were immediately needed to unambiguously verify a given diagnosis, information sampling stayed at a constantly high, albeit not perfect, level for all blocks in all three DOA groups. On average, participants checked somewhat more than 90% of these necessary parameters before sending a repair order. This decrease in checking relevant parameters while keeping a cross-check of necessary parameters at a high level can be seen as optimization of information sampling. Only those parameters were constantly sampled that were absolutely essential for verifying a

given diagnosis, whereas other information that was useful but not necessary became more and more neglected over time. That is, participants learned to sample only the information that was essential for verification. With this strategy, the demands for verifying an automated diagnosis could be reduced without damage to the diagnostic performance. However, a more detailed analysis revealed that this strategy of automation verification was moderated by DOA as well as the effort needed for cross-checking. Although verification of necessary system parameters was almost complete for system faults of low and medium complexity, it decreased for high-complexity errors, which required a complex cross-checking procedure including not only assessing system parameters but also analyzing the effects of control actions. This kind of automation bias, which resembles what has been referred to as complacency in the context of supervisory control, was particularly observed in the group working with the most highly automated aid (AI support). When looking at which part of the verification was not completed, we noticed that participants supported by the most highly automated aid primarily omitted necessary control actions. This finding is especially interesting against the background that the AI group was the only group that was supported for the fault management implementation and never had to implement control actions after sending a repair order. This behavior seems to generalize to their verification behavior during the diagnostic phase before sending a repair order.

From this analysis of automation verification behavior, it might be concluded that complacency-like issues in interaction with automated aids remain limited to very specific circumstances. However, the results suggesting this conclusion are qualified by another finding of the present study that suggests that just looking at system information to cross-check an automatically generated diagnosis does not prevent the occurrence of commission errors. Up to half of the participants working with the automated aid committed a commission error when the automated aid generated a false diagnosis for the first time. This effect emerged independent of DOA and independent of how many system parameters had been checked to verify a given diagnosis. A comparison of information sampling behavior between participants who did and did not commit a commission error actually revealed only few differences with respect to the number of system parameters accessed.

Out of the 18 participants who followed the wrong recommendation, only 7 could be called complacent in the sense that they had not verified AFIRA completely before sending a repair order. The other 11 participants had checked all the necessary information before decision making and still followed the aid's incorrect recommendation. However, a significant difference between both subgroups was found in the time spent per system parameter to evaluate it. Participants who correctly recognized that the aid's advice was wrong showed a sharp increase in the average time needed to process a sampled system parameter in case its information did not fit the diagnosis of the aid, compared with trials in which the aid's diagnoses were correct. Obviously, these participants became aware of the incongruity of the system parameters and the aid's diagnosis and invested more time to evaluate the system data. This finding is in line with

Schriver, Morrow, Wickens, and Talleur (2008), who found that experts allocate more attention to failure-relevant cues when a failure was present and that more attention allocated to failure-relevant cues was associated with higher decision accuracy. In contrast, participants who did not detect the wrong diagnosis, although they had sampled some or all of the contradictory system parameters, did not show any difference in the time spent for evaluating the system's raw data whether the aid's diagnosis was correct or wrong. Given this finding, the commission error committed by these participants does not seem to be related to a decision bias in terms of discounting of contradictory information. Similarly, an explanation in terms of confirmation bias in processing the sampled information or difficulties of comprehension is also unlikely, given that the participants were very well trained and that the system parameter always contradicted the aid's diagnosis in an unequivocal way. Rather, the commission errors in this group seem to be attributable to a sort of "looking-but-not-seeing" effect, analogous to what in other contexts has been referred to as "inattentional blindness" (Mack & Rock, 1998), a phenomenon whereby attention and eye movements are dissociated and information in the environment can be missed even if fixated. This finding is particularly interesting. It provides evidence that the commission errors associated with complete (optimal) information sampling found in the present study were not owed to a misweighting of contradictory information. Instead, it seems that they were related to a more subtle sort of automation-induced bias in information processing that was reflected not in an obvious neglect of automation verification but in a withdrawal of attentional resources from processing the available (and looked-at) system data. That is, even the participants who did not detect the aid's failure obviously continued to cross-check the aid's recommendation as instructed in the training but did not invest attention in this task anymore. Such effect would fit an earlier suggestion of Duley, Westerman, Molloy, and Parasuraman (1997), who also found a similar effect in a supervisory control task supported by an automated aid. It further would fit recent results of Sarter, Murmaw, and Wickens (2007). In their study, pilots in a simulator were found to look at the mode indicator of their flight management system but nevertheless committed a mode error, which suggested that they did not see what they had looked at. However, the data of the present experiment are not fully conclusive for supporting this conclusion. Thus, a second experiment was run to explore this effect in more detail. In addition, the second experiment was used to investigate how the different effects of automation bias develop over time, to what extent these effects are related to the operator's trust in the system, and how this dynamic development is affected by the practical experience an operator has made with a given system.

Experiment 2

Two sets of questions concerning automation bias effects in interaction with automated decision aids were addressed in this experiment. The first set concerned a better understanding of why operators sometimes followed a wrong

recommendation of an automated aid despite seeking out all parameters necessary to detect that the aid's advice was wrong. The results of the first experiment suggested that at least a few of these errors are related to a kind of looking-but-not-seeing effect, whereby operators maintain their usual strategies of automation verification but stop processing the sampled information attentively. This would reflect a new source of automation bias. In Experiment 2, this issue was investigated by implementing a kind of situation awareness assessment immediately after participants had sent a repair order following a false recommendation of their aid. Specifically, it was explored to what extent they were aware of what they did to verify the automatically generated diagnosis of their aid and to what extent they were aware of the system's parameters they had accessed for automation verification. This analysis allowed for directly investigating whether an observed commission error was related to incomplete automation verification, to automation verification without awareness, or to active discounting of contradictory cues.

The second set of questions was guided by the general idea that operators calibrate their trust in an automated system on the basis of what they know about the system and what experiences they have with using it (Lee & See, 2004; Merritt & Illgen, 2008; Seong & Bisantz, 2008). Specifically, it was analyzed how positive and negative experience with an automated aid would play together over time in determining the level of trust and automation bias. It was assumed that two feedback loops would need to be considered in this respect. The first one represents a positive loop, which is triggered by the experience that the automation provides valid advice. Repeated experience of this kind will successively increase trust in the system and eventually lead to a reduction of effort invested in cross-checks and automation verification. If this effort reduction does not yield any negative performance consequences (which is the more likely the more reliably the aid works), it might get reinforced and result in a self-amplifying process that continuously increases the level of automation bias (cf. the similar concept of "learned carelessness"; Luedtke & Moebus, 2005). However, a reverse effect was assumed to result from a concurrent negative feedback loop, which is mainly triggered by the experience of automation failures. More specifically, it was assumed that the negative feedback loop entails much stronger effects than the positive one. Evidence for this assumption is provided by many studies pointing to strong effects of automation failures on operators' trust and reliance (e.g., de Vries et al., 2003; Dzindolet et al., 2003; Lee & Moray, 1992, 1994; Madhavan, Wiegman, & Lacson, 2006). In particular, findings suggest that the experience of even a single automation failure can considerably reduce the trust of operators in a given system (Lee & Moray, 1992). First evidence that the experience of automation failures also affects the automation verification behavior in interaction with an automated aid has been provided by Bahner, Heuper et al. (2008). In the present experiment, we investigated the dynamic interplay of these feedback loops in more detail by analyzing how subjective trust, automation verification behavior, and the probability to commit a commission error change with the repeated experience that an aid works properly. Furthermore, it

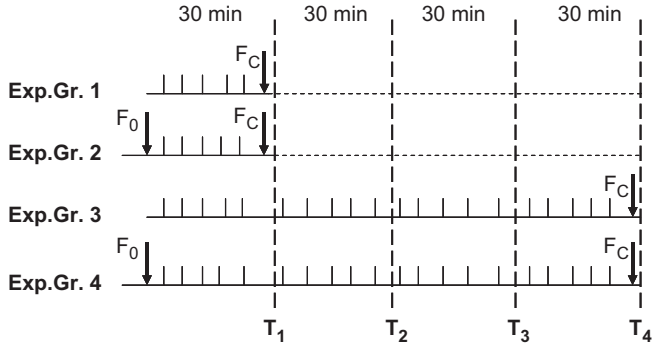

Figure 7. Time course of events for the four experimental groups (FC = critical automation failure of the aid at the end of the session for which issues of automation bias are observed; F0 = automation failure at the beginning of the session as part of the experimental treatment).

was of interest to what extent the dynamics of these effects were dependent on whether the operator had ever experienced an automation failure before.

Method

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

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 (AI support) was used for this study.

Design. The study involved four experimental groups that differed with respect to how long participants had worked with the aid until an automation failure eventually occurred and whether this automation failure was the first or second one the participants were exposed to. The time course of events for the four experimental groups is shown in Figure 7.

Participants of the first experimental group worked with the aid for one 30-min 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. The second experimental group worked according to an identical schedule with the only difference that the run already started with a first system fault for which the diagnosis provided by AFIRA was wrong. Thus, the automation failure at the end of the session represented the second automation failure for this group. A similar variation was realized for Experimental Groups 3 and 4 with the difference that participants of these groups worked for a considerably longer period (four blocks; 20 system faults) with the system before the critical automation failure at the end of the session occurred. 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- and experience-related effects on automation bias involved all four groups.

Dependent measures. Measures used to assess the level of automation verification included (a) AVT and (b) AVIS-N. The definition of these measures was the same as in the first experiment. Performance consequences of automation bias in terms of errors of commission were again quantified by the percentage of participants who followed the wrong diagnosis of the aid in case of an automation failure 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 had been proposed by AFIRA, (b) which parameters they had sampled to verify the aid's advice, and (c) what the critical relations were between the parameters accessed (the relation between parameters provides the critical information needed to disambiguate similar system failures). This questioning was done 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, an assessment was made of how many participants committing a commission error made this error because of (a) an incomplete automation verification, operationally defined like AVIS (see earlier definition); (b) a complete automation verification without awareness, that is, a situation whereby they indeed looked at all information needed to verify the aid's diagnosis but were not able to report what they had seen; or (c) a discounting of contradictory information, a situation whereby they looked at all necessary parameters and were able to report the contradictory information but nevertheless had followed the wrong diagnosis of the aid.

Subjective trust in the diagnostic function of AFIRA was assessed directly by asking the participants how trustworthy they thought AFIRA was ("How much did you trust in the assistance system AFIRA?"). Respondents answered on a 10-point Likert-type scale ranging from *not at all* to *absolutely*. To avoid any demand characteristics, we "hid" the specific question relevant for the study in a larger questionnaire consisting of 18 questions that asked for subjective ratings of trust and estimated reliabilities not only for AFIRA but for all subsystems of AutoCAMS (e.g., oxygen and nitrogen subsystems).

Procedure. The experiment consisted of two familiarization and practice sessions and one experimental session distributed across 3 days. Procedural details of the practice session and the familiarization part of the experimental session were the same as for the first experiment. Participants were randomly assigned to one of the four experimental groups. Independent of the specific experimental group

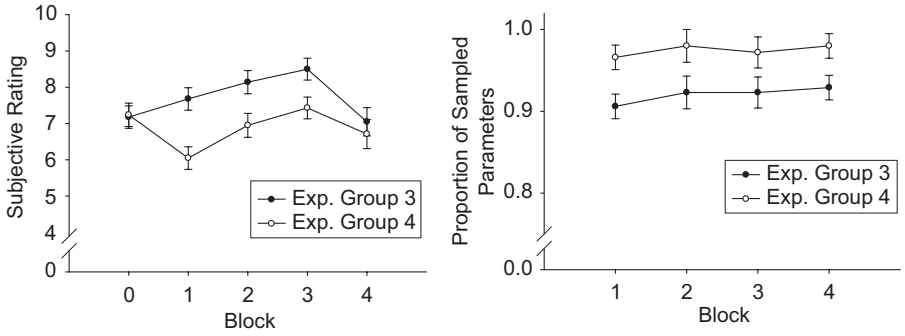

Figure 8. Left graph: Subjective trust ratings across experimental blocks for Experimental Groups 3 and 4 (Block 0 = subjective trust rating after training with the aid). Shown are means and standard errors of responses to the question "How much did you trust the assistance system AFIRA [Automated Fault Identification and Recovery Agent]?" Right graph: Automation verification information sampling across blocks for Experimental Groups 3 and 4. Shown are means and standard errors of AVIS-N (automation verification information sampling of necessary system parameters) measure.

and the length of the experimental session (one 30-min block for Groups 1 and 2 and four blocks for Groups 3 and 4), all participants were instructed that the whole experiment would include a total of five 30-min 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 end of the experiment. For Groups 3 and 4, the blocks were separated by short breaks (~3 min). After the automation failure at the end of the session, the simulation stopped and participants had to answer the AVQ. 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 ratings of its reliabilities were collected before each 30-min block and at the end of the session.

Results

Subjective trust in automation. Effects of positive and negative experience with AFIRA on subjective trust were explored on the basis of data from Experimental Groups 3 and 4. As expected, the dynamics of trust development in these groups were highly dependent on the kind of experience the participants made with the aid. Even more important, negative experience with the aid seemed to affect subjective trust more than did positive experience. This difference becomes evident from the time course of effects shown in Figure 8 (left). Immediately after familiarization and training with AFIRA (Block 0), participants of both groups showed a comparatively high level of trust in the correct functioning of the aid. For participants of Group 3, this level even increased across the first three blocks,

as they repeatedly experienced that the aid worked properly. However, the first experience of an automation failure at the end of Block 4 led to a sharp decrease of trust in this group, down to a level that was even slightly lower than the initial trust. A different picture emerged for participants of Experimental Group 4, who were exposed to a first automation failure already in the beginning of the experimental session. This experience caused a significant and sharp decline of trust that was still visible at the end of the first block, despite the fact that the aid meanwhile had worked properly again for five events. Although trust ratings recovered slowly across the next two blocks (10 events) when the aid worked correctly, they never reached the level of the other group's ratings. After the experience of a second failure at the end of Block 4, trust ratings dropped again considerably yet less than after the first failure. A 2 (groups) × 5 (block) ANOVA of these effects revealed significant main effects of group, $F(1, 41) = 4.62$, $p < .04$, and block, $F(4, 164) = 10.43$, $p < .01$, as well as a significant Group × Block interaction, $F(4, 164) = 5.56$, $p < .01$.

Automation verification. To explore whether the effects seen in subjective trust ratings also would be reflected in differences in automation verification behavior, we compared to what extent participants of Groups 3 and 4 sampled all the system parameters necessary to cross-check the automatically generated diagnosis of AFIRA before confirming it. Only events for which AFIRA provided a correct diagnosis were considered for this analysis. The effects are shown in Figure 8 (right). As becomes evident from this figure, the experience of a failure of the aid at the beginning of the experimental session entailed a significant effect on automation verification (AVIS-N) that persisted across the entire time of the experiment. Participants with an early failure experience were significantly less biased in interaction with the aid than were participants without failure experience. On average, they sampled 97.4% of the system parameters that were necessary to completely verify the aid's diagnoses. In contrast, participants without failure experience checked only 92.0% of the critical information. A 2 (group) × 4 (block) ANOVA revealed a significant group effect, $F(1, 42) = 6.82$, $p < .02$. Neither the block effect nor the Group × Block interaction was significant. Similarly, no significant group effect was found for AVT.

Commission errors. Table 1 provides an overview of the number of participants who committed a commission error when the aid surprisingly proposed a wrong diagnosis at the end of the experimental session. As becomes evident, the risk of committing such error was considerably higher for the group of participants who did not have prior experience of an aid's failure. In this case, 20.4% of the participants committed a commission error. This rate contrasted with a significantly lower error rate (4.5%) for participants who were already exposed to the aid's first failure at the beginning of their session, $\chi^2(1) = 5.10$, $p < .03$. Somewhat contrary to expectations, the number of valid diagnoses prior to the automation failure did not entail any significant effects on automation bias, $\chi^2 < 1$.

TABLE 1. Number of Participants Who Committed a Commission Error When the Aid Failed at the End of the Session

Prior Experience of a False Diagnosis	Correct Diagnoses Prior to the False Diagnosis		Total
	5	20	
No	6 (27.3%)	3 (13.6%)	9 (20.4%)
Yes	0 (0%)	2 (9.1%)	2 (4.5%)
Total	6 (13.6%)	5 (11.4%)	11 (12.5%)

Note. Percentages in brackets reflect the proportion in relation to the number of participants in the different cells ($n = 22$, $n = 44$, and $n = 88$, respectively).

Microanalyses of commission errors. Out of the 11 participants who followed the wrong automation advice at the end of the experiment, 6 showed a behavior resembling that of a complacent operator in supervisory control, as they made the commission error because they did not check all the information that would have been necessary to verify the aid's diagnosis. The other 5 participants followed the wrong automation advice despite checking all parameters that were necessary to realize that the automatically generated diagnosis was wrong. However, 4 of these participants seemed to have conducted these cross-checks without, or with less, attention. This finding was revealed by the results of the questionnaire that was administered after they had falsely confirmed the aid's diagnosis. Although all 5 participants in fact had checked all necessary system information to verify the aid's diagnosis, 4 of them were not able to recall correctly what they had seen. Three of these participants stated that the nitrogen flow they had checked 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. Another participant was not able to recall a critical relation between two parameters even though the log file revealed that he had looked at both. Only 1 of the 11 participants committed the error despite being aware of all the contradictory system information. However, he failed to give a clear reason for his decision. In contrast, out of the 77 participants who had correctly identified the aid's wrong diagnosis, only 4 were not able to recall all necessary parameters that they had cross-checked before.

Discussion

One of the goals of the second experiment was to investigate to what extent positive and negative experience in interaction with an automated aid would determine the level of trust, the degree of automation verification, and the strength of automation bias in terms of commission errors. The assumption that two feedback loops were active in determining the subjective level of trust could

be confirmed. However, the strength of these loops seems to be considerably different. This finding is suggested by the different time courses of trust effects induced by positive and negative experience. About 20 repeated positive incidents were needed to compensate for a decline of trust induced by the single automation failure that occurred early in time during work with the aid. This finding is in line with earlier results of Lee and Moray (1992), who have studied the dynamics of trust development in a supervisory control task. The two proposed feedback loops also determined the level of automation verification and the risk of commission errors in interaction with the automated aid. Participants who had already had the experience of an automation failure turned out to be less biased and less prone to commit a commission error when the aid failed a second time. This finding confirms similar results reported by Bahner, Hueper, et al. (2008) and suggests that direct experience of automation failures may provide an effective countermeasure for automation bias effects.

Whereas the effects of a single automation failure on subjective trust seemed to recover (albeit slowly) over time if the aid worked properly again afterward, a similar effect was not observed for automation verification information sampling behavior. Regaining the initial trust level was not reflected in the participants' cross-checking behavior, which persisted at a nearly perfect level and thereby reduced the probability of a commission error. This finding suggests that the impact of the negative feedback loop is more enduring on the behavioral level than on the subjective trust level.

Another study objective aimed at a better understanding of the proposed looking-but-not-seeing effect as a possible cause of commission errors. Overall, the proportion of participants committing a commission error (n = 11 out of 88, i.e., 12.5%) was smaller than in previous studies. However, among only participants for whom the automation failure at the end of the session was the first failure (Experimental Groups 1 and 3), the percentage rose to 20.5% (n = 9 out of 44), which corresponds to the figure reported by Bahner, Hueper, et al. (2008). Similar to the first experiment, only about half of the participants (n = 6) who committed a commission error did so because they did not check all information needed to verify the aid's recommendation. The other half of the participants (n = 5) actually checked all relevant information needed to identify the wrong diagnosis but, nevertheless, followed the incorrect advice. As was revealed by the situation awareness questionnaire, only 1 of these latter participants could correctly report what the system parameters indicated. The others had indeed looked at all the parameters but obviously had not processed the information attentively. This finding supports the looking-but-not seeing hypothesis derived from the results of the first experiment and suggests that automation bias can be associated with three different effects: (a) a withdrawal of attention in terms of incomplete cross-checks of information, (b) an active discounting of contradictory information, and (c) an inattentive processing of the contradictory information analogue to a looking-but-not-seeing effect. The latter effect is in line with earlier results from automation monitoring (e.g., Duley et al., 1997; Sarter et al., 2007) and enlarges the set of already known sorts of automation bias.

Summary and Conclusions

The present research provides detailed insights into the human performance consequences of automated decision aids and their dependence on the kind of function allocation operationally defined in terms of DOA. As expected, the provision of automated aids resulted in clear performance benefits that were directly dependent on the DOA, that is, more highly automated aids led to higher performance improvements than did less automated aids. In the first study, this finding was reflected in shorter FIT as well as better performance in stabilizing the simulated life-support system in states of failure. However, these benefits of automated support were not without costs. Performance costs were mainly reflected in return-to-manual difficulties as well as issues of automation bias. Both sorts of costs showed at least some relation to the kind of function allocation. Difficulties of return-to-manual performance remained limited to the most highly automated aid, which provided support not only for cognitive processes involved in fault identification but also for the implementation of appropriate manual control actions. Issues of automation bias in terms of neglect of automation verification were primarily found for the most highly automated aid when automation verification included comparatively complex procedures. These findings provide support for earlier assumptions that a lower DOA might provide advantages to a higher one in this respect (e.g., Endsley & Kaber, 1999; Endsley & Kiris, 1995). However, this advantage of a lower DOA does not hold for preventing risks of commission errors as well. Commission errors in case of a first automation failure occurred independent of the DOA. The more detailed analysis of the origins of these errors in the second experiment revealed that the vast majority of them could be explained by participants' withdrawing attention from the automated processes, directly reflected either in insufficient verification or, more subtle, in inattentive information processing. Furthermore, the results provide evidence that these effects represent phenomena that are directly based on the practical experience an operator has in interaction with an automated system as well as on individual differences, which obviously make some individuals more prone to automation bias effects than others. The latter is suggested by the fact that only a minority of the participants who never had the experience that the aid can fail committed a commission error although almost all of them showed some evidence of neglect of automation verification.

Altogether the pattern of effects found in the present research supports a framework model of complacency and automation bias, which has recently been described in detail by Parasuraman and Manzey (2010). This model conceptualizes complacency and at least most kinds of automation bias as reflecting automation-induced attentional phenomena that "result from a complex interaction of personal, situational, and automation-related characteristics" (Parasuraman & Manzey, 2010, p. 403). Practical conclusions that might be drawn from this work relate to the advantages of medium levels of automation—such as AS support in the present study—as a sort of compromise for balancing performance benefits and costs of automation. In addition, the significance of practical experience with automation failures for preventing issues of automation bias might be taken

into consideration for concepts of training and familiarization of operators with automated decision aids (cf. Bahner, Hueper, et al., 2008; Parasuraman & Manzey, 2010).

Limitations of the present research relate to the typical characteristics and constraints of laboratory experiments. Although the experimental task used for this research represented a sort of microworld, which makes it more complex than usual laboratory tasks, and the participants in this research, that is, engineering students, seem to be more or less similar to the target group of operators typically involved in the use of such systems, other aspects of the current research have only limited ecological validity. This limitation holds in particular for assessing the impact of time-related and experience-related effects on return-to-manual issues and automation bias effects. Whereas in the real world, these kinds of effects usually develop across months and years, laboratory experiments require simulating these dynamics within a couple of hours. It is difficult to assess how this limitation might have affected the strength of effects observed in the given studies. In particular, the finding of comparatively weak return-to-manual effects might underestimate the performance consequences of automation support in the real world.

Acknowledgments

This research was sponsored by Research Grants MA 3759/1-1 and MA 3759/1-2 provided by Deutsche Forschungsgemeinschaft. Thanks are due to Marcus Bleil for programming the experiment and data recording software and to Sabine Jatzev and Jasmin Hannighofer for their help with data collection.

References

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 52nd Annual Meeting* (pp. 1330–1334). Santa Monica, CA: Human Factors and Ergonomics Society.

Bahner, J. E., Hueper, 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.

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.

Duley, J. A., Westerman, S., Molloy, R., & Parasuraman, R. (1997). Effects of display superimposition on monitoring of automation. In *Proceedings of the 9th International Symposium on Aviation Psychology* (pp. 322–328). Columbus, OH: International Symposium of Aviation Psychology.

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

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.

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.

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.

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.

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.

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.

Luedtke, A., & Moebus, C. (2005). A case study for using a cognitive model of learned carelessness in cognitive engineering. In G. Salvendy (Ed.), *Proceedings of the 11th International Conference of Human-Computer Interaction*. Mahwah, NJ: Lawrence Erlbaum.

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., Bleil, M., Bahner-Heyne, J. E., Klostermann, A., Onnasch, L., Reichenbach, J., & Röttger, S. (2008). *AutoCAMS 2.0. manual*. Retrieved from http://www.aio.tu-berlin.de/?id=30492

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.

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

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.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *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.

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.

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.

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. (2000). Function allocation: Algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, *52*, 203–216.

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

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

Dietrich Manzey is a university professor of work, engineering, and organizational psychology at the Institute of Psychology and Ergonomics, Berlin Institute of Technology, Germany. He received his PhD in experimental psychology at the University of Kiel, Germany, in 1988 and his habilitation in psychology at the University of Marburg, Germany, in 1999.

Juliane Reichenbach is a PhD candidate in the Department of Psychology and Ergonomics of Berlin Institute of Technology. She has obtained a master in psychology in 2004 from University of Regensburg. The current article represents a part of her dissertation work.

Linda Onnasch is a research fellow in the Department of Psychology and Ergonomics of Berlin Institute of Technology, where she has obtained a master in psychology in 2009. She is currently working on a PhD addressing issues of trust in automation.