

Misuse of automation: The impact of system experience on complacency and automation bias in interaction with automated aids

Juliane Reichenbach, Linda Onnasch and Dietrich Manzey
 Berlin Institute of Technology
 Berlin, Germany

The study investigates how complacency and automation bias effects in interaction with automated aids are moderated by system experience. Participants performed a supervisory control task supported by an aid for fault identification and management. Groups differed with respect to how long they worked with the aid until eventually an automation failure occurred, and whether this failure was the first or second one the participants were exposed to. Results show that negative experiences, i.e., automation failures, entail stronger effects on subjective trust in automation as well as the level of complacency and automation bias than positive experiences (correct recommendations of the aid). Furthermore, results suggest that commission errors may be due to three different sorts of effects: (1) a withdrawal of attention in terms of incomplete cross-checks of information, (2) an active discounting of contradictory system information, and (3) an inattentive processing of contradictory information analogue to a “looking-but-not-seeing” effect.

INTRODUCTION

Human interaction with automated systems often involves the risk of misuse of automation, i.e. an uncritical reliance on its proper functioning (Parasuraman & Riley, 1997). One important aspect of automation misuse is reflected in an insufficient monitoring or checking of automated functions, a phenomenon which commonly has been referred to as complacency (Parasuraman, Molloy & Singh, 1993). Originally complacency has been identified as an issue in supervisory control of autonomous processes. A typical example involves pilots who rely on the proper functioning of their autopilot so much that they neglect to monitor and check it appropriately. Important performance consequences of complacency may include a loss of situational awareness and an elevated risk of missing automation failures.

However, complacency-like effects can emerge in other fields of human-automation interaction as well. For example, Mosier & Skitka (1996) have introduced the concept of automation bias. They argue that decision aids may be misused by taking their outcome “...as a heuristic replacement for vigilant information seeking and processing” (p. 205). One kind of error resulting from this effect includes commission errors where an operator follows an aid’s advice even though it is wrong. According to Skitka, Mosier & Burdick (1999), “commission 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). The latter alternative reflects a decision bias in a strict sense. However, the former alternative, i.e., following the aid’s recommendation without verification, seems to reflect a decision bias effect which, on a behavioral level, involves a withdrawal of attention that resembles complacency effects in supervisory control.

Empirical evidence for a link between complacency and automation bias has been provided by a recent set of studies (Bahner, Hueper & Manzey, 2008; Bahner, Elepfandt & Manzey, 2008; Manzey, Reichenbach & Onnasch, 2008). In these studies, the participants had to perform a supervisory

control task which required them to monitor an autonomously running life support system and to intervene whenever they detected a system fault. This task was supported by an automated aid. In case of system faults it provided the human operator with an automatically generated diagnosis and recommendations for fault management. Complacency in interaction with this aid was operationally defined by the extent to which the operators cross-checked the automatically generated diagnoses before they accepted it and intervened in the system. Between 20% and 75% of the participants in these studies were found to commit a commission error when the aid, after some time of proper functioning, surprisingly provided a wrong diagnosis. Detailed analyses of the information sampling behavior revealed that these operators showed a higher level of complacency in their interaction with the aid than those who detected the failure of the aid. However, not all of the commission errors could be related to an obvious complacency effect. Up to 50% of the participants committing a commission error followed the aid’s wrong advice despite seeking out all system information needed to detect that the aid’s advice was wrong.

The current study capitalizes on this research. Using the same experimental paradigm as in the research referred to above, it is explored to what extent positive and negative experiences with an automated aid play together in determining the level of trust, complacency and strength of automation bias. It is assumed that two feedback loops need to be considered in this respect. The first one represents a positive loop which is triggered by the experience that the automation provides a valid advice. Repeated experience of this kind will successively increase the 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 which continuously increases the level of complacency and automation bias (cf. the similar concept of “learned carelessness”; Luedtke & Moebus, 2004). However, a reverse effect is assumed to result from a concurrent negative

feedback loop which is mainly triggered by experience of automation failures. This is suggested by findings showing that even the experience of a single automation failure can considerably reduce the trust of operators in a given system (Lee & Moray, 1992). For example, after the experience of failures during training operators are less complacent when working with an automated aid (Bahner et al., 2008a). In the present experiment the dynamic interplay of these feedback loops is investigated 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 is of interest to what extent the dynamics of these effects are dependent on whether or not the operator has ever experienced an automation failure before.

The second question concerns a better understanding of why operators sometimes follow a wrong recommendation of an automated aid despite seeking out all parameters necessary to detect that the aid's advice was wrong. On first sight this might be taken as evidence for discounting contradictory information. Yet, a closer inspection of the data from Manzey et al. (2008) suggested that at least some of these errors were more likely related to a kind of "looking-but-not-seeing effect", where operators maintain their usual strategies of information sampling but stop to process the sampled information attentively. This would reflect a new variant of automation bias effect. In the present experiment this issue is investigated by analyzing to what extent an observed commission error is related to incomplete automation verification, to automation verification without awareness, or to an active discounting of contradictory cues.

METHOD

Participants

88 engineering students (65 male, 23 female; mean age: 24.05 yrs) participated in the study. Participants were paid € 70 for completing the study.

Apparatus: AutoCAMS 2.0

A "microworld" simulation of a supervisory process control task was used for the experiment (AutoCAMS 2.0, Manzey, Bleil, Bahner-Heyne, Klostermann, Onnasch, Reichenbach & Röttger, 2008). This system simulates an autonomously running life support system of a spacecraft consisting of five subsystems that are critical to maintain atmospheric conditions in the cabin with respect to different parameters (e.g. oxygen, pressure, carbon dioxide). During normal operation, all of these parameters are automatically kept within target range. However, due to malfunctions in the system (e.g. blockage of a valve, defective sensor) parameters can go out of range. 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 ("red light"). A time counter starts displaying how much time has elapsed since the occurrence of the fault. In order 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 seconds; during this time the operator is required to control the affected subsystem manually. If the repair order sent was correct, the master alarm turns green and all subsystems run autonomously again. In case of a wrong repair order, the alarm stays red and manual control is required until a correct repair is initiated and completed.

In the present experiment, the operator's task is supported by an automated aid (Automated Fault Identification and Recovery Agent, AFIRA). For each occurring system fault, AFIRA provides an automatically generated diagnosis. Upon confirmation by the operator, it executes all steps necessary to manage a given fault and initiate an appropriate repair. In order to verify the aid's diagnosis before confirming it, the operator has access to all important raw data providing information about the current system state. These include tank levels and flow rates for oxygen and nitrogen, and a history graph for each of the five subsystems. However, this information is not always visible but has to be activated for a 10s view by a mouse click on the tank, flow meter or history graph, respectively.

In addition to the primary task, two concurrent secondary tasks have to be performed. The first one is a prospective memory task in which participants are required to check and record the carbon dioxide values every 60 seconds. The other one is a simple 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 as fast as possible. This icon appears in random intervals roughly twice per minute.

Design

The study involved four experimental groups which differed with respect to how long they had worked with the aid until eventually an automation failure 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 different experimental groups is shown in figure 1.

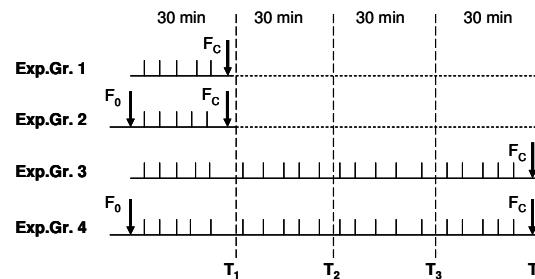


Figure 1: Time course of events for the four experimental groups (Fc: critical automation failure of the aid at the end of the session for which issues of automation bias are observed; F0: automation failure at the beginning of the session as part of the experimental treatment)

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

essentially same schedule with the only difference that the run started with a system fault for which the diagnosis provided by AFIRA was wrong. 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 (4 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 effects on automation bias involved all four groups.

Procedure

The experiment consisted of two familiarization and practice sessions and one experimental session distributed across three different days. The practice session on the first day lasted approx. 4 hours and included familiarization and practice with the AutoCAMS system. Participants were trained to manually, i.e., without automation support, identify and manage seven possible faults. On the second day, all participants had to perform a 45 min test trial which served to test their acquired skills according to a predefined criterion. Only the participants who passed this test were accepted to participate in the experiment.

Each experimental session started with an introduction to AFIRA. This familiarization included a description of the aid's function as well as a short practice trial. During this practice trial AFIRA always provided correct diagnoses and recommendations. However, participants were informed that the aid's reliability though being high would not be perfect. They were warned to always cross-check the proposed diagnoses before confirming it. After this introduction the experimental run started. For this run the participants were randomly assigned to one of the four different experimental groups. Independent of the specific experimental group all participants were instructed that the whole experiment would include a total of five 30 min blocks. This was done in order to assure that all participants worked with the same attitude and expectation, and were not able to anticipate the real end of the experiment.

After the automation failure at the end of the session (first failure for groups #1 and #3; second failure for groups #2 and #4), the program stopped as soon as the participant decided to either follow the aid's advice or disagreed with AFIRA's diagnosis. Participants were then asked questions about their approach of automation verification. Specifically, they had to provide information about which diagnosis had been proposed by AFIRA, which parameters they had sampled in order to verify the aid's advice, and what the critical relations were between the parameters accessed (the relation between parameters provides the critical information needed to disambiguate similar system failures). This was done in order to check to what extent the participants were aware of the steps they had performed and the system information they had accessed.

Ratings of subjective trust in the different components of the AutoCAMS system (e.g. oxygen, nitrogen, carbon dioxide subsystem) and AFIRA, as well as ratings of

its reliabilities were collected before each 30 min block and at the end of the session.

Dependent Measures

Measures used to assess the level of complacency included (1) *automation verification time (AVT)*, and (2) *automation verification information sampling (AVIS)*. AVT was defined as the time interval [s] from the appearance of the master warning until confirming or vetoing AFIRA, independent of whether this decision was right or wrong. AVIS was defined as the percentage of system parameters accessed (via mouse click) which were necessary to completely verify a given diagnosis provided by AFIRA. Only parameters accessed between the occurrence of the master warning and conforming or vetoing AFIRA were considered for this measure.

Automation bias was analyzed by the *percentage of participants committing a commission error*, defined as the percentage of participants who followed the diagnosis in case of an automation failure at the end of the experiment. In addition, the underlying determinants of commission errors were analyzed by assessing how many participants committing a commission error made this error because of

- a) *an incomplete automation verification*: operationally defined like AVIS (see above),
- b) *a complete automation verification without awareness*: number of participants who indeed looked at all information needed to verify the aid's diagnosis but were not able to report what they had seen in the SA inquiry,
- c) *a discounting of contradictory information*: number of participants who looked at all necessary parameters and were able to report the contradictory information in the SA inquiry 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. Respondents answered on a 10-point Likert-type scale ranging from *not at all* to *absolutely*.

RESULTS

Subjective Trust in Automation

Effects of positive and negative experience with AFIRA on subjective trust were explored based on 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 entail much stronger effects on subjective trust than positive experience. This becomes evident from the time course of effects shown in figure 2. 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 over the first three blocks as they repeatedly made the experience 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 which 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 which 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 over the next two blocks (10 events) where the aid worked correctly, they never reached the level of the other group. 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) x 5(Block) ANOVA of these effects revealed significant main effects of Group, $F(1,41)=4.62$, $p<.04$, and Block, $F(4,164)=10.43$, $p<.001$, as well as a significant Group x Block interaction, $F(4,164)=5.56$, $p<.001$.

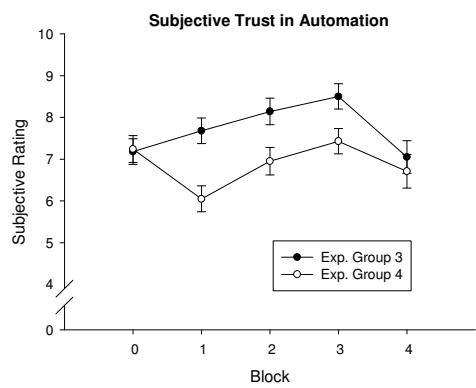


Figure 2: Time course of subjective trust ratings across experimental blocks for participants of experimental groups #3 and #4 (block 0 = subjective trust rating after training with the aid)

Automation Verification

In order to explore whether the effects seen in subjective trust ratings also would be reflected in differences in automation verification behavior, it was compared to what extent participants of group #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 3. 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) which persisted over the entire time of the experiment. Participants with an early failure experience were significantly less complacent in interaction with the aid than participants without failure experience. On average they sampled 97.4% of the system parameters which were necessary to completely verify the aid's diagnoses. In contrast, participant without failure experience only checked 92.0% of the critical information. A 2(Group) x 4(Block) ANOVA revealed a significant Group effect, $F(1,42) = 6.82$, $p<.02$. Neither the

Block effect, $F(3,126)=1.11$, nor the Group x Block interaction, $F(3, 126) < 1$, was significant. No significant Group effect was found for automation verification time, $F<1$.

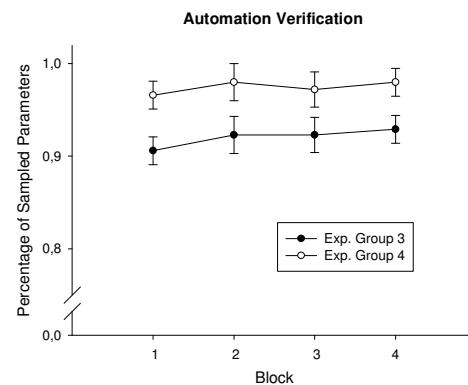


Figure 3: Time course of automation verification information sampling across blocks for participants of experimental groups #3 and #4.

Automation Bias

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 this sort of automation bias 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 contrasted to a significantly lower error rate (4.5%) for participants who were already exposed to a first failure of the aid 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 (percentage) of participants who committed a commission error when the aid failed at the end of the session.

Prior experience of a false diagnosis	N of correct diagnoses prior to the false diagnosis	Total
No	6 (27.3%)	9 (20.4%)
Yes	0 (0%)	2 (4.5%)
Total	6 (13.6%)	5 (11.4%)

Micro analyses of Commission Errors

Out of the 11 participants who followed the wrong automation advice at the end of the experiment, only six could be classified as being complacent in a "classical" sense, as they made the commission error because they did not check all the information that would have been necessary to verify the correctness of the aid. The other five participants followed the wrong automation advice despite checking all parameters that were necessary to realize that the automatically generated diagnosis was wrong. However, four of these participants seemed to have conducted these cross-checks without or with

less attention. This was revealed by the results of the questionnaire that was administered after they had falsely confirmed the aid's diagnosis. Although all five participants in fact had checked all necessary system information to verify the aid's diagnosis, four of them were not able to recall correctly what they had seen. Three of these participants stated that the nitrogen flow they had checked was on standard level although it actually was much lower which is an indicator for a specific system fault. Another participant was not able to recall a critical relation between two parameters even though the logfile revealed that he had looked at both. Only one of the eleven participants committed the error despite being aware of all the contradictory system information. However, he failed to give a clear reason for this 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 which they had cross-checked before.

DISCUSSION

The goal of the present study was to investigate to what extent positive and negative experience in interaction with an automated aid determine the level of trust, the degree of automation verification (i.e. complacency), and the strength of automation bias in terms of commission errors. Another study objective aimed at a better understanding of the proposed "looking-but-not-seeing effect" as a possible cause of commission errors. In this regard, four main conclusions can be drawn from the results.

(1) The assumption that two feedback loops interact dynamically in determining the subjective level of trust could be confirmed. However, the strength of these loops seems to be considerably different. This is suggested by the different time courses of trust effects induced by positive and negative experience. About 20 repeated positive experience were needed to completely compensate for a decline of trust induced by a single automation failure that occurred early in time during working with the aid. This is in line with earlier results of Lee and Moray (1992) who have studied the dynamics of trust development in a supervisory control task.

(2) The two proposed feedback loops also determined the level of complacency and risk of commission errors in interaction with the automated aid. Participants who had already made the experience of an automation failure turned out to be less complacent and less prone to commit a commission error when the aid failed a second time. This confirms similar results reported by Bahner et al. (2008a) and suggests that direct experience of automation failures may provide an effective countermeasure for complacency and automation bias effects.

(3) Whereas the effects of a single automation failure on subjective trust seem to recover (albeit slowly) over time if the aid works properly again afterwards, 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 suggests that the impact of the

negative feedback loop is more enduring on the behavioral level than on the subjective trust level.

(4) One of the most interesting aspects extracted from the present study is the idea of different causes for automation bias. Only half of the participants committing a commission error behaved complacent in a "classical" sense, i.e., they did not check all necessary information needed to verify the aid's recommendation. The other half of the participants actually checked all relevant information needed to identify the wrong diagnosis but, nevertheless, followed the incorrect advice. However, only one of these participants could correctly report what the system parameters indicated. It seems therefore, 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, Westerman, Molloy & Parasuraman, 1997) and enlarges the set of causes of automation bias.

REFERENCES

- Bahner, J.E., Hueper, A.C., & Manzey, D. (2008a). Misuse of automated decision aids: Complacency, automation bias, and the impact of training experiences. *International Journal of Human-Computer Interaction*, 66, 688-699.
- Bahner, J.E., Elepfandt, M. & Manzey, D. (2008b). Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: HFES.
- Duley, J.A., Westerman, S., Molloy, R., & Parasuraman, R. (1997). Effects of display superimposition on monitoring of automation. *Proceedings of the 9th International Symposium on Aviation Psychology* (pp. 322-328). Columbus, OH: ISAP.
- 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, The Netherlands: Elsevier.
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35, 1243-1270.
- Luedtke, A. & Moebus, C. (2005). A case study for using a cognitive model of learned carelessness in cognitive engineering. In G. Salvendy (Ed.), *Proc. of the 11th International Conference of Human-Computer Interaction*. Mahwah: Erlbaum.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2008). Performance consequences of automated aids in process control: The impact of function allocation. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: HFES.
- Manzey, D., Bleil, M., Bahner-Heyne, J.E., Klostermann, A., Onnasch, L., Reichenbach, J., & Röttger, S. (2008). *AutoCAMS 2.0. Manual*. Available from www.aio.tu-berlin.de/?id=30492 [19 February 2009].
- Mosier, K. L. & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman, & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 201-220). Mahwah, NJ: Lawrence Erlbaum Associates.
- Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-259.
- Parasuraman, R., Molloy, R., & Singh, I.L. (1993). Performance consequences of automation induced "complacency". *The International Journal of Aviation Psychology*, 2, 1-23.
- Skitka, L.J., Mosier, K.L., Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51, 991-1006.