System reliability, performance and trust in adaptable automation

Alain Chavaillaz a, *, David Wastell b, Jürgen Sauer a

a Department of Psychology, University of Fribourg, Fribourg, Switzerland
b Division of Operations Management and Information Systems, Nottingham University Business School, Nottingham, UK

Article info

Article history:
Received 27 February 2013
Received in revised form 24 June 2015
Accepted 9 July 2015
Available online 28 August 2015

Keywords:
Reliability
Adaptable automation
Performance
Trust

Abstract

The present study examined the effects of reduced system reliability on operator performance and automation management in an adaptable automation environment. 39 operators were randomly assigned to one of three experimental groups: low (60%), medium (80%), and high (100%) reliability of automation support. The support system provided five incremental levels of automation which operators could freely select according to their needs. After 3 h of training on a simulated process control task (AutoCAMS) in which the automation worked infallibly, operator performance and automation management were measured during a 2.5-h testing session. Trust and workload were also assessed through questionnaires. Results showed that although reduced system reliability resulted in lower levels of trust towards automation, there were no corresponding differences in the operators' reliance on automation. While operators showed overall a noteworthy ability to cope with automation failure, there were, however, decrements in diagnostic speed and prospective memory with lower reliability.

© 2015 Published by Elsevier Ltd.

1. Introduction

While automated support systems (e.g. autopilots in aircraft, car navigation systems) provide great benefits to human operators, they sometimes give inappropriate advice or make incorrect decisions that may result in serious performance problems, even leading to fatalities (see Parasuraman and Riley, 1997). Since such support systems have become very common at work, an increasingly important part of the operators’ job concerns monitoring these systems and reacting adequately to such potential failures. A key factor is therefore the reliability of such systems and the effects it may have on important variables, such as trust and performance (Parasuraman et al., 1993a; Sheridan, 2002). The issue of reliability is of particular interest in the context of advanced support systems that allow for non-static allocation of tasks between the human and the machine, as encapsulated in the concepts of adaptable and adaptive automation.

1.1. Adaptable and adaptive automation

Automation refers to a partial or complete transfer of operator tasks to a machine, shifting operator status from active to passive agency (e.g. Wickens et al., 2004). Task allocation can either be static (i.e. a fixed task allocation between the operator and the machine, e.g. Sheridan, 2002), or be adjusted over time and tailored to operators’ competences and current workload (e.g. Sauer et al., 2012). In adaptable automation, task allocation is decided by the operator, whereas in adaptive automation this decision is taken by the technical system (e.g. Parasuraman and Wickens, 2008). Overall, when compared to static automation, there seem to be benefits of both adaptable and adaptive automation with regard to operator performance (e.g. Parasuraman et al., 1993b; Kaber and Riley, 1999; Inagaki, 2003; Sauer et al., 2012).

The question of task allocation is not simply one of choice between the extreme poles of ‘no automation’ and ‘full automation’, but may also entail more fine-grained decisions regarding the optimal level of automation (LOA). Various models have been proposed to distinguish between different LOA (e.g. Sheridan and Verplank, 1978; Endsley, 1999; Parasuraman et al., 2000). Each LOA denotes the degree to which the operator is involved in information acquisition, information analysis, decision-making and decision execution (Parasuraman et al., 2000). For example, 10 LOAs were proposed in the seminal model of Sheridan and Verplank (1978), ranging from full manual to fully automated control, with intermediate levels in which operators can veto decisions taken by the system. While LOA switches are decided by operators in adaptable automation, in adaptive automation measurable criteria are used automatically as switch points; these include the...
operators’ psychophysiological state, performance, or critical events (Kaber and Endsley, 2004).

The high flexibility associated with adaptable automation may be considered an advantage since it allows balancing out varying levels of workload (e.g. extreme over- or underload) by switching LOA to increase or reduce support provided by the automatic system (Wickens et al., 2004). Furthermore, the principles of adaptable automation map well onto psychological stress models advocating the importance of increasing operator control to reduce the negative effects of strain (e.g. Karasek and Theorell, 1990). For example, Karasek and Theorell’s demand-control model proposes that higher levels of operator control allows a better management of high demands. Against the background of such models, offering operators free choice in adapting LOA to their needs appears to be a sensible approach.

1.2. System reliability and trust

An important issue in the design of automation is the question of how system reliability affects important outcome variables. In the field of static automation, there is a considerable number of studies on the impact of variations in system reliability. Most work has shown that under decreasing system reliability, operator performance deteriorates across a range of work environments, including flight simulation (e.g. Bailey and Scerbo, 2007), in-vehicle navigation (Ma and Kaber, 2007), waste processing facility (Wiegmann et al., 2001), and military applications (Rovira et al., 2007). Although some studies have not shown such negative effects (e.g. Parasuraman et al., 1993a), they are outnumbered by studies which do (for an overview, see Wickens and Dixon, 2007).

In the context of adaptable and adaptive automation, there are few studies examining system reliability. Moray et al. (2000) investigated the direct impact of reliability with an adaptive system. The task in this study involved the management of a central heating simulation, aided by a support system which helped in the diagnosis of faults and the choice of repair action. The reliability of this support system varied from 70% to 100%. Operators showed better performance under automated compared to manual control mode when the system was totally reliable, and vice versa when system reliability was low. Although the study demonstrated that operators were able to evaluate automation usefulness according to its reliability, the event-driven adaptive automation mode did not allow operators to adapt the automation level to their current need, since the LOA was predefined for each type of fault. In contrast, adaptable automation would provide such opportunities freely to the operator, allowing their actual preferences to be determined, according to such factors as their level of trust towards the automation.

Originally developed as a concept in social psychology, trust has been successfully applied to the context of human–machine interaction (e.g. Lee and Moray, 1994). It may be defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (Lee and See, 2004, p. 54). In a meta-analysis, Hancock et al. (2011) were able to identify three factors that influence trust towards automation, namely the environment (e.g. team collaboration and tasking), the operator (abilities and personality), and the machine (attributes and performance). They showed that environmental and operator characteristics had a smaller influence than the machine-related characteristics (such as automation performance). Lee and Moray (1992) also pointed out that performance of the automated system reflects its reliability, predictability, and ability in helping operators achieve task goals. The level of trust (as an attitude) may be assumed to influence the reliance of operators (i.e. their actual behaviour) on automation, and a number of studies have shown that both trust and reliance are influenced by system reliability (e.g. Wiegmann et al., 2001; Bailey and Scerbo, 2007; Ma and Kaber, 2007). However, system reliability does not always affect attitude and behaviour in such a congruent manner. For example, Rovira et al. (2007) showed an effect only on objective measures such as performance and use of automation, whereas other studies have shown the opposite impact, affecting subjective measures such as trust exclusively (Dzindolet et al., 2003). A similar dissociation of subjective trust and objective trust-related behaviour was also observed in a study by Manzey et al. (2012), which suggested that subjective trust ratings responded to variations in automation reliability more closely than automation management behaviour (e.g. information sampling of critical system parameters). The findings of these previous studies suggest that operators are sensitive to automation reliability but its concrete effects may manifest themselves in different ways, dependent on the work environment and other factors (e.g. the simulation being employed).

Although operators seem to be sensitive to automation reliability, this does not necessarily imply that they actually match their trust to the actual level of reliability (i.e. trust calibration). Indeed, a highly accurate calibration of trust is rarely observed (e.g. Wiegmann, 2002; Wiegmann et al., 2001), not even after prolonged use of the system (Muir and Moray, 1996). The limited ability of operators to calibrate their trust may lead to non-optimal use of automation, including both under and over-reliance (e.g. Lee and See, 2004; Wiegmann et al., 2001). For instance, Bahner et al. (2008) observed that operators who relied on the support system sampled less information than needed. It is notable that all these previous studies on reliability and trust compared a manual mode with static automation. One possibility to improve trust calibration might be the implementation of a more flexible form of automation, such as adaptable or adaptive automation (Bailey and Scerbo, 2007).

Automation reliability and its impact on the operator (cf. Moray et al., 2000) may also be moderated by the operators’ self-confidence in their own capabilities of managing the system. Lee and Moray (1994) showed that when self-confidence was high, operators preferred to use the manual control mode, whereas there was a strong preference for even unreliable automation when self-confidence was low. These results show the importance of taking self-confidence into account as a moderating variable (Moray et al., 2000). While self-confidence enjoys some stable qualities like a personality trait, it also has state qualities and may be influenced by automation design (Sauer et al., 2012). For example, adaptable automation may provide more opportunities for operators to practise their technical skills than adaptive automation, which could enhance their self-confidence.

1.3. Present study

The aim of the present study was to investigate the effects of automation reliability on a range of outcome measures, notably operator performance, automation management, and trust. These effects were examined using a complex process control simulation. An extended testing session was used to observe possible changes over time. The range of LOA offered here (LOA 1–5) was also larger than in much previous work, which has typically offered only simple binary choices (i.e. manual control vs. automation). The influence of expertise on support system reliability was also addressed by using familiar and novel fault scenarios. Finally, because the effects of automation failure are not immediate (Lee and Moray, 1992), after-effects of support system failures were assessed in the form of trust but also objective measures such as reliance and performance. According to the research literature (Lee
and Moray, 1992; Itoh et al., 1999), operator trust recovers more quickly from a single acute failure than from chronic failures.

The process control simulation (AutoCAMS) serves as a model for a complex work environment. Since it encompasses multiple tasks of two different priority levels, it permits assessment not only of operator performance but also of workload in the form of secondary task performance; it has been shown (Hockey, 1997) that under high workload, secondary tasks suffered more from performance impairment than primary tasks. In addition to objective measures, subjective ratings on trust, self-confidence and workload were also taken. To investigate the impact of automation reliability on trust, three reliability levels (60, 80, and 100% reliable) were selected (cf. Ma and Kaber, 2007; Wiegmans et al., 2001). Reliability levels corresponded to the proportion of misdiagnoses proposed by the support system for the fault scenarios encountered during the experimental session.

The following predictions were made: (a) Under low automation reliability, operators would generally select lower LOAs, show degraded performance (especially for the secondary tasks, Hockey, 1997) and a lower trust level; their performance would also decrease more strongly over time; (b) Operators would select higher LOAs when faced by a novel than a practised fault; (c) The impact of a support system failure would persist longer for low automation reliability, that is, performance and trust would be more degraded toward the end of the experimental session for more unreliable automation.

2. Method

2.1. Participants

Thirty-nine participants (10 females), aged from 18 to 35 yrs (M = 22.8, SD = 3.2), took part in the study. All were students with science or engineering backgrounds to ensure that they had a comparable technical competence regarding real process control tasks. They were paid CHF 120 (about € 100) for their participation.

2.2. Experimental design

A mixed 3 × 10 design was employed in this study. The level of system reliability was a between-subjects factor (high, medium, low). The second independent variable, fault scenario (10 fault states) was a within-subjects factor, in which the effects of fault familiarity, time-on-task, and after-effects of a support system failure were tested by means of planned comparisons (see 2.8).

2.3. Simulation environment

A PC-based simulation environment, called AutoCAMS 2.0 (Cabin Air Management System, see for details Manzey et al., 2008), was used as a model of a complex process control task. This simulation environment, modified and enhanced over many years by different research groups (e.g. Hockey et al., 1998; Lorenz et al., 2002; Manzey et al., 2008), simulated a life-support system of a space shuttle. The present study employed an enhanced version used by Sauer et al. (2012) allowing for the modelling of adaptable, as well as adaptive, automation.

The simulation contains five critical parameters (CO₂, O₂, pressure, temperature, and humidity) reflecting the air quality in the shuttle cabin. When functioning normally, automatic controllers ensure that these parameters remain within a defined target range. The main interface and its components are presented in Fig. 1. Operators had four tasks to accomplish. They were asked (a) to diagnose and fix any system disturbances as fast as possible and (b) to maintain the stability of the system throughout the experimental session by manually controlling the subsystems, if necessary. In addition to these two primary tasks, operators had to perform two secondary tasks: (a) a prospective memory task, for which they had to record the current level of the N₂ tank at every full minute, and (b) a probe detection task, for which they had to click as fast as possible on a symbol which appears at irregular intervals (on average about 30 s) to indicate the connection between ground control and the space shuttle (see Fig. 1). Failing in their tasks did not have any direct consequences for operators, but they were instructed that system instability would have negative consequences for astronauts in the space shuttle. All control actions performed by the operator and each change in the system were recorded in a file, was used for further analyses. Finally, a support system, called AFIRA (Automated Fault Identification and Recovery Agent), was available to provide assistance to the operator during the presence of system faults. It provided five different levels of support (see Table 1 for a detailed description of each LOA, including the information processing stages covered), which could be modified at any time by the operator to fit best his/her needs for task completion (i.e. free choice in selecting the automation level).

2.4. System reliability

System reliability was manipulated across the three experimental conditions by varying the percentage of correct diagnoses made by the support system. In the high-reliability condition, each diagnosis proposed by AFIRA was correct (100% reliability); in the medium-reliability condition, 80% of the diagnoses were correct; and in the low-reliability condition, there were 60% correct diagnoses. When the support system made a wrong fault diagnosis, the incorrect fault diagnosis chosen had a sufficiently similar symptom pattern to the actual system fault (e.g. when an O₂ leak occurred, AFIRA indicated an O₂ block) to ensure that it was sufficiently plausible, and therefore challenging, for the operator to detect AFIRA failures. Please note that there are several elements of system reliability, such as accuracy of fault detection, accuracy of fault diagnosis, and accuracy of implementing remedial action. Each aspect corresponds to a component of AFIRA that could fail. In the present study, only accuracy of fault diagnosis was manipulated whereas the other components were perfectly reliable.

2.5. Dependent measures

Measures recorded by AutoCAMS were supplemented by several questionnaires. Taken together, our evaluation methodology covered the areas of automation management, primary and secondary task performance, trust, self-confidence, and operator workload.

Automation management. Three dependent variables measured different aspects of automation management: (a) Automation reliance indicates the average LOA chosen; (b) the frequency of LOA changes indicates the degree of stability of the LOA chosen; (c) The level of manual control activity indicates the frequency of use of the different manual controllers and the frequency of manual intervention per minute.

Performance. Operator performance on the primary and secondary tasks was measured by five indicators. For the primary tasks, two indicators of diagnostic performance were taken: (a) the percentage of correctly diagnosed and repaired faults (i.e. fault identification accuracy) and (b) fault identification time, which referred to the time required by the operator to initiate the correct repair. The third indicator of primary task performance was (c) the stability of system state, which referred to the percentage of time that any parameter deviated from its target range. For the
secondary tasks, (d) prospective memory performance was measured as the frequency of operators failing to take tank level readings (i.e. percentage of omissions). For the probe detection, (e) the time needed to acknowledge the connection between the ground centre and the space shuttle was used.

**Trust.** Two questionnaires were used to measure operator trust in automation. The Checklist of Trust between People and Automation (CTPA) is a 12-item questionnaire (Jian et al., 2000). An item example is “I am confident in the system”. Ratings were made on a 7-point Likert scale (not at all – totally agree). This questionnaire was used after each experimental block to detect possible changes in trust over the course of the experiment. In the present study, the internal consistency of the scale was excellent (Cronbach’s alpha = .94). The Human Computer Trust (HCT) questionnaire represents a more elaborate scale comprising 25 items (Madsen and Gregor, 2000). An item example is “The system performs reliably”. The reported internal consistency is excellent (Cronbach’s alpha = .94). This questionnaire also used a 7-point Likert scale (not at all – totally agree). According to its creators, its validity requires significant experience with the system to be evaluated. For this reason, the HCT was only administrated at the end of the experimental session.

**Self-confidence.** Self-confidence was assessed by two questions adapted from Lee and Moray’s (1994) questionnaire on trust and self-confidence. The questions referred to the two primary tasks of the operator: “How confident were you in your fault diagnosis (very little – a great deal)?”, “How confident were you in your manual control of the system parameters (very little – a great deal)?” In the present study, the internal consistency of the adapted scale was acceptable (Cronbach’s alpha = .74).

---

Table 1

Implementation of the five levels of automation (based on Sheridan and Verplank, 1978) and their corresponding stage of information processing (based on Parasuraman et al., 2000, see also Manzey et al., 2012).

<table>
<thead>
<tr>
<th>Level of automation</th>
<th>Description</th>
<th>Information processing stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No assistance is provided by the system.</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>The system alerts operators of a fault occurrence.</td>
<td>Information acquisition (IAC)</td>
</tr>
<tr>
<td>3</td>
<td>The system alerts operators of a fault occurrence and proposes a diagnosis.</td>
<td>Information analysis and display (IAD) Action selection (AS)</td>
</tr>
<tr>
<td>4</td>
<td>The system alerts operators of a fault occurrence, proposes a diagnosis and a course of action to remedy to the fault.</td>
<td>Action implementation (AI)</td>
</tr>
<tr>
<td>5</td>
<td>The system alerts operators of a fault occurrence, proposes a diagnosis and a course of action to remedy to the fault, and executes the remedial action if operators accept it.</td>
<td></td>
</tr>
</tbody>
</table>

---

Fig. 1. Main interface of the modified version of AutoCAMS 2.0 (Manzey et al., 2008). Subsystems and facilities: (A) parameter in history graph, (B) functional schema of the cabin (with O2/N2 flow schema and flow meter readings, as well as symbols for the different subsystems), (C) targets of flow levels, (D) system clock, (E) “transmission control” icon (secondary task: probe detection), (F) N2 tank level recording facility (secondary task: prospective memory), (G) manual control panel, (H) repair facility, (I) subsystem failure indicator, (J) repair status (indicates the type of repair currently in progress), (K) control panel for selecting levels of automation, and (L) display of Automated Fault Identification and Recovery Agent (AFIRA).
Subjective workload. Operator workload was measured with the NASA-TLX (Hart and Staveland, 1988) that comprises 6 items (mental demands, physical demands, temporal demand, performance, frustration, and effort). In the present study, the internal consistency of the scale was good (Cronbach’s alpha = .86).

2.6. Procedure

After being welcomed in the laboratory, participants were told that the primary aim of the study was to investigate how human operators manage complex technical systems. They were informed that the experiment consisted of two sessions (training and testing), separated by a one-week interval.

2.6.1. Training

The training session lasted approximately 3 h (including a 15-min break) and covered all technical aspects of system management. Training took place in groups of up to four participants, who were supervised by up to three instructors. Each participant worked on his/her own computer. The training time was calculated on the basis of experience gained from previous experiments using AutoCAMS (e.g. Sauer et al., 2012). Participants were trained on six system faults (i.e. leak in the O2 valve, block in the N2 valve, heater stuck on, CO2 scrubber inefficient, dehumidifier stuck on, and malfunction of the nitrogen sensor), which they practised twice during the session. Participants were provided with a fault-diagnosis manual, which outlined in a step-by-step procedure how each fault could be rectified. In the first series of fault scenarios, participants had to operate the system in full manual control (i.e. LOA 1) in order to learn how to manually manage each subsystem, whereas the four other LOAs were practised in the second series.

To establish a high level of trust towards the system, each diagnosis proposed by AFIRA was correct. This meant that each participant experienced a support system of perfect reliability during the training session. This represents a noteworthy departure from our previous studies with AutoCAMS. In these prior studies, operators had been systematically exposed to automation failures during training, and had gained experience in managing the system under such adverse conditions (Sauer et al., 2011, 2012, 2013). As a result, a degree of mistrust had inevitably been instilled during training, which would be expected to influence the level of automation selected by operators in the test phase, when this choice was available. In the present experiment, we deliberately sought to establish the opposite predisposition, in order to assess the degree to which reliance in the prior work had been influenced by experiences in the training regime. In other words, we explicitly set out to assess the impact of unreliability on operators who were not expecting it and were equally inexperienced in handling it. In the present experiment, we expected to observe generally higher levels of reliance, manifest in a preference for a higher LOA.

At the end of the training session, participants were invited to fill in both trust questionnaires (CTPA and HCT) and the self-confidence scale. These data served as a baseline for the measures taken during the testing session.

2.6.2. Testing

The testing session lasted approximately 2.5 h (with a 15-min break) and consisted of a sequence of two blocks, followed by a series of questionnaires. Each block was 39 min long and contained five system faults (three practised and two novel ones). Each system fault had to be repaired within 330 s. This maximal duration was based on experience gained in previous research with AutoCAMS (e.g. Sauer et al., 2012). After this time, the fault was automatically cleared so that a new system fault could be set up. When a system fault was repaired, feedback about the successful repair was given to participants in the form of a green signal issued by AFIRA (at LOA 2 or higher, see Fig. 11). Furthermore, manual controllers were reset to automatic mode. The order in which fault scenarios were presented in each experimental block is shown in Table 2. In the high reliability condition, the support system always provided the correct diagnosis to the operator. In the medium reliability condition, the support system failed in the second fault scenario of each block while in the low reliability condition, AFIRA failed in the second and fourth fault scenarios. Participants were randomly assigned to an experimental condition. The sequence of faults chosen also allowed us to look at the impact of system reliability for both novel (block 1) and practised faults (block 2). Before starting the first block, participants were reminded again to complete all four tasks, but it was emphasised to give priority to system stabilisation and fault repair as the primary tasks. Furthermore, they were informed that they could change the automation level of AFIRA as they pleased. At the end of each block, participants filled in the NASA-TLX, the CTPA, and the self-confidence scale. At the end of the testing session, they completed the HCT. The fault-diagnosis manual given to participants during training was also made available to them during the testing session. Finally, in a debriefing session participants were asked some open questions about their management of AutoCAMS.

2.7. Data analysis

For the analysis of data produced by AutoCAMS, a mixed two-factorial analysis of variance (ANOVA) was carried out to determine the influence of support system’s reliability (3 levels) and fault scenario (10 levels: fault 1 to 10). When the latter factor was significant, planned comparisons were computed, each addressing an important research question: (a) fault familiarity (F1, F3, F5, F6, F7, F10 vs. F2, F4, F8, F9), (b) time-on-task (F1—F5 vs. F6—10), and (c) after-effects of automation failure (no after-effect: F2, F4, F7, F9 vs. after-effect: F3, F5, F8, F10). If the interaction reliability x fault was significant, planned comparisons were computed to explore the source of this, as indicated below. For the questionnaires analysis, a mixed two-factorial ANOVA was carried out with system reliability as between-subjects factor and time-on-task as within-subjects factor. Huynh-Feldt corrections were applied, if necessary, to adjust degrees of freedom.

To ensure that training did not induce undesirable differences in operator trust towards the support system, both trust questionnaires (CTPA and HCT) were administered at the end of the training session. As intended, the trust level was very similar across reliability conditions for both questionnaires (for CTPA, Mlow = 3.8, MMedium = 3.9, MHigh = 3.8, FCTPA(2,36) < 1; for HCT, Mlow = 4.9, MMedium = 4.7, MHigh = 4.9, FHCT(2,36) < 1).

Table 2
Sequence of practised and novel fault scenarios across experimental conditions and blocks. Automation failures are displayed in italics.

<table>
<thead>
<tr>
<th>Fault scenario</th>
<th>System reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
</tr>
<tr>
<td>Block 1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Practised</td>
</tr>
<tr>
<td>2</td>
<td>Novel</td>
</tr>
<tr>
<td>3</td>
<td>Practised</td>
</tr>
<tr>
<td>4</td>
<td>Practised</td>
</tr>
<tr>
<td>5</td>
<td>Novel</td>
</tr>
<tr>
<td>Block 2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Practised</td>
</tr>
<tr>
<td>7</td>
<td>Practised</td>
</tr>
<tr>
<td>8</td>
<td>Novel</td>
</tr>
<tr>
<td>9</td>
<td>Novel</td>
</tr>
<tr>
<td>10</td>
<td>Practised</td>
</tr>
</tbody>
</table>
3. Results

3.1. Automation management

Automation reliance. As the data in Table 3 shows, participants chose an average LOA of 3.8 across the testing session. ANOVA revealed that system reliability did not affect the level of automation chosen, F < 1, $\eta^2_{\text{partial}} = .02$. The analysis also showed that LOA was unaffected by fault scenario, F < 1, $\eta^2_{\text{partial}} = .02$. The interaction between reliability and fault scenario was also not significant, F < 1, $\eta^2_{\text{partial}} = .05$.

Frequency of LOA changes. The analysis revealed that operators took advantage of the benefits of adaptable automation by moderate activity in changing the LOA (see Table 3). The propensity to change LOA was not influenced by system reliability, F < 1, $\eta^2_{\text{partial}} = .03$. Similarly, there was no significant effect of fault scenario, and interaction between the two factors was not significant either, all F's < 1.20, $\eta^2_{\text{partial}} = .04$.

Manual control activity. Overall, operators were active in managing the control systems (see Table 3). During faults, they carried out, on average, 2.4 control actions per minute with the five control panels. As Table 3 shows, the main effect of reliability just failed to be significant, F(2,35) = 3.143, p = .056, $\eta^2_{\text{partial}} = .15$. A significant effect of fault scenario emerged, F(5.9,603.0) = 11.0, p < .001, $\eta^2_{\text{partial}} = .24$. Planned comparisons revealed that manual control was used less frequently during practised (M = 2.8) than novel faults (M = 2.5), F(1,35) = 6.1, p > .05, r = .39. No significant results were to be reported for the planned comparisons carried out on time-on-task and AFIRA after-effects, both F < 1. No significant interaction was observed, F < 1, $\eta^2_{\text{partial}} = .05$.

3.2. Primary task performance

Fault identification accuracy. A one-way ANOVA revealed that the percentage of repaired faults was significantly influenced by the support system reliability, F(2,35) = 4.62, p < .05, $\eta^2_{\text{partial}} = .21$ (see Table 3). Additional analyses revealed that the high reliability group carried out significantly more correct fault diagnoses than the other two, t(35) = 2.19, p < .05, r = .35. Furthermore, it emerged that the medium reliability group showed better diagnostic performance than the low reliability group, t(35) = -2.11, p < .05, r = .34.

Fault identification time. In addition to diagnostic accuracy, diagnostic speed was also analysed. Since a number of participants did not achieve a successful repair, a default value of 330 s (equals maximum duration of a fault state) was entered into the data file for uncompleted repairs. The two-way ANOVA revealed a significant main effect of system reliability, F(2,35) = 12.1, p < .001, $\eta^2_{\text{partial}} = .41$. Simple contrasts indicated that operators were significantly faster to repair faults in the high reliability group than the medium reliability group (M = 129.6 s, t(24) = 2.24, p < .05, r = .42), who were significantly faster than the low reliability group, t(23) = 2.73, p < .01, r = .43 (see Table 3). In addition, a main effect of fault reached significance, F(6.4, 226.3) = 25.3, p < .001, $\eta^2_{\text{partial}} = .42$. Planned comparisons showed a significant effect of both after-effects of incorrectly diagnosed faults and of practice, but the identification time remained constant between the first (M = 132.5 s) and the second block (M = 130.4 s), F(1,35) < 1. Operators identified faults slower when AFIRA may suffer from an automation failure (M = 169.2 s) than during the subsequent fault with a correct diagnosis of automation (M = 108 s), F(1,35) = 44.6, p < .001, r = .75. Moreover, practised faults were faster identified (M = 119.5 s) than novel faults (M = 149.4 s), F(1,35) = 77.7, p < .001, r = .83. Finally, the interaction was significant, F(12,9,226.3) = 4.67, p < .001, $\eta^2_{\text{partial}} = .21$. Planned comparisons confirmed that operators with a low system reliability identified more slowly faults during an automation failure (M = 229 s) than the subsequent (correctly diagnosed) faults (M = 149.1 s), whereas operators in the high reliability group did not show such a large difference between both types of faults (Mpractised diagnosis = 109.8 s; Mcorrect diagnosis = 89.6 s), F(1,35) = 8.43, p < .01, r = .44 (see Fig. 2). Furthermore, although operators in the low reliability group need in general more time in fault identification, the discrepancy between practised (M = 160.7 s) and novel faults (M = 195.0 s) is larger than for the medium (Mpractised = 116.6 s; Mnovel = 149.2 s) and the high reliability groups (Mpractised = 84.4 s; Mnovel = 107.5 s), respectively F(1,35) = 14.7, p < .001, r = .54, and F(1,35) = 7.86, p < .01, r = .43.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>System reliability</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td><strong>Automation management</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average LOA [1–5]</td>
<td>3.8 (0.9)</td>
<td>3.6 (0.8)</td>
</tr>
<tr>
<td>Frequency of LOA changes [no/fault]</td>
<td>1.6 (2.4)</td>
<td>1.1 (2.0)</td>
</tr>
<tr>
<td>Manual control activity [no/min]</td>
<td>2.7 (0.5)</td>
<td>2.8 (1.0)</td>
</tr>
<tr>
<td><strong>Primary tasks</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fault identification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy [%]</td>
<td>83.3 (1.2)</td>
<td>86.9 (8.6)</td>
</tr>
<tr>
<td>Time [s]</td>
<td>174.4 (79.0)</td>
<td>129.6 (71.5)</td>
</tr>
<tr>
<td>Unstable system state [%]</td>
<td>5.8 (4.8)</td>
<td>5.2 (4.2)</td>
</tr>
<tr>
<td><strong>Secondary tasks</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omission in prospective memory [%]</td>
<td>18.2 (19.4)</td>
<td>9.9 (14.0)</td>
</tr>
<tr>
<td>Reaction time in probe detection [s]</td>
<td>2.1 (0.7)</td>
<td>2.3 (0.9)</td>
</tr>
</tbody>
</table>
Management of unstable system state. Parameters deviated from their target state on average 5.10% of the time during a fault state. In contrast to performance in fault diagnosis, manual control performance was not affected by system reliability, F(1, 315) = 8.60, p < .001, ƞ²_p = .20. Planned comparisons showed, however, no significant effects of time-on-task, practice, and after-effect of AFIRA failure, all F’s < 1. No interaction was recorded, F < 1, ƞ²_p = .05.

3.3. Secondary task performance

Prospective memory. Concerning the percentage of omissions for this task (see Table 3), the main effect of reliability was significant, F(2,35) = 3.46, p < .05, ƞ²_p = .17. Operators in the low reliability group (M = 18.2%) failed to report the N₂ tank level significantly more often than the medium (M = 9.9%) and high reliability groups (M = 8.3%), respectively t(23) = 2.06, p < .05, r = .40, and t(24) = 2.46, p < .05, r = .45. The main effect of fault state was significant as well, F(9,315) = 3.94, p < .001, ƞ²_p = .10. Planned comparisons showed an effect of time-on-task and direct vs. after-effect of a support system failure, respectively F(1,315) = 10.5, p < .005, r = .48, and F(1,315) = 6.06, p < .05, r = .38, but no effect of practice, F(1,315) < 1. Operators failed more frequently to complete the task in the first (M = 14.4%) than in the second block (M = 9.8%). The percentage of misses was larger during an automation failure (M = 14.8%) than during the subsequent fault, for which AFIRA suggested a correct diagnosis (M = 10.7%). Finally, the interaction between reliability and fault scenario reached significance, F(18,315) = 2.48, p < .05, ƞ²_p = .12. Planned comparisons for time-on-task effects indicated that operators of the low reliability group had a larger decrease in their percentage of misses from the first (M = 22.6%) to the second block (M = 13.8%) than the medium and high reliability groups together (first block: M = 10.3%; second block: M = 7.9%), F(1,35) = 4.50, p < .05, r = .34 (see Fig. 3A). Interestingly, operators of the low-reliability group showed a greater proportion of misses during an automation failure (M = 23.8%) than during correct diagnosis of the automation (M = 14.2%) whereas the medium and high reliability groups had comparable proportions of misses in each case (support system failure: M = 10.4%; subsequent fault: M = 9.0%), F(1,35) = 5.32, p < .05, r = .36 (see Fig. 3B).

Probe detection. Only fault scenario significantly affected reaction time in the probe detection task, F(7,4,260.5) = 2.29, p < .05, ƞ²_p = .06, however no significant effects were revealed by the planned comparisons. The main effect of reliability and the interaction did not reach the significance level, both F < 1, ƞ²_p < .05. The descriptive data for the three reliability conditions are presented in Table 3.

3.4. Trust and self-confidence

Checklist of trust between people and automation. The analysis of the 12-item trust questionnaire (which was administered after each block) showed a significant main effect of reliability, F(2,36) = 5.71, p < .01, ƞ²_p = .24. Trust in the support system was significantly larger in the high-reliability group (M = 5.5) than in the low-reliability group (M = 4.1), t(24) = −3.05, p < .05, r = .53, and marginally larger than in the medium-reliability group (M = 4.6), t(24) = −2.17, p < .10, r = .41. No significant difference was observed between the low and medium-reliability condition.

No significant effect of time-on-task was found and no interaction was observed, F(2,36) = .68, p > .10, ƞ²_p = .04.

Human computer trust questionnaire. The more elaborate 25-item questionnaire (completed once at the end of the testing session) also found a significant effect of reliability, F(2,36) = 6.13, p < .01, ƞ²_p = .25. The high reliability group (M = 5.4) trusted the support system significantly more than the low-reliability group (M = 4), t(24) = 3.18, p < .05, r = .54. However, no significant difference was found between the medium-reliability group (M = 4.6) and the two other groups.

Self-confidence. A 3 × 2 × 2 mixed ANOVA was used to determine the impact of system reliability (3 levels), time-on-task (2 levels), and primary task (2 levels) on operator self-confidence. This last factor was not influence by system reliability; main effect: F(2,37) = 1.64, p > .10, ƞ²_p = .08; interaction reliability x time-on-task: F(2,37) = 3.05, p > .05, ƞ² = .14; reliability x primary task, F(2,37) = .18, p > .10, ƞ²_p = .001; 3-way interaction: F(2,37) = .13, p > .10, ƞ²_p = .001. However, operators felt more confident in manual control of air parameters (M = 7.6) than in fault diagnosis (M = 7.1), F(1,37) = 6.68, p < .05, ƞ²_p = .15. Furthermore, this difference increased over time, F(1,37) = 4.78, p < .05, ƞ² = .11. Detailed analyses revealed that self-confidence in manual control increased from Block 1 (M = 7.36) to Block 2 (M = 7.9), F(1,37) = 12.26, p < .001, ƞ²_p = .25, whereas it remained stable for fault diagnosis (M_block1 = 7.16, M_block2 = 7.12), F(1,37) = p > .10, ƞ²_p = .02. Finally, the main effect of time-on-time was not significant, F(1,37) = 1.33, p > .10, ƞ²_p = .04.

3.5. Operator workload

Although workload appeared to decrease, as expected, with increasing levels of system reliability (M_low = 8.5; M_medium = 7.6; M_high = 6.4), this difference was not significant, F(2,36) = 1.06, p > .10, ƞ²_p = .06. The main effect of time-on-task and its interaction did not reach the level of significance either, respectively F(1,36) = 1.23, p > .10, ƞ²_p = .03, and F(2,36) < 1, ƞ²_p < .01.

4. Discussion

The aim of the present study was to evaluate operator performance, trust, and use of adaptable automation under different reliability levels. Overall, operators managed the system rather effectively in the three reliability conditions, but there were
decrments in diagnostic speed and prospective memory with lower reliability, with the latter effect suggesting that performance was being maintained at the cost of higher mental workload. Although reduced system reliability resulted in lower levels of trust towards automation, there were no corresponding differences across the three groups in operator reliance on automation, suggesting an interesting dissociation between attitude and behaviour. This shows that operators have difficulty in accurately calibrating their level of trust according to the actual reliability of the system. Finally, the results showed a higher reliance on automation compared to previous work using the same simulation environment (e.g., Sauer et al., 2012).

Automation failures had a direct effect on operator performance in the form of longer fault identification times and more omissions for the prospective memory task. These observations generally confirmed the impact of system reliability on operator performance observed in the literature (e.g., Bailey and Scerbo, 2007; Dzindolet et al., 2003; Ma and Kaber, 2007; Moray et al., 2000; Rovira et al., 2007). It indicates that operators in the low-reliability group were more cautious (i.e., they took more time) before completing the correct repair than the high-reliability group. The fact that reliability neither influenced system stabilisation performance nor manual control activities suggests that operators under low-reliability automation accomplished the primary task of maintaining parameters in the target range equally well as operators in the other conditions. Furthermore, these findings are consistent with the nature of automation failure since performance decrements were confined to fault diagnosis which was the only function of the support system AFIRA that failed. This focus of attention on primary task duties might have led operators under low-reliability automation to neglect secondary tasks, especially prospective memory. Prospective memory as a complex task involving multiple steps is more cognitively demanding than probe detection and hence more vulnerable to interferences. Previous work often found a higher sensitivity of prospective memory performance than the probe detection task (e.g., Hockey et al., 2007). Overall, the results suggest that the maintenance of primary task performance under conditions of unreliability is achieved at the cost of higher cognitive workload.

The effects of system reliability on performance became more visible when examining the so-called after-effects of AFIRA failure. The planned comparisons revealed that reduced reliability produced these after-effects in that it also affected performance during the fault states immediately following the AFIRA failure. The fault states following the AFIRA failure are particularly diagnostic because all groups experienced exactly the same scenario under a fully reliable AFIRA. The first evidence for such after-effects of automation failure stems from the prospective memory data. The low-reliability group showed about 50% more prospective memory errors than the other two conditions. The preceding AFIRA misdiagnosis may have led the low-reliability group to expend more cognitive resources on the diagnostic task in expectation of another AFIRA misdiagnosis (which however did not occur). While this shift of resources to the primary task may be a suitable strategy, it was at the cost of secondary task performance (cf. Hockey, 1997). After-effects also manifested themselves in diagnostic speed, as the low-reliability group showed about 50% longer fault identification time than the high-reliability group.

A central feature of this study was the measurement of operators’ reliance on automation. Contrary to studies employing a dual automation mode (e.g., Moray et al., 2000; Riley, 1997), the set-up of the present study allowed operators to choose at will between five incremental LOAs. Surprisingly, operators did not adjust the LOA according to the actual reliability of the system, but opted rather for a high and constant LOA (around level 4) throughout the testing session. These observations are comparable to the results of Sauer et al. (2011, 2012) but contradict other findings, which showed a clear preference towards manual control under low reliability automation (Moray et al., 2000; Muir and Moray, 1996). Opting for a rather high LOA seems to be a sensible strategy even under low-reliability conditions given the technical performance of the adaptable support system AFIRA since the strategy seems to have optimised the balance of automation benefits and costs (see Manzey et al., 2012). The support system always provided a correct indication of the on-set of a system fault and also of the course of action to be taken to remedy it. Reduced system reliability meant that the type of system fault was sometimes misdiagnosed while on-set of fault and course of actions proposed were always correct. This suggests that operators may have still considered AFIRA to be a helpful tool overall even if one of its components occasionally failed.

Training may have also contributed (and perhaps even more strongly than the nature of automation failure) to the persistence of high automation reliance in the present study. In this experiment, there was no indication during training that the automatic systems could be fallible. This was in contrast to earlier experiments, in which participants experienced several automation failures in the training session (e.g., Sauer et al., 2012). Three previous experiments within the same simulator environment consistently showed that operators tended to opt for a lower LOA under adaptable automation, also offering free choice between LOA 1–5 (2.32 [Sauer et al., 2013], 2.62 [Sauer et al., 2011], and 3.08 [Sauer et al., 2012] compared to a LOA of 3.8 for the current study). The difference in the training regimes may not only have explained why operators chose a higher LOA in the present study, but also why they stayed with a high LOA throughout the testing session. It suggested that training had created a behavioural norm, developing into habitual behaviours which stabilised with continuous practice (cf. Aarts et al., 1998). This behavioural norm was further reinforced by the operators’ lack of prior experience in handling automation failures, which curtailed their ability to cope with system faults.

These findings are consistent with other studies which show the relative stability of automation management behaviour in the face of fluctuations in system reliability (Lee and Moray, 1994; Dzindolet et al., 2003; Manzey et al., 2012). In contrast to the behavioural measures, the subjective measure of trust was sensitive to automation reliability. Furthermore, as in previous research (e.g., Wiegmann, 2002), operators tended to under-trust the system according to automation reliability, namely when the system was 100% reliable, the trust rating was not maximal. However, it is possible that operators needed more time to build up trust to its maximum level. For example, Manzey et al. (2012) showed that trust increases only slowly in small steps when the system operates fully reliably (but plummets to a much lower level in the event of an automation failure). It can also be assumed that apart from reliability other relevant factors such as ability and predictability of automation (Lee and Moray, 1992) will have had an influence on trust ratings. The reduced levels of trust in the present study were also reflected in the operators’ reluctance to rely completely on an automated support system, as indicated by an automation reliance measure that did not reach its maximum level. This behaviour might signal either mistrust (Lee and Moray, 1994), or the will to keep control over the machine by staying involved in the process (e.g., Sauer et al., 2011). Moreover, as in Moray et al. (2000), operators displayed a constant level of self-confidence in system management in all three experimental groups.

The results of the present study reiterated previous research findings with regard to the influence of automation reliability on operators’ performance, trust and use of automation (e.g., Lee and See, 2004). Broadly speaking, we observed a complex pattern of effects, in which behavioural and attitudinal elements do not
arrange themselves in a simple causal order. Whereas operators are sensitive to automation misdiagnoses, reductions in trust do not directly manifest themselves in objective behaviour such as automation reliance (i.e. mean LOA selected). The kind of relationship between trust and behaviour is thus a complex one, which will depend on the characteristics of the work environment (e.g. Hancock et al., 2011) and other factors, such as the behavioural norms created in training. The fact that secondary task performance suffered from unreliable automation in the form of misdiagnoses raises the question of whether adaptive automation (i.e. machine controlled) would be a better alternative. While there is one study examining system reliability in adaptive automation, it used event-based adaptive automation (Moray et al., 2000). This might not be as good an option as performance-based adaptive automation to deal with reduced system reliability because the latter is better tailored towards the current operator state. There is a need to address the impact of system reliability also in the context of the various types of flexible automation by empirically examining their effects in the form of controlled experiments using complex simulations of real technical work environments.

Moreover, further research may be needed to focus on the impact of the precise nature of automation failures. In the present study, system reliability was manipulated by the frequency of misdiagnoses of the system. However, failure in a late information processing stage (e.g. at the level of action selection or action implementation) or a complete failure of the system already in an early stage (e.g. during information acquisition) might produce different patterns in operation performance, trust, and use of adaptable automation. Also, an increase in manual control activities may be predictable if an incorrect course of action is being proposed by the automation in order to stabilise the system again. Conversely, faults may remain undetected during complete automation failure if operators rely solely on automation (automation misuse, see Parasuraman and Riley, 1997). As a consequence, the appropriate measures to solve the situation may be not taken and therefore performance would drop. Furthermore, to better understand the concept of trust, the system should not only be rated as a whole, but its components should be rated separately. The present study already made a first attempt into this direction by measuring self-confidence in fault diagnosis and system stabilisation as two separate system functions. This also suggests a need to use instruments that encompass a larger number of facets of trust in human-machine systems. Ergonomics 35 (10), 1243–1253.


A. Chavaillaz et al. / Applied Ergonomics 52 (2016) 333–342


