To look and (not) see: Predicting the detection of automation failures based on the eye movements of human operators

Carmen Bruder
German Aerospace Center (D.L.R.) Aviation and Space Psychology
Hamburg, Germany
+49 40 513096 61
carmen.bruder@dlr.de

Paul Weber
German Aerospace Center (D.L.R.) Aviation and Space Psychology
Hamburg, Germany

Catrin Hasse
German Aerospace Center (D.L.R.) Aviation and Space Psychology
Hamburg, Germany
+49 40 513096 68
catrin.hasse@dlr.de

ABSTRACT
In order to detect automation failures in time, operators are required to monitor automated systems efficiently. The present study analyzed eye movements to predict whether or not subjects would detect an automation failure. Eye movements were recorded whilst subjects were monitoring an automated system where automation failures sometimes occur. The data imply that the eye movements of human operators effectively differ between operators who detect the automation failure and those who miss it. The findings are discussed in the context of personnel selection and incident reporting as used in air traffic control (ATC).

Keywords
Automation, monitoring, automation failures, eye movements, aviation, personnel selection

INTRODUCTION
Human-machine systems in aviation are mostly automated in many aspects. According to research on the future of aviation, such as the Single European Sky ATM Research (SESAR) Program, such developments mean the nature of human roles and tasks will inevitably change [22, 23]. Workshops with experienced pilots and air traffic controllers, which were conducted in order to gather their expectations about future tasks, roles, and responsibilities, indicated that monitoring in a highly automated workplace will pose challenges to future operators in aviation [5].

Automation failures occur even though the systems are becoming more and more reliable. If an automation failure happens, the human operator is asked to detect this failure efficiently. That is one of the ironies of automation as described by Bainbridge [1]. If the operator fails to identify the system failure, serious consequences can occur.

With respect to personnel selection, training, and incident reports, reliable eye-movement indicators could be useful for predicting whether or not a human operator perceives the automation failure. To select candidates who supervise automated systems efficiently, physiological indicators for failure detection performance would be a perfect complement to current tests based on behavioral indicators.

Finally physiological indicators could provide additional information in incident reports as used in ATC to classify human errors [e.g. 24, 19], to shed some light into the “black box” of cognitive processes. For example, eye movement indicators would help to classify the human error according to the Rasmussen performance levels [18], i.e. whether an automation failure was not perceived or it was perceived but not detected or reported. Lundberg et al. used unobtrusive eye tracking within a field experiment to explain loss of separation events over Swedish air space in 2011 [17]. They suggested that it is hard for air traffic controllers to handle the trade-off between spending visual scan time on conflict detection and simultaneously examining potential conflicts identified by the conflict detection automation [17].

Detecting automation failures
Automation failures occur in many ways. To name just a few of the failure types, automated systems can fail to perform an action, execute an action incorrectly, or do so at a wrong time. Analogously, Hollnagel (2000) defined categories of actions, such as correctly performed actions, erroneous execution, commission, and omission [14]. As the task of human operators is to supervise an automated system, errors of omission and commission can occur. Hollnagel (2000, p. 138) defined an error of omission as “the failure of carrying out an action during the time window when it was required” [14]. An error of commission is when an operator performs an inappropriate action to achieve the goal.

With respect to highly automated systems, an error of omission happens when the human operator does not detect the automation failure in time. Since human attention is limited and human-machine systems are often complex, human operators need to monitor the automated system, i.e.
allocate their attention adequately and switch between targets dynamically.

**Monitoring automated systems**

Billings (1997, p. 198) explained that operators in aviation “must closely monitor the behavior of their automated systems, but if an anomaly occurs, they must sometimes take very prompt action [2].” The increase in automation requires operators who can monitor in such a way as to enable them to detect automation failures in time and to take control if automation fails [8]. That is very important for situations where the automated system provides no warning when an automation failure occurs.

Wickens et al. [27] concluded that automation might affect system performance as new skills may be required, and that human operators might not have been adequately selected and trained for these changes. Therefore in order to prepare for the future selection and training of aviation operatives, it is imperative to understand the task of monitoring. Operational monitoring comprises using one’s senses to follow meaningful information from an automated system responsibly, even when there is no direct need for action. It involves being prepared to take control of a system at any time, for example in case of malfunction [8].

**Measuring monitoring with eye movements**

A variety of studies support the idea that eye movements offer an appropriate means for measuring the efficient and timely acquisition of visual information [e.g. 9, 16, 25, 26, for an overview see 15]. Based on this research, eye movement parameters were identified which reflect the human monitoring performance [10]. These parameters were associated with accurate failure detection [11], accurate manual control in case of automation failure [13]. Hasse et al. examined monitoring behavior as a team task [12]. Bruder et al. [4] investigated the link between eye movement parameters and the monitoring behavior of experts, and also compared the monitoring behavior of experts with novices [5]. Grasshoff et al. [10] predicted more than 53% of the variance of the failure detection performance by means of eye tracking parameters. When comparing existing selection procedures with eye tracking measurements of job applicants it could be shown that monitoring performance refers mainly to something like a complex skill, requiring and incorporating attentional planning processes [10].

Though the results of previous studies revealed valuable insights into eye movements during the process of monitoring, they failed to show how the scanning behavior of operators during the detection of automation failures can be described in contrast to situations where operators fail to detect automation failures. In the present study, the eye movement behavior was traced in more detail, i.e. with respect to specific monitoring behavior around the automation failure.

In the present study, monitoring patterns for accurate failure-detection and missed automation failures were compared. In principle, this combination can arise in four different forms:

- If an automation failure is monitored and detected, the eye movement patterns for successful failure detection are identified.
- If the automation failure is not monitored and not detected, eye movement patterns for the error of omission are found.
- If the automation failure is monitored but not detected, it is the looking-but-not-seeing effect.
- Finally, the automation is not monitored but detected, which is caused by peripheral target detection [7].

The looking-but-not-seeing effect as a principle of inattentive blindness is a major side effect of top-down processing, where targets unrelated to the specific task are not monitored [20]. The looking-but-not-seeing effect could result from performing parallel tasks, which is common when dealing with complex automated systems. Human operators performing another task look but do not recognize information unrelated to the task they are currently occupied with [6]. The looking-but-not-seeing effect could also be a side effect of top-down processing, as human operators overlook a target if they do not expect it [21].

With the looking-but-not-seeing effect and the peripheral target detection in mind, the following research questions should be addressed here:

1. To what extent can the detection of automation failures be predicted by the corresponding eye movements?
2. How strong is the looking-but-not-seeing effect?

**METHOD**

An empirical study requiring the monitoring of an automated system was undertaken with applicants for jobs in aviation.

**Simulation tool**

A simulation tool called Monitoring Test (MonT) was designed to measure the monitoring performance of applicants for future jobs in aviation who have no prior experience as a pilot or an air traffic controller. For this purpose, a simplified and abstract simulation of traffic flow was developed. MonT therefore depicts the basic processes involved in monitoring rather than presenting an ecologically valid traffic situation. The same monitoring abilities are, however, required to complete the tasks successfully (see Figure 1).

As this tool presents a simplified and abstract simulation of traffic flow, it was appropriate for use with test subjects who have no prior experience as a pilot or air traffic controller. The simulation consisted of two separate systems namely the top system and the bottom system, which were fully identical but operated entirely independently of one another. They were both represented on the same screen. Each system contained two areas that
represented storage space where an actual and a target value were displayed, as well as two entries and two exits.

Objects entered the system through one of the entry points and then moved along the connections between the entry, area, and exit. These connections consisted of segments. Objects moved forward one segment per time unit, which comprised three seconds. Objects entering an area were either stored in this area or were forwarded to the corresponding connection. In both cases the current value of the area was updated automatically.

The objects were controlled automatically, i.e. the decision whether an object should enter an area or move to a corresponding connection was made by the automated system. The automated system indicated the number of objects that should be forwarded by using the input fields.

The task of the automated system was to bring all actual values into agreement with target values. Automation failures could happen whenever objects entered an area. Three types of failures were distinguished:

- The automated system failed to send objects onward.
- The automated system sent the wrong number of objects.
- The automated system sent objects at the wrong time.

**Task of the test subject**
The test subject’s task was to monitor the automated system and report any automation failures as soon as they were recognized. In order to decide whether or not a failure had occurred, the test subject had to learn the rules of the automatic system as well as the different types of failures that could happen (see above). The test subject was required to report failures by clicking on the input fields that were incorrectly edited by the automated system. Frequency and response time were analyzed.

**Design of monitoring scenarios**
For experimental purposes, as well as for the long-term objective of using the scenarios for personnel selection and training, scenarios presented a variety of failures rather than simulating real conditions. The monitoring task consisted of six scenarios; a single automation failure happened in two scenarios and three or four automation failures happened in two scenarios. In the other two scenarios, the automatic system worked accurately (two distractor scenarios).

Each scenario lasted two to three minutes, beginning with an orientation phase where the display was frozen. The duration of the orientation phase was fixed and lasted 15 seconds. After finishing the orientation phase, the simulation started flowing dynamically.

**Eye-tracking equipment**
Each subject was seated in front of a 19-inch LCD computer display at a distance of approximately 60 cm (see Figure 2). The LCD display had a refresh rate of 60 Hz, contrast ratio of 2000:1, lag of 2ms, and a resolution of 1280 x 1024 pixels. Eye movements were remotely recorded by the binocular Eyegaze Analysis System and the ‘Bright Pupil Method’, manufactured by LC Technologies, Inc. The system operated at 120 Hz and was combined with the simulation tool MonT to ensure that both systems used the same timestamp.

**Participants**
The experiment was conducted with a sample of 101 applicants, 71 for air traffic control training at DFS (Deutsche Flugsicherung GmbH) and 30 for pilot training at DLH (Deutsche Lufthansa). They were 17 to 27 years old (M=20.00, SD=2.36) and 69.5% were male. 84% claimed to have experience with computer games. The computer experience was rated on a Likert scale from 1 (no experience) to 5 (very experienced) with an average of 3.9 (SD=.67, ‘experienced’). Experiments were conducted in conjunction with the regular selection process at the German Aerospace Center without influencing the selection outcome. Subjects received 25 € for their participation in the two hour experiment.

**Procedure**
Two subjects performed the experiment at the same time, but each with a separate computer and eye-tracking system. A room divider was installed between the subjects to
prevent visual contact between them. After providing subjects with a demographic questionnaire, they were instructed to perform the monitoring task. Before the monitoring scenarios were presented, subjects performed a knowledge test and four exercises. At the end, the subjects completed two questionnaires measuring their complacency potential and their attitude towards technology. Finally, subjects were asked about their impressions of the experiment.

**Measurements**

Two instances of failure detection were separated: A failure was detected successfully if a subject reported this failure by clicking on the input fields that were incorrectly edited by the automated system. A failure was missed if the subject did not report the failure within the corresponding timeframe.

**Defining relevant AOIs**

The definition of relevant AOIs was directly connected to the automation failures and the operations of the automatic system that encounter an automation failure (see Figure 3). Perceiving these AOIs at the right time should indicate ideal monitoring behavior and provide the subject with the ability to detect automation failures. As a first step, AOIs were determined for every monitoring scenario. AOIs included connections, input fields, entry values, actual and target values. Then, for each automation failure within a scenario, selected AOIs were defined as relevant (‘relevant AOIs’). Thus, the status of AOIs changed from being irrelevant to relevant.

![Figure 3](image)

**Figure 3. Definition of all AOIs (dark grey) and AOIs defined as relevant (light grey) exemplified by a time unit an automation failure occurred in the right area of the bottom system**

The definition of time frames for the organization of eye-tracking data was guided by instances of automation failure. Figure 4 shows how instances of automation failure are related to corresponding eye movements. Once the automatic system began to control the traffic dynamically, eye movement parameters for all stages of an automation failure were defined — before, during, and after the automation failure occurs: two time units before (-2), one time unit before (-1), when the automation failure occurs (0), one time unit after (+1), and two time units after the automation failure (+2). Additionally, the eye tracking data on relevant AOIs within the orientation phase (OP) were analyzed separately.

![Figure 4](image)

**Figure 4. The link between system events and the corresponding eye-tracking data exemplified by a scenario with two automation failures**

**Eye Tracking Parameters**

In relation to the predefined AOIs and monitoring phases, the following eye tracking parameters were analyzed:

- Total time until the first fixation: Time (in ms) until the first fixation falls on a relevant AOI within a time unit (tttff).
- Total fixation count: Number of fixations on relevant AOIs (tfc).
- Relative fixation count: The ratio between the number of fixations on relevant AOIs and all fixations within a time unit (rfc).
- Total mean fixation duration: Mean fixation duration on a relevant AOI and all mean fixation duration within a time unit (tmfd).
- Relative mean fixation duration: The ratio between mean fixation duration on a relevant AOI and all mean fixation duration within a time unit (rmfd).
- Total gaze duration: Total duration of all gazes on relevant AOIs (tgd).
- Relative gaze duration: The ratio between gaze duration on relevant AOIs and all gaze durations within a time unit (rgd).

Relative parameters ranged from 0 to 1, with 0 indicating that no eye movements fell on predefined AOIs in relation to all eye movements within a time period, and with 1 indicating that all eye movements fell on the predefined AOIs within a time period.

**RESULTS**

Data from 101 subjects were reported, each of whom processed six scenarios with a total of nine automation failures. In sum, eye-tracking data and failure-detection data from 761 automation failures were included in the
Predictability of automation failure detection by eye tracking data

To address the first research question, a logistic regression analysis was conducted to predict the detection of automation failures when monitoring automation failures (using the eye movements on relevant AOIs as a predictor). In the contingency table (see Table 1) the outcome variable “failure detection performance” (failure reported versus failure not reported) is paired with the predictor variable “monitoring of automation failure” (automation failure detected or not detected) predicted by the eye movements on relevant AOIs. The regression model predicted the failure detection performance using seven eye tracking parameters (ttff, tfc, rfc, tmfd, rmf4, tgd, rgd) and six stages of an automation failure (OP, -2, -1, 0, +1, +2).

A test of the full model against a constant only model was statistically significant, indicating that the monitoring of automation failures measured by eye movements on relevant AOIs as a set reliably distinguished between accurate failure-detection and failures not detected (chi square = 288.227, p < .001 with df = 42). Nagelkerke’s R² of .422 indicated a moderately good relationship between prediction and grouping. Prediction success overall was 76.1% (78.8% for successful detection and 73.4% for missed failures).

Looking-but-not-seeing effect

To take a closer look to the looking-but-not-seeing effect, accurate failure-detection and missed failures were paired with relative fixation counts on relevant AOIs, but only for the time unit in which the automation failure occurs (TU=0). Table 2 contains the contingencies between failure-detection performance and monitoring the automation failures.

A test of the full model against a constant only model was statistically significant (chi square = 156,866, p < .001 with df = 1). Nagelkerke’s R² of .248 indicated a low correlation between prediction and grouping. Prediction success overall was 67.3% (97.0% for success and 37.1% for missing).

Table 2. Contingency table consisting of failure detection performance paired with “monitoring of automation failure” measured by relative fixation counts only for the time unit in which the automation failure just occurs (TU=0).

<table>
<thead>
<tr>
<th>predictor</th>
<th>monitoring of automation failure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>non eye movements on relevant AOIs</td>
</tr>
<tr>
<td>failure detection performance</td>
<td>failure NOT reported</td>
</tr>
<tr>
<td>automation failure NOT detected</td>
<td>276</td>
</tr>
<tr>
<td>automation failure detected</td>
<td>81</td>
</tr>
</tbody>
</table>

The data show substantial differences between accurate failure detection and missed failures. Using monitoring behavior to predict accurate failure detection was very reliable: 97 percent of accurately detected failures were able to be predicted from eye movement patterns. However, it was not possible to predict missed automation failures from eye movement patterns: 62.9 percent of missed automation failures were monitored, e.g. eye movements were measured on relevant AOIs when the automation failure occurred, but the failure was not reported.

DISCUSSION

To address the differences in monitoring behavior between detected automation failures and those that were missed, this study analyzed monitoring behavior within a simplified and abstract simulation. In the present study, the eye movement behavior was be traced in detail, i.e. with respect to specific monitoring behavior around the time that the automation failure occurs.

About three quarters of automation failure detections were predicted successfully by the corresponding eye movements. If the eye movement patterns for successful failure detection occurred, the automation failure was monitored and detected. Similarly, if eye movement patterns related to the error of omission were found, the automation failure was not monitored and not detected.

This leads to the conclusion that human operators’ monitoring patterns are a useful indicator for their success in the detection of automation failures. Using MonT in combination with eye tracking data provides a useful, objective test for future personnel selection. As eye movements are operational measures of immediate cognitive processes [16], the test outcome cannot be faked. However about one fifth of the automation failures were not classified correctly from the corresponding eye movements.

**Table 1.** Contingency table consisting of outcome variable “failure detection performance” paired with predictor variable “monitoring of automation failure” measured by eye movements

<table>
<thead>
<tr>
<th>predictor</th>
<th>failure detection performance</th>
<th>monitoring of automation failure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>automation failure NOT detected</td>
<td>eye movements on relevant AOIs</td>
</tr>
<tr>
<td>automation failure NOT reported</td>
<td>140</td>
<td>237</td>
</tr>
<tr>
<td>automation failure detected</td>
<td>12</td>
<td>372</td>
</tr>
<tr>
<td>overall percentage</td>
<td>67.3</td>
<td></td>
</tr>
</tbody>
</table>
movements. Two types of discrepancies between monitoring behavior and detection performance happened:

1. The automation failure was NOT monitored but detected.
2. The automation failure was monitored but NOT detected.

Discrepancy 1 could be seen in light of peripheral target detection [7], i.e. eye tracking measures fixations, and therefore does not capture peripheral vision. Discrepancy 1 may also be affected by methodological shortcomings concerning the definition of relevant AOIs for failure detection. As mentioned in the measurements section, it was defined WHEN (in which specific time unit) human operators should pay attention to WHAT (which certain areas) to detect the automation failure. Therefore, discrepancy 1 can be an effect of the individual differences in the timing of fixating on relevant information. Thus, the predictive value should be improved by adjusting the definition of AOIs which are relevant to detect automation failures.

Discrepancy 2 can be interpreted as the result of the looking-but-not-seeing effect [20]. The human operator looks at information relevant for identifying an automation failure, but fails to detect the automation failure. A reason could be that human operators overlook the automation failure if they do not expect an automation failure. In line with Maples et al. (2008), this may reflect a side effect of top-down processing based on the expectations made during the orientation phase [21]. The looking-but-not-seeing effect could also be caused by parallel tasks which are typical for complex human-machine systems. Human operators performing another task look at but do not recognize information unrelated to the task they are currently occupied with [6]. In the present study, this could be the case, as the participants had to monitor an automated system that sometimes had overlapping actions.

However, when predicting failure detection using eye tracking data only for the time unit when the automation failure occurs, the proportion of discrepancy 1 decreases (to under 5%), but the proportion of discrepancy 2 increases (to over 60%). This clearly shows that it is important to include the monitoring behavior before, during, and after an automation failure in order to reliably predict failure detection. This is noteworthy considering that missing an automation failure (discrepancy 2) has more severe consequences than discrepancy 1.

OUTLOOK
Predicting failure detection by using eye tracking within dynamic simulations is an innovative strategy and enables the development of new approaches for personnel selection, training, and incident reports. Learning from the differences in monitoring automated systems between successful and unsuccessful failure detection will be helpful in learning how to select successful trainees and providing them with appropriate training.

Nevertheless, discrepancies between eye-tracking indicators and failure detection shows still some methodical shortcomings of predicting automation failures based on human operators’ eye movements. Further research will improve the reliability of eye-movement indicators by adjusting the definition of information that is relevant for detecting automation failures. This includes WHEN (in which specific time unit) human operators should pay attention to WHAT (which certain areas) to detect the automation failure. These improvements further calibrate the diagnostic relevance of eye tracking when predicting the detection of automation failures based on human operators’ eye movements.

REFERENCES


