



Triggering Cockpit Alerts Using an Eye-Tracking-Based Measure of Monitoring Performance

Simon Schwerd¹ and Axel Schulte

Institute of Flight Systems, University of the Bundeswehr Munich, Germany

Abstract: This study explored the potential for enhancing pilot performance via an alerting system that adapts according to an eye-tracking-based measure of monitoring. The novel measure combines gaze analysis with system state assessment to estimate the pilot's understanding of the current system state. On this basis, an alerting system was developed to direct pilot attention to unnoticed system state changes, thereby improving system state monitoring. In a flight simulator study involving 10 participants in a generic jet cockpit, we compared the adaptive alerting system with a no-assistance condition. Although alerting improved the participants' performance in two tracking tasks, it adversely impacted performance in a third task. Nonetheless, alerting resulted in a decrease in both variance and mean detection time of critical changes. Participants' subjective ratings were generally positive, yet they criticized the lack of transparency of the alerting mechanism. Alerts triggered based on eye-tracking and system state show potential for improving operator task performance. Nonetheless, for the system to reach its full performance potential, it is critical that the operator understand the principles underlying the alert triggers. False positives and alert design were identified as key areas for improvement to maintain user trust and task flow.

Keywords: monitoring, eye-tracking, pilot behavior, warning systems, adaptive automation

Human error remains a significant contributor to accidents across various domains, such as driving and aviation (Kelly & Efthymiou, 2019; National Highway Traffic Safety Administration, 2018). In aviation, the loss of situation awareness (SA) has been identified as the most common factor leading to hazardous events (Kharoufah et al., 2018). Thus, an essential task for operators is to efficiently monitor the dynamic states of systems, continuously comparing their expectations with the displayed information. Failure to monitor relevant aspects of the work environment can lead to a mental picture of the situation that deviates from the actual state, resulting in poor decision-making and errors of omission (Endsley, 1995; Silva & Hansman, 2015). In civil aviation, the importance of monitoring is reflected by the role of the pilot non-flying, who is responsible for ensuring safe operation by monitoring the state of the aircraft while the pilot flying is engaged in flying. To increase safety, cockpit procedures and pilot training require different types of monitoring activities such as passive, active, periodic, and mutual monitoring (Civil Aviation Authority, 2013; Federal Aviation Administration, 2017).

Alerting systems adaptive to monitoring behavior could enhance safety by detecting poor monitoring performance and providing timely notifications to the operator regarding crucial system statuses (Feigh et al., 2012; Rouse, 1988).

This approach requires an objective measure of monitoring performance that is used to trigger alerts in the cockpit. Currently, there are not many studies that developed and investigated such adaptive systems in closed-loop studies. Before stating the goal of this study, we review the few studies known to the authors. Bosse et al. (2009) proposed a model that utilizes gaze measurement and display features to estimate operator attention. They adjusted the saliency of task-relevant objects on a map display according to the mismatch between actual and desired attention, which led to a significant improvement in task performance. Similarly, Fortmann and Mengerhausen (2014) developed an eye-tracking-based adaptive interface for an unmanned aerial vehicle (UAV) monitoring task. They adapted the saliency of key objects on the tactical map based on a measure of SA, resulting in faster detection of UAV malfunctions and intruders. Schwarz and Fuchs (2017) used a combination of physiological, behavioral, and performance measures to identify critical operator states, which triggered different adaptation strategies for their task environment. Compared to workload- and performance-based adaptation, SA-based adaptation prevented performance decrements by guiding the operator's attention to the most relevant parts of the task environment. Lounis et al. (2020) developed an eye-tracking-based support system for a real aircraft

cockpit. By comparing pilot dwell times on various cockpit indicators with a standard gaze behavior database, they detected poor monitoring and issued a vocal alert. Even though this system managed to redirect attention to critical flight instruments, it did not result in performance improvements and was subjectively rated poorly due to false alarms. The authors suggested that integrating flight parameters could enhance the system's usability.

These studies show the effective use of gaze-based measures in adaptive systems. However, it is crucial to include both gaze and system states to pinpoint the right situations for intervention. Currently, there is no generic measure of good monitoring performance applicable to various task environments, and most studies have not incorporated system state into their measure. Therefore, our research aims to contribute in two ways:

We aim to define a monitoring measure that combines eye-tracking and system state, applicable across different task contexts.

We aim to employ this new measure to trigger alerts in an aircraft cockpit that will enhance pilot monitoring performance in a representative task setting.

Measurement of Monitoring Performance

Various metrics exist that measure monitoring by capturing the temporal, spatial, and sequential dynamics of gaze behavior (Peißl et al., 2018; Ziv, 2016). These are typically linked to task-relevant areas in the workplace, known as "areas of interest" (AoI), such as fixation duration on a specific display. Through these metrics, studies found differences between novice and expert aircraft pilots and identified poor monitoring as a cause for automation surprises (Lounis et al., 2021; Sarter et al., 2007).

Measures of SA – which is often linked to monitoring – involve asking contextual information from participants at specific experiment times. The most popular methods, such as freeze-probes, online probes, or posttrial questionnaires, are not suitable for real-time measurement (Salmon et al., 2009; Zhang et al., 2020). Thus, there are studies that used eye-tracking-based implicit measures to infer SA during operation, as monitoring is a crucial activity for data gathering. For example, Moore and Gugerty (2010) showed in an early study that the time spent fixating on relevant AoI predicted SA scores in an air traffic control task, provided that fixations were optimally distributed. Excessive focus on one AoI degraded the overall situational picture, a finding replicated in other domains (Hasanzadeh et al., 2018; van de Merwe et al., 2012). Winter et al. (2019) later revealed a correlation between viewing behavior, task state, dynamics, and performance. They found performance strongly linked to correct sampling timing, defined by the current state of a

moving dial, and concluded that integrating system state into eye-tracking-based online SA measurements could be beneficial.

On the basis of these findings, we introduce a measure called "awareness deviation". This measure describes how well the operator is aware of a system state by comparing fixated values with the current system state. We then show how this measure can help in setting off alerts in an aircraft cockpit.

Awareness Deviation as a Measure of Monitoring Performance

Our goal is to design a system that estimates a pilot's awareness of the aircraft status, steering their attention to crucial details when their knowledge of the system state deviates from the ground truth. This system should tell the difference between various state details (e.g., altitude or position), and the awareness attached to each one. To achieve this, we introduce the concept of *awareness deviation* (AD), which can also be interpreted as a measure of the perceptual level of SA (Level 1) by quantifying the difference between the current state value and what the system assumes is the pilot's last perception of each state (Endsley & Jones, 2012). The metric AD was first introduced and tested by the authors in two previous studies (Schwerd & Schulte, 2020, 2021), on the basis of which this follow-up study was developed. With this AD measure, we can alert the pilot and direct their attention to any unnoticed changes in the system state. Figure 1 gives a full picture of the adaptive mechanism, which includes two inputs (system state and eye-tracking measurement) and produces a specific alert for the pilot. We break down the five modules of the process in more detail in the sections that follow.

Fixation Filter

A conventional eye-tracking system provides gaze positions in the cockpit. All gaze samples are filtered for fixations in the *fixation filter* (see Figure 1, Part 1). For the sake of simplicity, we assume that the pilot perceives every piece of information when he fixates the display of the information within a margin of 2° on a cockpit indicator. From a fixation, we infer the AoI and extract the value displayed at the time of measurement. Note that with our assumption based on the *eye-mind hypothesis* (Just & Carpenter, 1980), we accept false-positive measurements, where a fixation and perception of information do not align (e.g., a "blank stare" into the void). A more detailed analysis of this problem is given by Schwerd and Schulte (2022).

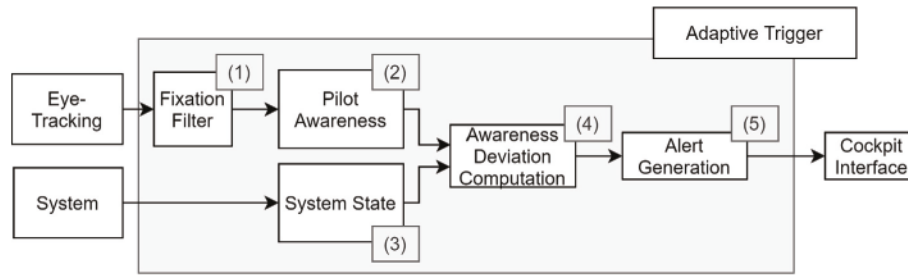


Figure 1. Data flow of awareness deviation measurement for specific information.

Pilot Awareness and System State

Given the eye-tracking measurements, the *pilot awareness* module holds a set of elements describing the state of the system the pilot is aware of (see Figure 1, Part 2). Each element of this set is a system state associated with the displayed value at the most recent fixation (e.g., “altitude: 10,300 ft”). Similarly, a *system state* is generated containing all system states associated with the actual value at the current time (see Figure 1, Part 3). The system state is continuously updated with current system state values.

Awareness Deviation Computation

In the *awareness deviation computation*, we compute the difference between system state and pilot awareness (see Figure 1, Part 4). This difference is computed on each state on the display separately, which creates a deviation value for every state. Since system states are usually encoded by different types, such as modes (e.g., automation modes), texts (e.g., warnings), times (e.g., waypoint timing), or positions (e.g., map information), we defined a function $\text{dist}(c_p, c_s)$ for different types in order to compute the difference between pilot value c_p and system values c_s , respectively (see Table A1 in the Appendix).

We then adjust $\text{dist}(c_p, c_s)$ by a normalizing constant c_n to quantify when a deviation is significant enough to be considered a substantial distance. A substantial distance is later used as a threshold for triggering alerts. For a given system state such as aircraft altitude, the constant c_n is not universal – it varies based on the task, as different tasks have different accuracy requirements. For example, awareness of flight altitude needs to be more precise during landing and take-off than in transit situations in high flight levels.

As a final processing step to compute the awareness deviation $\text{AD}(c_p, c_s)$, we apply an exponential function to normalize the value between 0 and 1:

$$\text{AD}(c_p, c_s) = 1 - e^{-\frac{|\text{dist}(c_p, c_s)|}{c_n}}, \quad (1)$$

where c_p is the value in the pilot’s awareness, c_s is the system value, and c_n is the normalizing value. The function has the following attributes: (1) for $\text{dist}(c_p, c_s) = 0$, the deviation

AD is 0. (2) With growing $\text{dist}(c_p, c_s)$, AD grows rapidly until the difference reaches the normalizing constant c_n after which AD converges to 1.0 for $\text{dist}(c_p, c_s) \rightarrow \infty$. We chose this system dynamic since it has advantages in statistical processing because deviations greater than what is considered large in the task context (defined by c_n) are not overweighted. Apart from the processing reason, the rationale behind the exponential formula is that at a certain distance between the pilot and system value, it does not matter if the difference grows even further.

Figure 2 shows an exemplary course of the AD for aircraft speed. In this example, we chose a normalizing constant of $c_n = 30$ kt.

Initially, the speed is at 300 kt (indicated by the system value) and the pilot is aware of this value (indicated by the pilot value), and therefore the distance and AD are zero. When the speed changes, the distance between the pilot and system value increases. Accordingly, the AD increases within its normalized limits. Following our assumptions, the pilot’s fixation within a margin of 2° on the speed indicator updates their knowledge about a speed of 400 kt. When the pilot updates their knowledge, distance and AD are set to 0. After 3 s, the speed changes again without the pilot’s awareness. At 22.5 s, the second fixation updates the knowledge about a current speed of 300 kt.

Alert

The ADs for all system states are passed to the alert generation (see Figure 1, Part 5). An alert is generated when the deviation value exceeds a limit for a predefined duration. We selected this limit to be .6, which indicates a $\text{dist}(c_p, c_s)$ of approximately the normalizing constant. We wait to trigger an alert for a predefined duration to prevent premature alerts the moment after the state has changed.

Table A2 in the Appendix denotes the trigger algorithm for a single state. For every alert, a normalizing constant c_n and a delay time (time-until-notification) must be selected. Note that there is no check on whether the pilot has actually made an error before an alert is triggered. We chose to evaluate AD-adaptive alerts independent of the desired flight parameters because we were interested

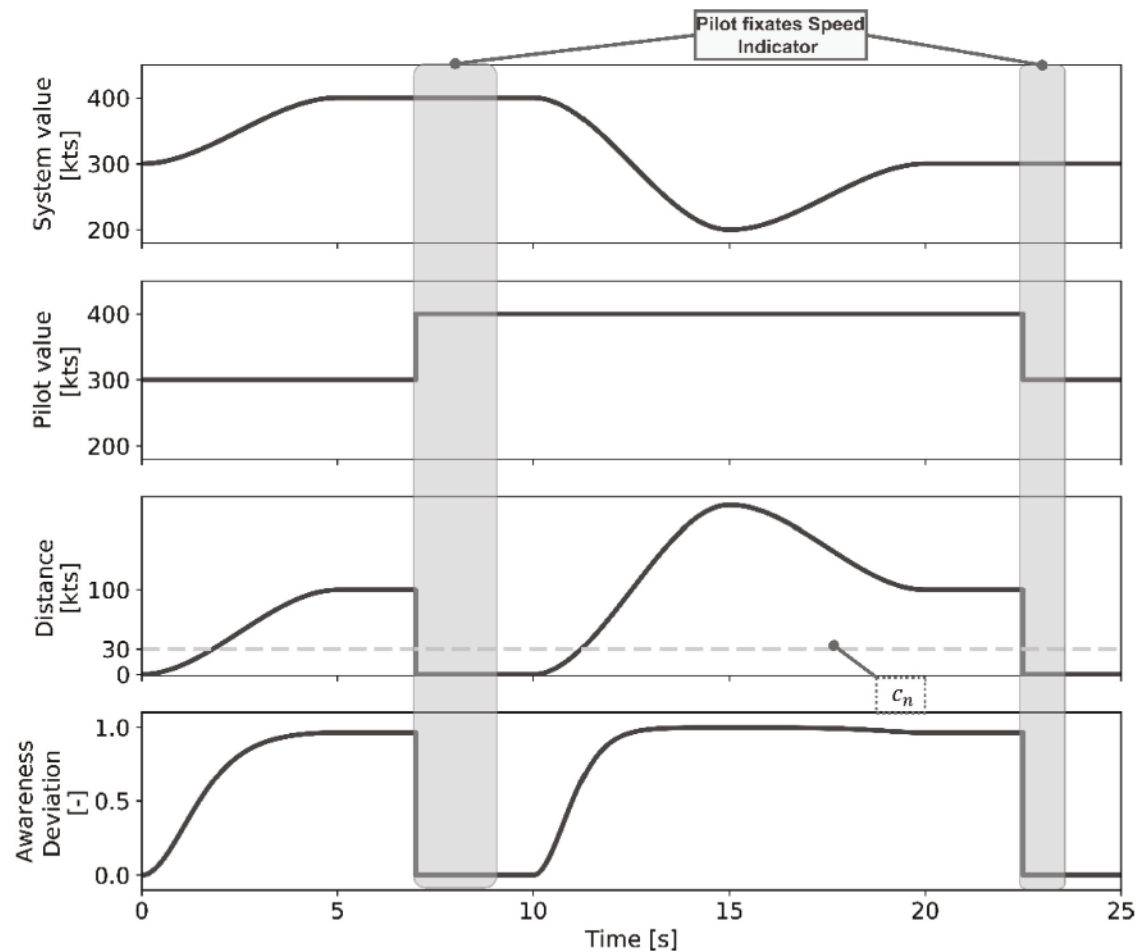


Figure 2. Example for course of values for system, pilot, distance function, and awareness deviation function for aircraft airspeed. There are two fixations on the relevant area of interest (speed indicator) at $t = 7$ s and $t = 22.5$.

in whether it is possible to trigger useful alerts using only the proposed monitoring performance. The rationale for this was that in some situations it is not trivial to extract the current performance of the pilot, for example, when the system does not know the desired flight parameters, and early notification of poor monitoring performance could prevent future performance degradation. But we acknowledge that it may be preferable to include pilot performance in the alert-triggering mechanism if performance measures are readily available to the system.

Experiment

The goal of the experiment was to evaluate the AD metric as a suitable trigger for an alerting system. It was designed according to the guidelines of the ethics committee of the University of the Bundeswehr Munich. Our hypothesis was that our metric is correlated with pilot errors, and we assumed that triggering alerts based on AD would improve monitoring performance compared to no assistance.

Setup

Trials were conducted in a cockpit simulator resembling a generic fast-jet cockpit with three touchscreens and a head-up display (HUD; see Figure 3). For eye-tracking, we used the commercially available four-camera system by SmartEye (Smart Eye Pro 0.3 MP, 60 Hz, best accuracy $< 0.5^\circ$) and the fixation classifier implemented in the software SmartEye Pro 8.2 with a gaze angular velocity threshold set to $\dot{\alpha} \leq 2^\circ$.

Participants and Procedure

We conducted the experiment with 10 male participants ($\mu_{\text{age}} = 36.9$ years ± 12.5). Since alert was triggered based on the gaze behavior of the pilot, the system should be adaptive to different levels of experience. Therefore, we recruited pilots with differing flight hours (between 500 hr and 29,000 hr on civil aircraft) reflecting a broad range of expertise. The participants hold either a commercial pilot license ($n = 5$) or a private pilot license ($n = 5$, at



Figure 3. Cockpit simulation environment with integrated eye-tracking system.

least German Glider pilot licence (SPL) including Touring Motor Glider licence (TMG)), but had no prior experience in using our simulator. The pilots participated voluntarily and were not compensated in any way.

Each participant received a presentation about the experiment and provided written consent about the trial. Then, each pilot received training in the experimental task for approximately 1 hr. During this training, they encountered all aspects of the experimental task with equal difficulty and frequency. Then, the eye-tracking was calibrated with mean accuracy of 1.04° ($SD_{acc} = 0.72^\circ$). After that, the participants conducted two 30-min trials with equal difficulty,

first without and then with alerting. Participants were not briefed about the alerting mechanism, but were only told that there is an alerting system active in the second trial. After the second trial, all participants answered a questionnaire to evaluate the assistance system.

Tasks

In both trials, the experimental task included monitoring of continuous parameters and discrete states of the aircraft-cockpit interface with all task-relevant AoIs shown in Figure 4.

The participants had three continuous tracking tasks without using an autopilot: First, they had to track a route displayed in their tactical map (see Figure 5), where only the route leg to the next waypoint was visible (as displayed in the top right of Figure 4, AoI 3). After reaching a waypoint, a new route leg appeared with an unknown heading.

Second, the participant had to track a specified altitude and speed, both displayed in the HUD (Figure 4, AoI 1 and 2). The target altitude and speed were changed by a text message with a frequency of 0.1 times per minute, which was indicated by a green light in the left screen (Figure 4, AoI 5). At unknown times, the aircraft altitude or speed was disturbed by a simulated gust that changed the aircraft position or speed to a value beyond the target range. We triggered disturbances of either altitude or speed with an average frequency of 3.5 times per minute, but there was no disturbance of both variables at the same time. We simulated bad visibility to minimize the influence of spatial visual perception in the outside view (see HUD in

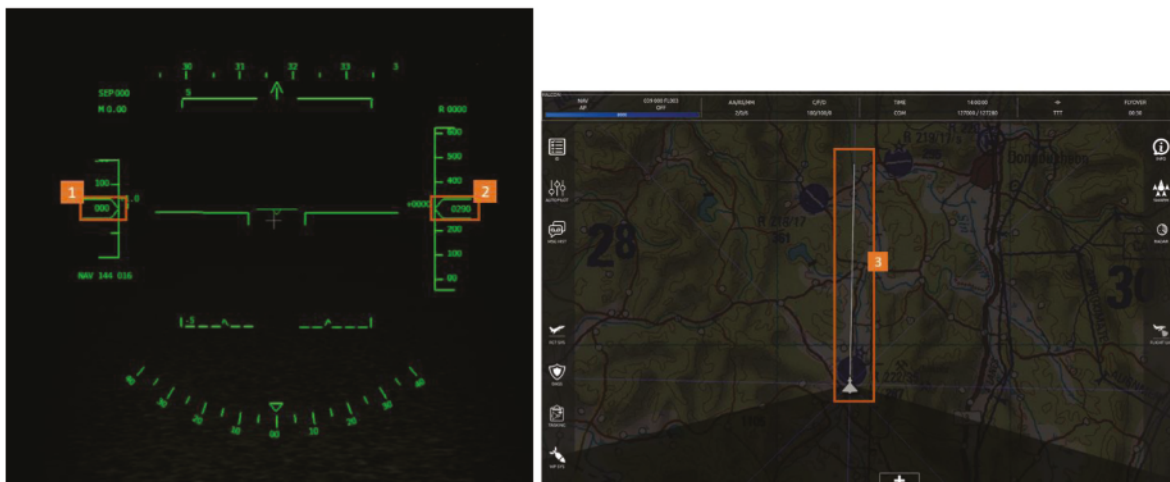


Figure 4. Simulator display setup with area of interest (top left: head-up display; top right: center display; bottom: side displays). Areas of interest (AoI) marked in orange displayed the following information: (1) altitude indicator, (2) speed indicator, (3) position of aircraft and route to next waypoint, (4) warning system (red on new warning), (5) message system (green on new message), (6) fuel system with button controls for procedures, and (7) engine system with controls for procedure.

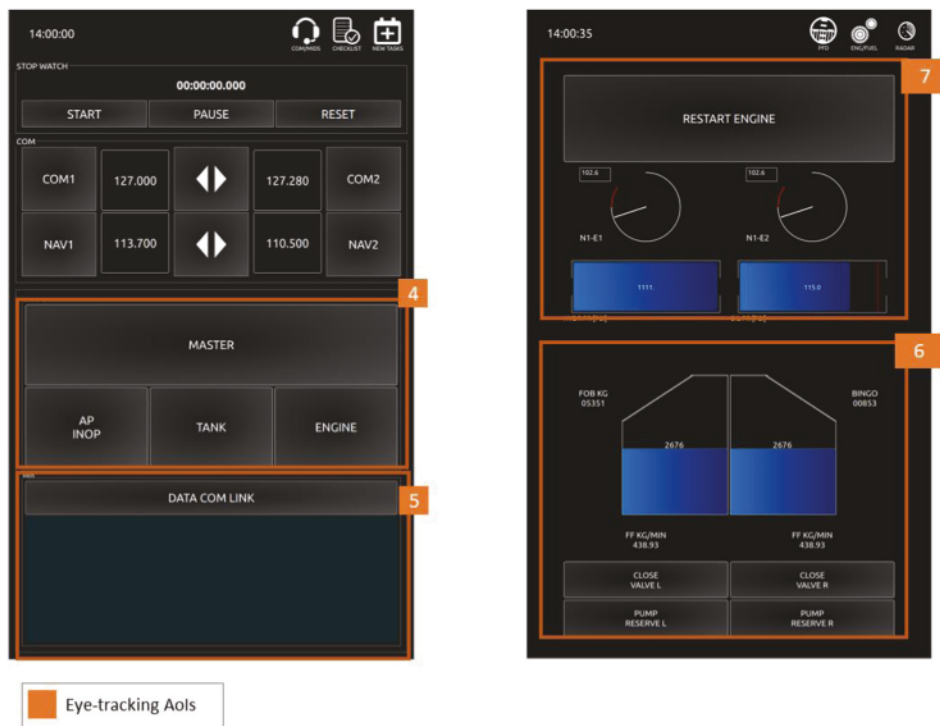


Figure 4. (Continued)

Figure 4). Therefore, change of speed and altitude could only be recognized by gazing at the displayed values.

Third, two aircraft warnings were triggered, which were visible as red indications on the left screen (Figure 4, Area 4). Participants had to react with trained procedures to these warnings. The procedures comprised the activation of system modes (Figure 4, Areas 6 and 7) or a change of thrust.

Configuration of Adaptive Mechanism

We applied the AD measurement to each of the five experimental tasks: tracking route, speed, and altitude, acknowledging a new message, and following a warning procedure. Distance functions for speed and altitude were numerical distances. For deviation in route tracking, we computed the geometrical distance between aircraft and the nearest route point as the system value. When a participant fixated the AoI (see AoI 3 in Figure 4), we presumed they fully understood the exact distance and calculate the numerical distance. For mode-like states of a new message or warning, we used the distance function as outlined in Table A1 in the Appendix.

We used a synthesized voice for the alerts, providing the pilot with key warnings such as “ALTITUDE” or “SPEED.” For route, altitude, and speed, we set the normalizing constants c_n to the tracking limits describing the required

Table 1. Constants for adaptive functions in altitude, speed, route, message, and warning

Information	Normalizing constant	Delay [s]
Altitude	200 [ft]	2
Speed	30 [kt]	2
Route	3,000 [m]	3
Warning	1 [-]	5
Message	1 [-]	5

tracking accuracy briefed to the pilots (e.g., altitude had to be tracked within 200 ft). For warning and message, we used $c_n = 1$ because these indications are binary – it is either “there is a new message” or “there is no new message.” Before the experiment, we set delay values that would provide the right balance between too-early and too-late alerts for each task. Table 1 presents the constants c_n and delay times for each information relevant to the experimental task.

Figure 6 shows an example of an alert to illustrate the behavior of the adaptive system. In this scenario taken from the experiment, a simulated disturbance changes the speed of the aircraft (shown in gray). The participant does not notice this change, which leads to a rising AD in speed. When the AD exceeds the threshold for more than 3 s, a vocal alert is activated. In response, the pilot immediately checks the speed indicator (see bottom graph in Figure 6: “Fixation on Speed Indicator”), which brings the AD back down to 0.

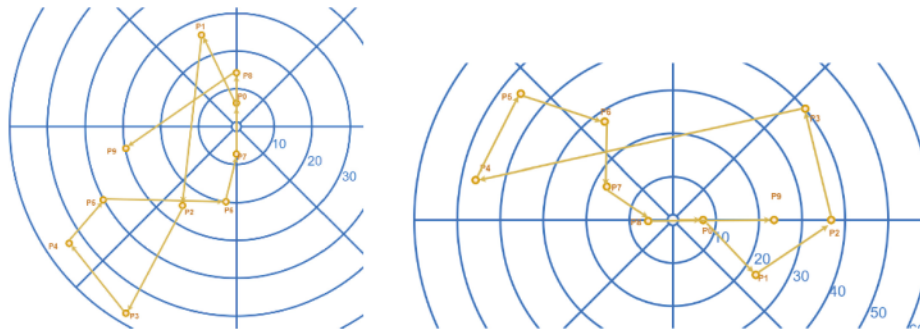


Figure 5. Schematic illustration of both scenarios with waypoints. The participants started at the center of the circle.

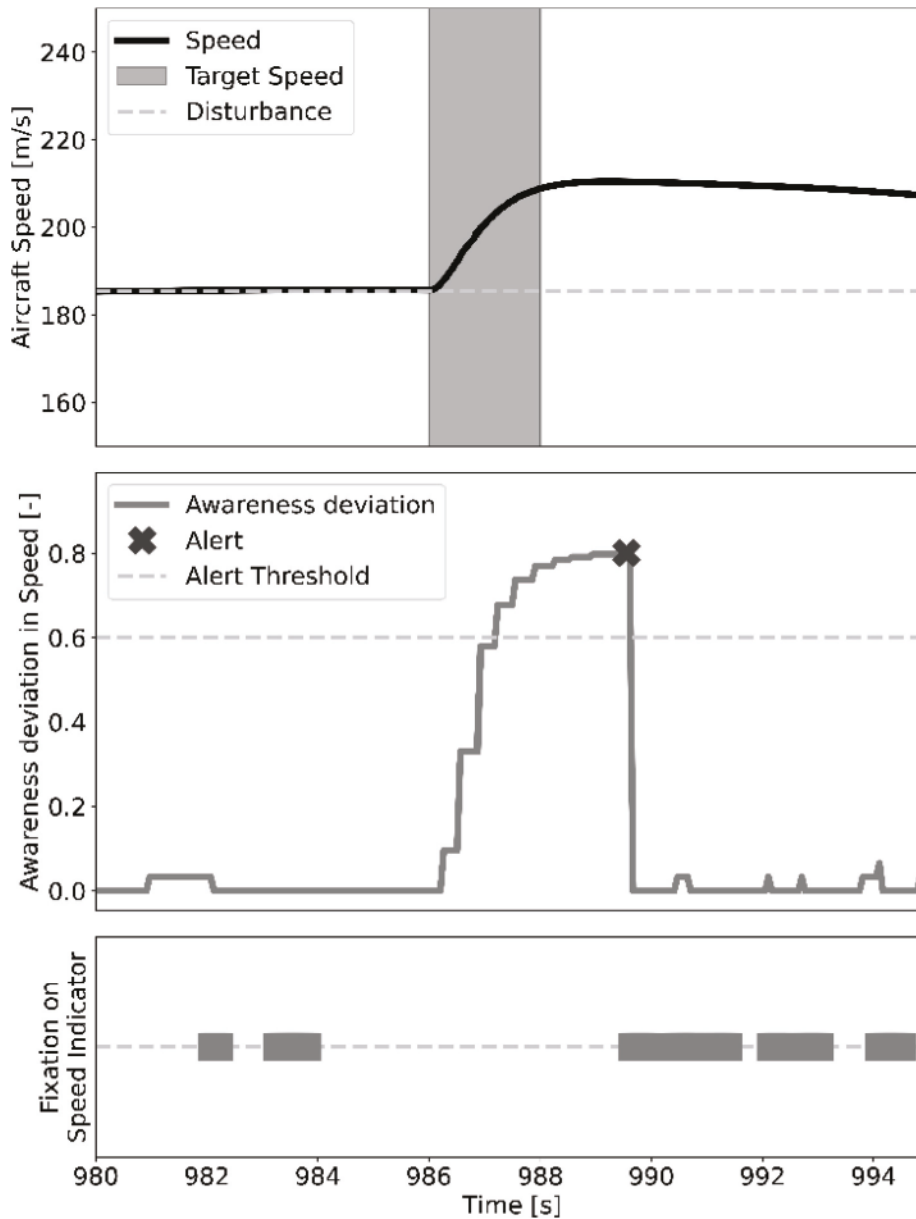


Figure 6. Example of awareness adaptive alert (here: speed). These data are taken from the experimental results. The course awareness deviation is not smooth, because the sample frequency of system value updates is 2.5 Hz.

Data Analysis

Task performance for the tracking tasks was quantified as follows: We calculated the mean error rate as the root mean square (RMS) of the difference between the target and actual values. If the actual value fell below the briefed tracking limits, we set the error to 0. The tracking limits were 200 ft for altitude, 30 kt for speed, and 2 nautical miles (NM) for the route.

During the trial, we collected data on AD, task performance, alerts, and gaze. We analyzed these data with Python Pandas (Reback et al., 2021) and conducted statistical tests with Python SciPy (Virtanen et al., 2020).

We applied the Shapiro-Wilk test to check whether the data were normally distributed. If they were, we used Pearson's correlation, indicated by the variable r , and a dependent t test to verify the significance. If not, we used Spearman's rank correlation, indicated by the variable r_s , and Wilcoxon signed-rank tests. We used Levene's test to ensure equal variances. Correlation strength was categorized as follows: weak ($r \leq .1$), medium ($.1 < r \leq .7$), and strong ($r > .7$). We set the significance level to $p = .05$.

Results

In the following we present different aspects of the results: First, we analyze the relationship between AD and tracking performance in the baseline condition (without alerts) and the relationship between number of alerts and tracking performance in the experimental condition (with alerts). Second, we compare how different gaze and performance measures changed with the presence of alerts. Third, we present the subjective feedback of the adaptive system given by the participants.

Relationship Between AD and Performance

To assess our method of measuring the monitoring performance, we compared AD with pilot performance in three tracking tasks without alerts in a control setting. Figure 6 shows the mean error and mean AD throughout all tracking tasks for the control condition. Figure 7 displays the mean error and mean AD over the complete trial in all tracking tasks for the control condition.

We found a strong correlation of $r = .88$ (significant, $p < .001$) between the mean altitude tracking error and AD in aircraft altitude. Speed data also demonstrated a strong positive correlation of $r = .75$ (significant, $p < .01$) with AD in speed. These findings align with a previous study with different participants that experimentally validated

operator measurement in a similar task (Schwerd & Schulte, 2020).

The far-right plot in Figure 7 presents the mean AD and error in route tracking. While 60% of participants did not make any errors in this task, there was a moderate positive correlation coefficient between AD and route error of $r_s = .65$, which is significant.

These data indicate a strong link between AD and tracking performance in speed and altitude, with a somewhat weaker relationship in route tracking.

Relationship Between Error and Number of Alerts

In the adaptive setting, we compared the number of activated alerts with the error rate. The results can be seen in Figure 8. There was a strong correlation between the number of alerts and error rate in both altitude ($r = .8$, $p < .01$) and speed ($r = .77$, $p < .01$). These correlations suggest that the AD measure is meaningful in terms of speed and altitude tracking performance because it triggers alerts in the event of an error without directly measuring errors. However, these data also show that our configuration of adaptive alerts may not prevent individual pilot errors.

As in the control condition, there is a floor effect in the route tracking task. Some participants did not make any errors in route tracking. However, there was a mediate correlation of $r_s = .67$ ($p < .04$) between the number of route alerts and the mean error in route tracking.

Interestingly, some participants triggered alerts without making any errors in route tracking. These false alarms occurred because the adaptive system relied solely on AD to activate alerts, without checking whether there was a genuine error in relation to the experimental task. This result shows that performance can be good even if the pilot is not monitoring a system value frequently due to estimation of the system value via a mental model.

Comparison of Tracking Performance Between Conditions

Figure 9 compares the average tracking error in both the control and experimental conditions. Errors in tracking speed and altitude were reduced in the adaptive condition. For speed error, which followed a normal distribution, the paired t test marked the decrease as significant, $t(9) = 5.53$, $p = .0004$. The change in altitude error, however, was not significant ($ps < .084$). Contrary to speed and altitude, the error in route tracking increased with adaptive alerting (not significant, $ps < .11$). The average performance shown in Figure 9 reflects two aspects of task performance: First, the participants' ability to perform the tracking tasks without any disruption to the aircraft altitude and speed.

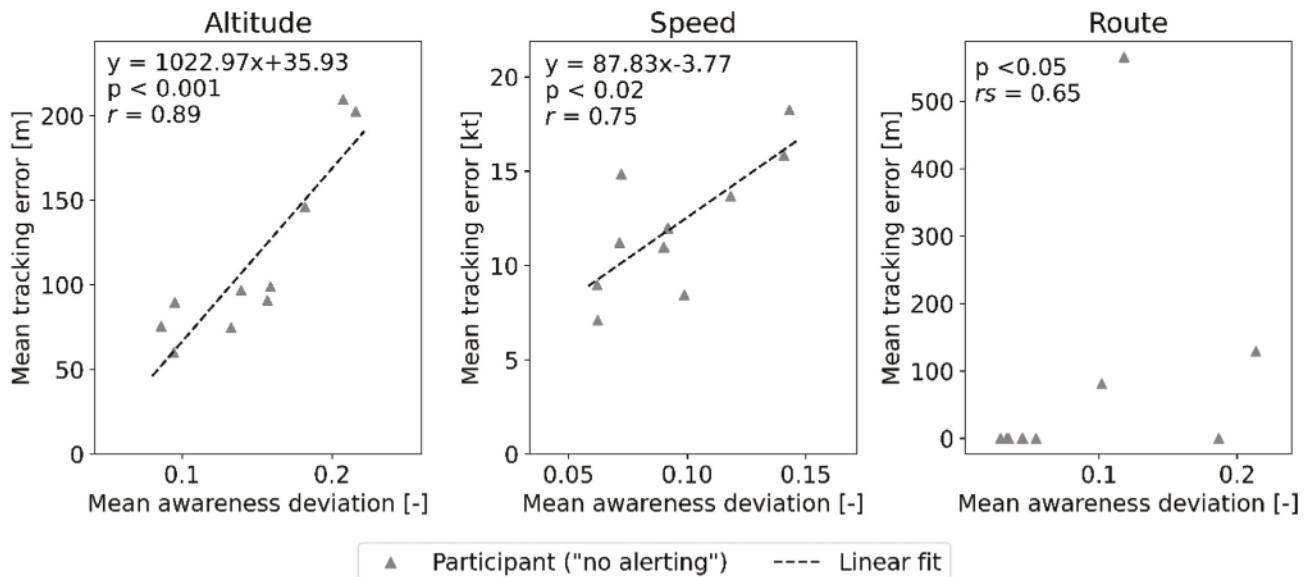


Figure 7. Correlation between awareness deviation and error in tracking for the condition "no alerting" ($n = 10$).

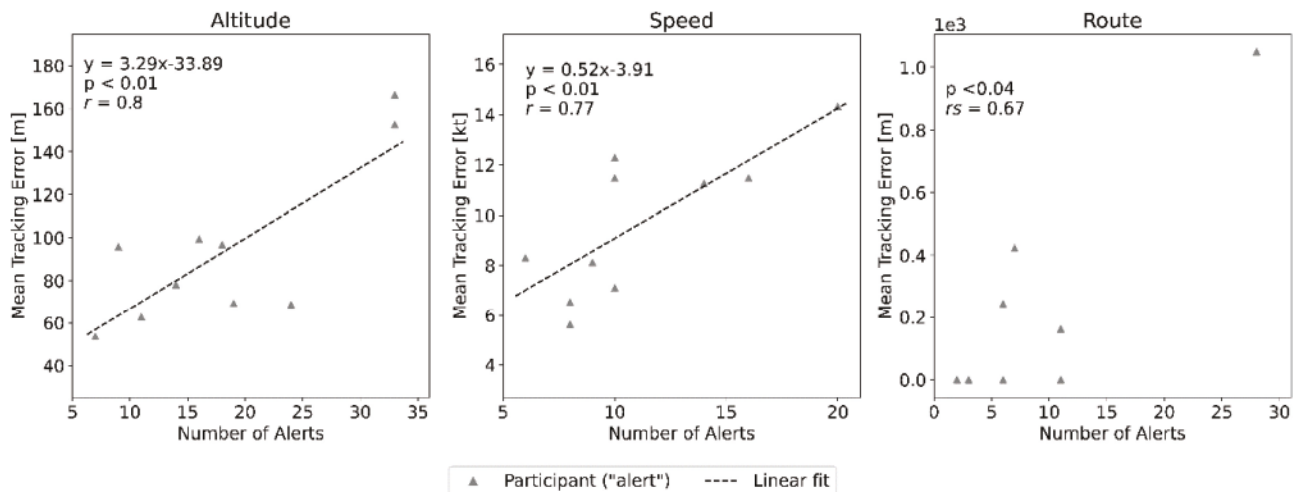


Figure 8. Correlation between number of alerts and mean error ($n = 10$).

Second, their ability to promptly recognize a disturbance and revert to the target value.

A deeper understanding of the difference between the two conditions can be gained by examining the time it took participants to notice an error in the aircraft state after it had been induced.

Comparison of Detection Performance Between Conditions

We defined detection time as the period between a change (disturbance, new waypoint, or warning) and the moment the participant first glanced at the relevant AoI (see

Figure 4). Table 2 presents mean and standard deviation values, and Figure 10 provides a letter-value box plot of the distribution. For altitude and speed, the average time to detect changes saw a slight decrease in the adaptive condition. On the other hand, there was a nonsignificant increase in the mean detection time for the route.

For altitude and speed, the adaptive alerting also reduced the standard deviation. The reduction between conditions was significant (altitude: $p < .01$, speed: $p < .03$). This indicates that the adaptive system effectively detected situations where participants failed to notice a value change in a system state. The number of instances with extended detection times was reduced by alerts triggered once the

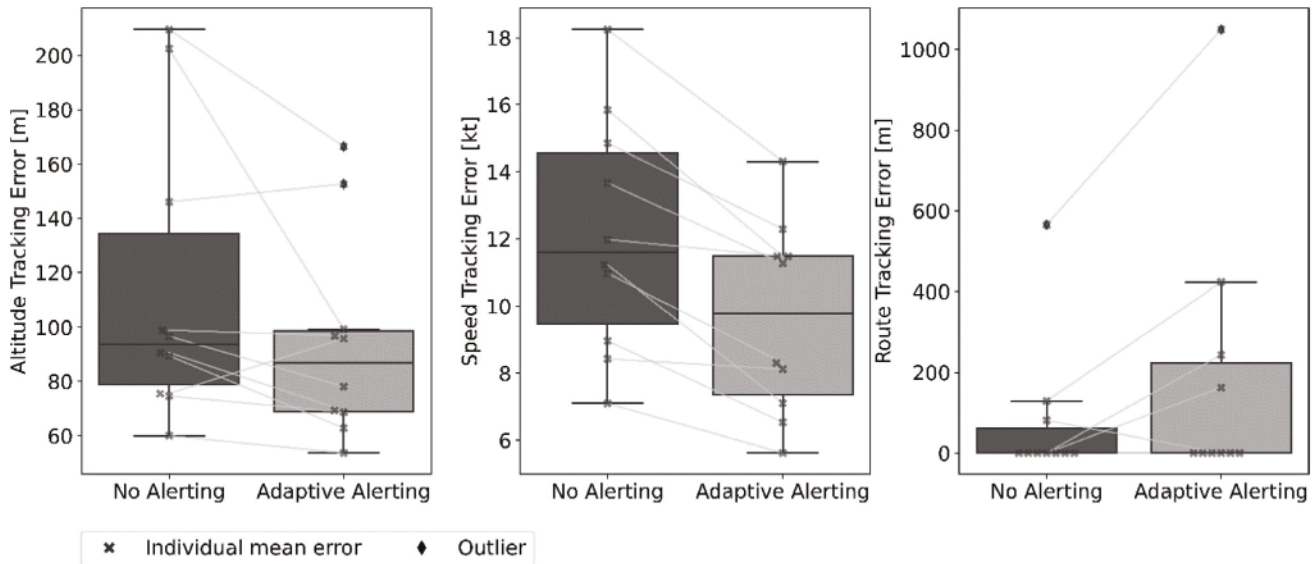


Figure 9. Average error rate in the tracking tasks ($n = 10$). "X" markers indicate means of individual participant connected over two conditions by a gray line.

Table 2. Detection times for tracking tasks

Condition	Altitude detection [s]		Speed detection [s]		Route detection [s]	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
No alerting	0.93	1.04	1.25	1.56	1.01	1.84
Adaptive alerting	0.83	.75	1.09	1.19	1.70	2.74

Note. Altitude, $n = 939$; speed, $n = 823$; route, $n = 129$. *M* = mean; *SD* = standard deviation.

Table 3. Detection times for message and warning

Condition	Warning detection [s]		Message detection [s]	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
No alerting	3.96	11.86	7.93	22.49
Adaptive alerting	1.23	0.94	3.18	3.45

Note. Warning, $n = 20$; message, $n = 54$. *M* = mean; *SD* = standard deviation.

AD measurement exceeded the threshold for more than the specified time delay.

By contrast, route tracking had a different effect. Although the median detection time dropped, the standard deviation increased, but the change was not significant ($p < .09$). It is worth noting that the median times were below the alert trigger delay, and thus the system could not have improved them.

Table 3 displays the average detection times for new messages and warnings. There was a reduction in both the mean and standard deviation in the experimental condition. This effect is largely due to the outliers visible in Figure 11, which also shows that the primary impact of the adaptive system was to reduce variance. The number

of samples was limited because warnings only occurred twice during one trial, and new messages arrived five to six times.

Comparison of Fixation Count and Duration on AoI Between Conditions

Fixation count and durations on AoI for both conditions are plotted in Figure 12 and reflect the system state dynamic in the experiment. Altitude and speed indications are fixated the most and longest compared to AoIs that are relevant to tasks with lower demand for continuous monitoring. The data show small changes between experimental conditions,

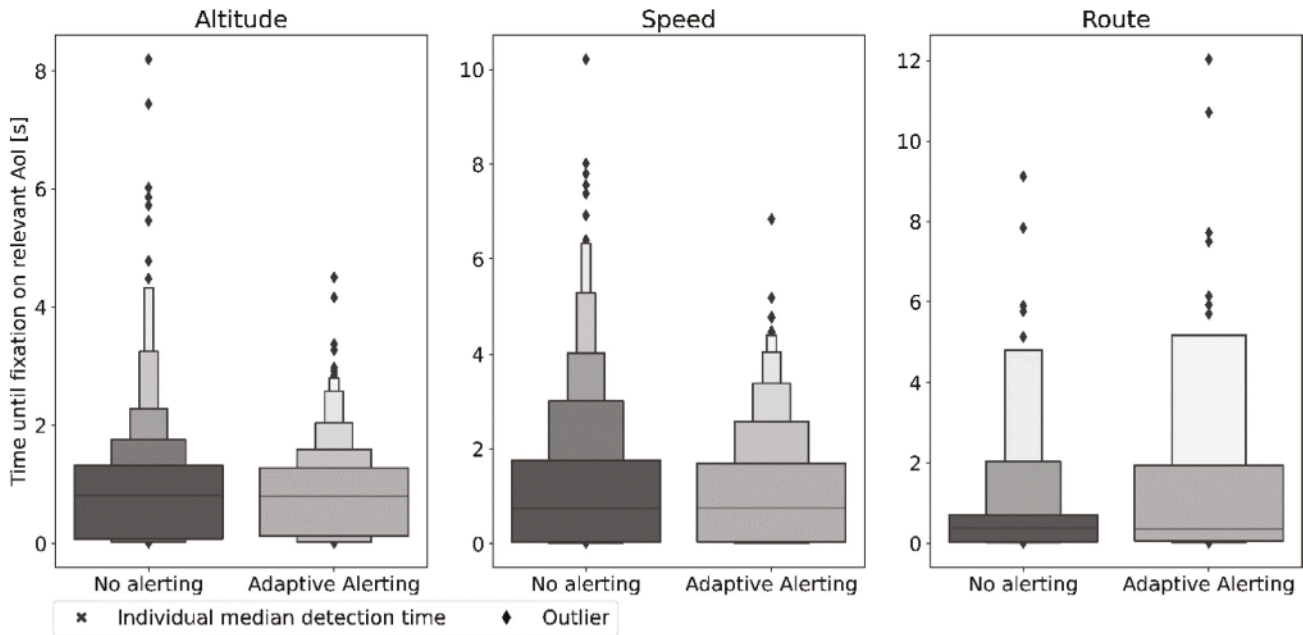


Figure 10. Time until detection of disturbance in a target parameter. Altitude, $n = 939$; speed, $n = 823$; route, $n = 129$.

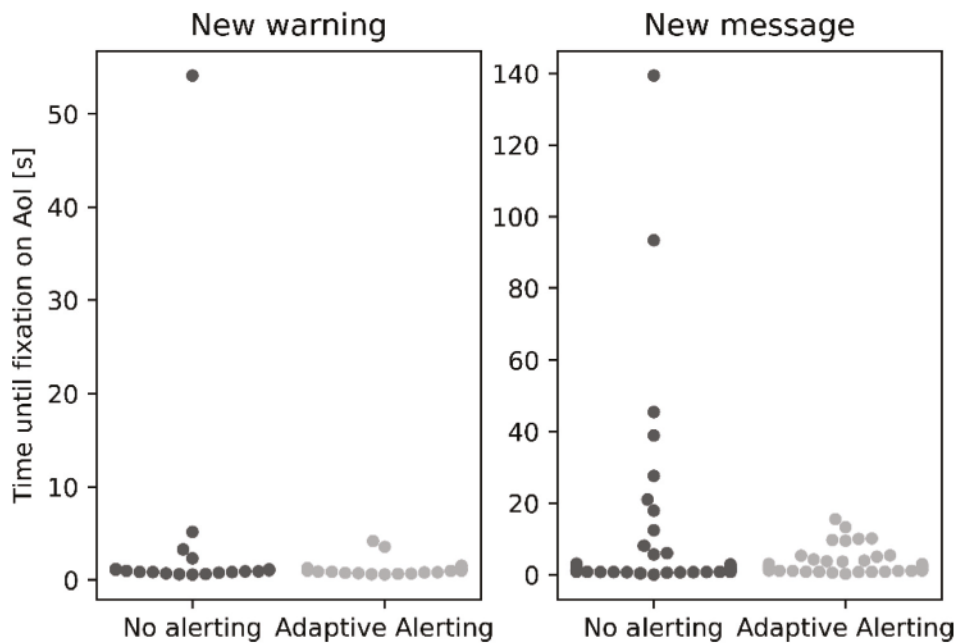


Figure 11. Time until detection of new warning (left) and new message (right).

most prominent in fixation count for altitude, speed, and route and in fixation duration for fuel, route, speed, and warning. Both metrics decreased for the route AoI, which indicates that the alerts moved attention away from this task, which aligns with the performance measures.

Subjective Rating

After the second trial, participants were asked to evaluate the system through a nonstandardized questionnaire (see

Figure 13). Overall, most participants found the alerts useful and felt they were activated at an appropriate frequency. They also did not feel disturbed or distracted by the alerts during other tasks.

However, the average participant rating revealed low levels of trust and transparency. In a debriefing following the trial, some participants expressed confusion about why an alert was triggered, particularly when their flight parameters were within the desired limits. This led to a loss of trust after a few false alerts. Another criticism was that

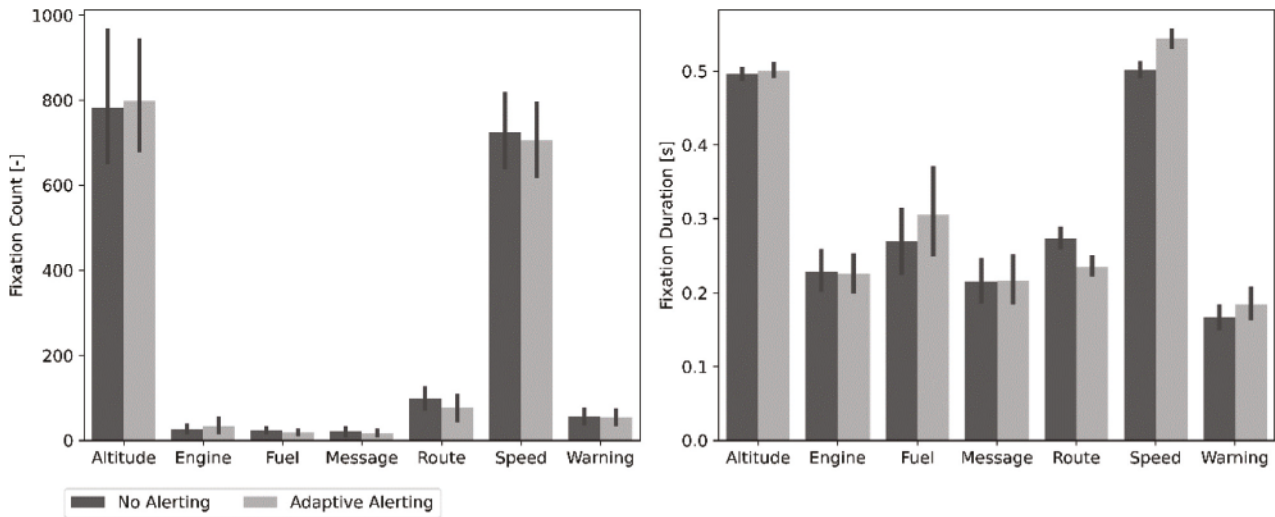


Figure 12. Bar plot of mean fixation count and fixation duration for all areas of interest and both experimental conditions.

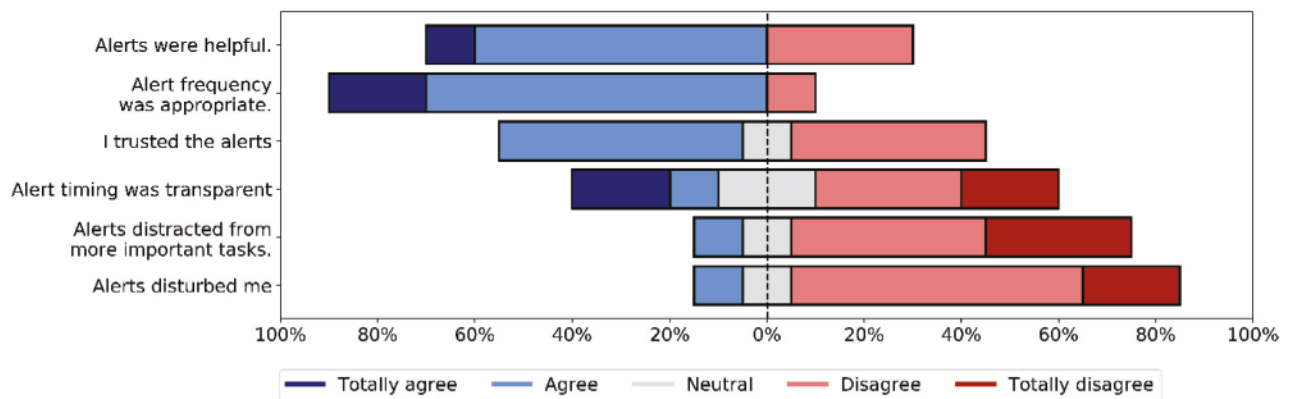


Figure 13. Subjective rating of adaptive alerting ($n = 10$).

auditory alerts should be reserved for critical situations, not minor deviations in flight parameters.

Discussion

In summary, the AD measure was effective in revealing notable features about the participants' performance, with results from the control condition showing a correlation with pilot errors. The alerts not only improved average performance but also reduced variance, particularly in the "unpredictable" tasks such as altitude tracking, which was disturbed by virtual gusts. Furthermore, using AD as an adaptive trigger was successful in generating sensible alerts in the experimental condition without requiring a direct error check. While checking for errors would be trivial for the experiment, this is not the case in every real-world application.

While the performance increase for speed and altitude tracking was consistent, performance in route tracking declined. Our initial explanation is that the alerts for speed and altitude may have distracted pilots from tracking the route. We noticed a higher number of task alerts for participants with lower performance (see Figure 8). If this low performance was due to high workload, additional alerts could potentially further impair the participant's performance. Another possible explanation could be that pilots rapidly lost trust in the route alert, as it was sometimes triggered without an actual error (as shown in Figure 8). This led participants to disregard the route alert.

From these results, we draw the first two conclusions from our study: First, unpredictable shifts in system states are more suited for adaptive alerting than predictable changes. Hence, a predictable shift in values should involve a longer delay before triggering an alert, as the pilot is likely

aware of this change. Second, false positives pose a serious issue for user trust. This must be addressed carefully when designing such systems. A similar issue has also been observed by Lounis et al. (2020).

Even though the performance of almost all participants improved, their subjective feedback was mixed. We suggest two reasons for this: Firstly, the assumptions about the direct relationship between fixation and perception do not hold in every situation, therefore AD was overestimated while the pilots were already aware of the state. This led to the false alerts, which contributed to the negative evaluation, particularly in route tracking where alerts were frequently triggered without a need. Secondly, pilots were not briefed about the logic behind the adaptive system before the experiment, leading to confusion when alerts were issued without clear reason. The combination of a high number of false alerts and a lack of transparency may be responsible for this mixed feedback. This aligns with findings from Dorneich et al. (2016), where aviation professionals ranked low transparency and predictability of adaptive systems in the cockpit as the main risk factors to flight safety and user acceptance. To boost user acceptance, we recommend training participants on the adaptive system. Moreover, the design of such systems could be enhanced by incorporating models of transparency, as suggested by Chen et al. (2018).

When it comes to alert design, we found our vocal alerts to be quite intrusive for the multitask experiment. The audio notifications disrupted the participants' workflow. This kind of design might be suitable for drawing attention to critical safety information, but it should be toned down for less crucial data.

Generally, managing interruptions is a key aspect in designing alert systems (McFarlane & Latorella, 2002). As an alternative, the system could have used visual alerts to emphasize changes in parameters, similar to the approach by Fortmann and Mengerlinghausen (2014). Another option could have been to delay alerts until the participant had finished their current task, as suggested by Katidioti et al. (2016).

Limitations

There are limitations to the validity of the experiment. First, the training effect was not eliminated since the sequence of trials was "no alerting" followed by "adaptive alerting." This could account for the improvements in performance. However, it is evident that the alerts influenced pilot performance since we did not see performance increase in all tasks (e.g., worse performance in route tracking). Further, the number of outliers where pilots fail to notice some change for a particularly long time is not a matter of training, but rather a situational error, which has the biggest potential for this alerting approach. Additionally, we had an extensive training of the experimental task before the

trial, which could help to reduce the confounding potential of the unbalanced trial sequence.

Second, there was no support system in the control condition. This is somewhat an unfair comparison, since any kind of support has a good chance of improving performance compared against no support. This study lacks the evidence that the adaptive trigger is superior to the situational trigger checking for errors.

Third, the experiment was conducted with only 10 participants. A follow-up study should include more participants to gain statistical power. This would be particularly important to the validity of the results related to the reduction of rare outliers.

In general, the approach to measure AD is very simple and susceptible to both false-positive and false-negative classifications of perception. The former could be explained by inattentive blindness (Kennedy et al., 2017) or working memory limitations (Cak et al., 2020) while the latter could be caused by nonfocal attention and perception with peripheral vision (Ramón Alamán et al., 2020). However, we think that this study is a step toward the implementation of more advanced adaptive systems for aircraft cockpits.

Conclusion

Even with the known limitations of the AD measurement approach, the study demonstrated that adaptive alerting was able to trigger useful alerts. While the subjective ratings were mixed, we have explored potential reasons for these results and proposed various ways to improve the system design.

A remaining challenge is the robust estimation of pilot attention on a professional flight deck. In the current study, we were able to assign different system values to individual experimental tasks and position the corresponding (AoIs a significant distance apart. This made identifying the pilot's focus relatively straightforward. However, in a real cockpit, indicators are positioned closely (e.g., on a primary flight display, integrate different information (e.g., symbols on a tactical map), or are overlaid with other information (e.g., HUD). Hence, robust identification of attended information is not trivial and remains an open research question.

In general, the integration of adaptive systems into safety-critical environments, like aircraft, requires more conceptual research and rigorous testing. The notion of system adaptivity needs to be clear and transparent and should provide tangible benefits to the operator to be fully effective, as outlined by Dorneich et al. (2016). Most laboratory experiments, including the one detailed here, often operate under the assumption that the operator is not fully aware of the logic behind the adaptation, or at the very least, not thoroughly trained in the use of the adaptive system. This assumption is worth questioning, as the integration of adaptive systems

into safety-critical environments would undoubtedly require comprehensive training, which would certainly include information about the system's adaptive logic.

As a result, it is unclear how operators might alter their behavior when they are supported by a behavior-adaptive system and what impact this would have on the efficiency of the overall human-machine system. Would operators rely too heavily on the adaptive system, possibly leading to overtrust and decreased vigilance? Or would they use the system as a tool to enhance their capabilities, leading to better overall performance? Further research is needed to answer these questions and guide the design of effective, reliable, and trustworthy adaptive systems.

References

- Bosse, T., van Lambalgen, R., van Maanen, P.-P., & Treur, J. (2009). Attention manipulation for naval tactical picture compilation. In *2009 IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology* (pp. 450–457). IEEE. <https://doi.org/10.1109/WI-IAT.2009.194>
- Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology & Work*, *22*(1), 85–94. <https://doi.org/10.1007/s10111-019-00551-w>
- Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- Civil Aviation Authority. (Ed.). (2013). *Monitoring matters – guidance on the development of pilot monitoring skills*. CAA.
- Dorneich, M. C., Rogers, W., Whitlow, S. D., & DeMers, R. (2016). Human performance risks and benefits of adaptive systems on the flight deck. *The International Journal of Aviation Psychology*, *26*(1–2), 15–35. <https://doi.org/10.1080/10508414.2016.1226834>
- Endsley, M. R. (1995). A taxonomy of situation awareness errors (No. 2). *Human factors in aviation operations*, 3. Avebury Aviation, Ashgate Publishing Ltd.
- Endsley, M. R. & Jones, D. G. (Eds.). (2012). *Designing for situation awareness: An approach to user-centered design* (2nd ed). CRC Press.
- Federal Aviation Administration. (Eds.). (2017). *Standard operating procedures and pilot Monitoring duties for flight deck crewmembers* [120–71B]. https://www.faa.gov/regulations_policies/advisory_circulars/index.cfm/go/document.information/documentid/1030486
- Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024. <https://doi.org/10.1177/0018720812443983>
- Fortmann, F., & Mengerlinghausen, T. (2014). Development and evaluation of an assistant system to aid monitoring behavior during multi-UAV supervisory control. In C. Stary (Ed.), *Proceedings of the 2014 European conference on cognitive ergonomics – ECCE '14* (pp. 1–8). ACM Press. <https://doi.org/10.1145/2637248.2637257>
- Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2018). Examining the relationship between construction workers' visual attention and situation awareness under fall and tripping hazard conditions: Using mobile eye tracking. *Journal of Construction Engineering and Management*, *144*(7), Article 4018060. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001516](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001516)
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. <https://doi.org/10.1037/0033-295X.87.4.329>
- Katidioti, I., Borst, J. P., Bierens de Haan, D. J., Pepping, T., van Vugt, M. K., & Taatgen, N. A. (2016). Interrupted by your pupil: An interruption management system based on pupil dilation. *International Journal of Human-Computer Interaction*, *32*(10), 791–801. <https://doi.org/10.1080/10447318.2016.1198525>
- Kelly, D., & Efthymiou, M. (2019). An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *Journal of Safety Research*, *69*, 155–165. <https://doi.org/10.1016/j.jsr.2019.03.009>
- Kennedy, K. D., Stephens, C. L., Williams, R. A., & Schutte, P. C. (2017). Repeated Induction of inattention blindness in a simulated aviation environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1959–1963. <https://doi.org/10.1177/1541931213601969>
- Kharoufah, H., Murray, J., Baxter, G., & Wild, G. (2018). A review of human factors causations in commercial air transport accidents and incidents: From 2000–2016. *Progress in Aerospace Sciences*, *99*, 1–13. <https://doi.org/10.1016/j.paerosci.2018.03.002>
- Lounis, C., Peysakhovich, V., & Causse, M. (2020). Flight eye tracking assistant (FETA): Proof of concept. In N. Stanton (Ed.), *Advances in intelligent systems and computing. Advances in human factors of transportation* (Vol. 964, pp. 739–751). Springer International Publishing. https://doi.org/10.1007/978-3-030-20503-4_66
- Lounis, C., Peysakhovich, V., & Causse, M. (2021). Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study. *PLoS One*, *16*(2), Article e0247061. <https://doi.org/10.1371/journal.pone.0247061>
- McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, *17*(1), 1–61. https://doi.org/10.1207/S15327051HCI1701_1
- Moore, K., & Gugerty, L. (2010). Development of a novel measure of situation awareness: The case for eye movement analysis. *Proceedings of the Human Factors and Ergonomics Society*, *54*.
- National Highway Traffic Safety Administration. (2018). *Crash stats: Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey*. <https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812506>
- PeiBl, S., Wickens, C. D., & Baruah, R. (2018). Eye-tracking measures in aviation: A selective literature review. *The International Journal of Aerospace Psychology*, *28*(3–4), 98–112. <https://doi.org/10.1080/24721840.2018.1514978>
- Ramón Alamán, J., Causse, M., & Peysakhovich, V. (2020). *Attentional span of aircraft pilots: Did you look at the speed?* <https://doi.org/10.3929/ethz-b-000407659>
- Reback, J., McKinney, W., Jbrockmendel, van den Bossche, J., Augspurger, T., Cloud, P., Gfyoung, Hawkins, S., Sinhrks., Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Garcia, M., Patrick, J., Schendel, J., ... H-Vetinari. (2021). *pandas-dev/pandas: Pandas 1.2.3* [Computer software].
- Rouse, W. B. (1988). Adaptive aiding for human/computer control. *Human Factors*, *30*(4), 431–443. <https://doi.org/10.1177/001872088803000405>
- Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring situation awareness in complex systems: comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500. <https://doi.org/10.1016/j.ergon.2008.10.010>
- Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and

- eye-tracking data. *Human Factors*, 49(3), 347–357. <https://doi.org/10.1518/001872007X196685>
- Schwarz, J., & Fuchs, S. (2017). Multidimensional real-time assessment of user state and performance to trigger dynamic system adaptation. In D. D. Schmorow & C. M. Fidopiastis (Eds.), *Lecture notes in computer science. Augmented cognition, neurocognition and machine learning* (Vol. 10284, pp. 383–398). Springer International Publishing. https://doi.org/10.1007/978-3-319-58628-1_30
- Schwerd, S., & Schulte, A. (2020). Experimental validation of an eye-tracking-based computational method for continuous situation awareness assessment in an aircraft cockpit. In D. Harris & W.-C. Li (Eds.), *Lecture notes in computer science. Engineering psychology and cognitive ergonomics. Cognition and design* (Vol. 12187, pp. 412–425). Springer International Publishing. https://doi.org/10.1007/978-3-030-49183-3_32
- Schwerd, S., & Schulte, A. (2021). Measuring the deviation between ground truth and operator awareness in a UAV management scenario: An eye-tracking approach. In *AIAA Scitech 2021 Forum*. American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2021-1579>
- Schwerd, S., & Schulte, A. (2022). Experimental Assessment of Fixation-Based Attention Measurement in an Aircraft Cockpit. In D. Harris & W.-C. Li (Eds.), *Lecture notes in computer science: Engineering Psychology and cognitive ergonomics* (Vol. 13307, pp. 408–419). Springer International Publishing. https://doi.org/10.1007/978-3-031-06086-1_32
- Silva, S. S., & Hansman, R. J. (2015). Divergence between flight crew mental model and aircraft system state in auto-throttle mode confusion accident and incident cases. *Journal of Cognitive Engineering and Decision Making*, 9(4), 312–328. <https://doi.org/10.1177/1555343415597344>
- van de Merwe, K., van Dijk, H., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology*, 22(1), 78–95. <https://doi.org/10.1080/10508414.2012.635129>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Winter, J. C. F. de., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, 21(1), 99–111. <https://doi.org/10.1007/s10111-018-0527-6>
- Zhang, T., Yang, J., Liang, N., Pitts, B. J., Prakah-Asante, K., Curry, R., Duerstock, B., Wachs, J. P., & Yu, D. (2023). Physiological measurements of situation awareness: A systematic review. *Human Factors*, 65(5), 737–758. <https://doi.org/10.1177/0018720820969071>
- Ziv, G. (2016). Gaze behavior and visual attention: A review of eye tracking studies in aviation. *The International Journal of Aviation Psychology*, 26(3–4), 75–104. <https://doi.org/10.1080/10508414.2017.1313096>

History

Received August 3, 2023

Revision received February 8, 2024

Accepted March 1, 2024

Published online May 13, 2024

Acknowledgments

This study is based on research conducted as part of the Simon Schwerd's PhD carried out in the Institute of Flight Systems of the Bundeswehr University, Munich, Germany. We would like to thank Julius Hoffelner, who supported by carrying out parts of the experiment as part of his master's thesis.

Publication Ethics

Informed consent was obtained from all participants included in the study. All procedures in studies involving human participants were performed in accordance with the ethical standards of the institution's Human Research Ethics Committee and the UniBwM.

Open Data


Data not available due to legal restrictions. Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, and thus supporting data are not available.

Funding

Open access publication enabled by University of the Bundeswehr Munich.

ORCID

Simon Schwerd

 <https://orcid.org/0000-0001-6950-2226>

Simon Schwerd

Institute of Flight Systems
University of the Bundeswehr Munich
Werner-Heisenberg-Weg 39
85579, Neubiberg
Germany
simon.schwerd@uniwb.de



Simon Schwerd received his MSc in mechanical engineering from the Technical University of Munich, Germany. After that, he earned a doctoral degree in aerospace engineering from the University of the Bundeswehr Munich, in 2023. His research interests include flight crew cognitive state estimation and adaptive systems for aircraft cockpits.



Axel Schulte received his Doctor of Engineering degree from the University of the Bundeswehr Munich, Germany. Since 2006, he has been a professor of aircraft dynamics and flight guidance there, and since 2008, the Director of the Institute of Flight Systems. His research interests include human-autonomy teaming and cognitive automation in aviation.

Appendix

Table A1. Distance functions for information categories

Types of information	Distance function $\text{dist}(x_1, x_2)$	Example
Numeric (integer, floating-point)	$x_2 - x_1$	Aircraft altitude, engine temperature, magnetic heading
Text	$\begin{cases} 0.0 & \text{if } x_1 \equiv x_2 \\ 1.0 & \text{otherwise} \end{cases}$	Text message
Booleans		Gear fully extended
Modes		Autopilot hold/acquire status
Positional	Great circle distance in meters between x_1 and x_2	Aircraft position, airport position

Table A2. Pseudocode of trigger mechanism

ADAPTIVE-ALERTING (information, c-norm, time-until-notification, threshold = .6)

time-of-high-AD = NIL

WHILE adaptive-alert-active

 pilot-value = LAST-FIXATED-VALUE-OF (information)

 system-value = SYSTEM-VALUE-OF (information)

 aw-deviation = COMPUTE-AW-DEVIATION (pilot-value, system-value, c-norm)

IF aw_deviation > threshold

 NOW = GET-TIME ()

IF time-of-high-deviation == NIL

 time-of-high-AD = now

IF now - time-of-high-deviation > time-until-notification

 TRIGGER-ALERT (information)

 time-of-high-AD = NIL

ELSE

 time-of-high-AD = NIL