Applied Ergonomics 125 (2025) 104461

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

When automation fails - Investigating cognitive stability and flexibility in a multitasking scenario

Sophie-Marie Stasch *, Wolfgang Mack

Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85579, Neubiberg, Germany

ARTICLE INFO

Keywords:
automation reliability
cognitive control
multitasking

ABSTRACT

Managing multiple tasks simultaneously often results in performance decrements due to limited cognitive resources. Task prioritization, requiring effective cognitive control, is a strategy to mitigate these effects and is influenced by the stability-flexibility dilemma. While previous studies have investigated the stability-flexibility dilemma in fully manual multitasking environments, this study explores how cognitive control modes interact with automation reliability. While no significant interaction between control mode and automation reliability was observed in single multitasking performance, our findings demonstrate that overall task performance benefits from a flexible cognitive control mode when automation is reliable. However, when automation is unreliable, a stable cognitive control mode improves manual takeover performance, though this comes at the expense of secondary task performance. Furthermore, cognitive control modes and automation reliability independently affect various eye-tracking metrics and mental workload. These findings underscore the need to integrate cognitive control and automation reliability into adaptive assistance systems, particularly during the perceive stage, to enhance safety in human-machine systems.

1. Introduction

1.1. Multitasking and cognitive control

Concurrent multitasking, defined as the simultaneous execution of at least two tasks (Salvucci and Taatgen, 2008; Fischer and Plessow, 2015; Koch et al., 2018), is a routine aspect of an operator's day-to-day activities in the human factor's context. Examples include pilots operating an aircraft, air traffic controllers guiding a large number of planes at an airport, or operators in a power plant managing a variety of control systems. To manage information overload within the limits of cognitive resources, operators often rely on task prioritization strategies (Hoover, 2008). A well-known example is the A-N-C-M (aviate – navigate – communicate - manage) axiom, advising pilots to prioritize the most safety-critical aviate task in mental overload situations.

Cognitive control, a collection of processes involved in generating and maintaining context-appropriate tasks and suppressing irrelevant goals (Gratton et al., 2018), plays an important role in effective task prioritization and is closely related to attentional processes (Mackie et al., 2013). Research on multitasking behavior shows that cognitive control is subject to the stability-flexibility-dilemma (Musslick et al., 2018; Goschke and Bolte, 2014; Dreisbach and Fröber, 2019). This dilemma describes the antagonistic demands on cognitive control operators need to balance in a concurrent multitasking scenario and remains rather unresearched in the field of human factors. While cognitive stability allows for efficient goal pursuit and the ability to ignore irrelevant distractions, it can also impair the ability to switch tasks efficiently. Cognitive flexibility allows for quick reactions to unexpected events and facilitates task switches, but can also increase distractibility by task irrelevant information. Balancing these opposing demands is therefore critical for effective multitasking.

Specifically, the meta-control state model (Hommel, 2015) explains how the cognitive system can switch between stable and flexible control modes based on goal modulation and mutual inhibition between competing task representations. Hereby, cognitive stability (persistence) is achieved by either strengthening the top-down influence of a goal—favoring one representation over others—or by increasing mutual inhibition, which suppresses alternative representations and reduces the likelihood of spontaneous task switching. Conversely, cognitive flexibility is facilitated by reducing either the goal's influence or the inhibition between alternatives, allowing for easier switching between task representations. Importantly, the number of voluntary task switches is

* Corresponding author.
E-mail address: sophie-marie.stasch@unibw.de (S.-M. Stasch).

https://doi.org/10.1016/j.apergo.2024.104461
Received 28 August 2024; Received in revised form 17 December 2024; Accepted 21 December 2024
Available online 3 January 2025

considered as a correlate of cognitive control, whereby a higher number of voluntary task switches is linked to cognitive flexibility and a lower number of voluntary task switches is linked to cognitive stability (Fröber and Dreisbach, 2017).

Recent research suggests that cognitive stability and cognitive flexibility may function as independent dimensions rather than as opposing ends of a continuum (Egner, 2023; Geddert and Egner, 2022). For simplicity, this study refers to a stable but inflexible control state as the "stable control mode" and a flexible but unstable control state as the "flexible control mode," without ruling out the existence of other combinations. Furthermore, current research on the stability-flexibility dilemma is predominantly rooted in basic cognitive science, with a focus on its effects on multitasking performance in simple stimulus-response tasks. Only few studies have explored how this dilemma manifests in more applied contexts.

For instance, Stasch and Mack (2023a) and Stasch et al. (2024) manipulated the stability-flexibility-dilemma using a gamification method (Stasch and Mack, 2023b) in a low-fidelity flight environment. Their findings demonstrate the dilemma's effect on performance, mental workload and different eye-tracking metrics when participants were performing the experimental task. Participants in a stable control mode performed better at the prioritized task at the cost of the subtask's performance, whereas a flexible control mode enhanced subtask performance at the expense of the main task. Notably, the tasks were performed manually without assistance through —a feature prevalent in modern human-machine systems. This points to a research gap, as the influence of automation on such performance trade-offs remains currently underexplored.

1.2. Automation reliability

Nowadays, most human-machine systems are equipped with automation functions at varying automation levels (Parasuraman et al., 2000) intended to reduce task load. However, these systems are not infallible and can become unreliable. Reasons for that include, but are not limited to, sensor failures, sudden external changes, cyber-attacks, incorrect system configurations, neglected maintenance or inter-system communication failures. When automation fails, the operator must manually assume control of the task, which can significantly impact human-machine performance (Vogelpohl et al., 2018; Eriksson and Stranton, 2017).

The negative effects of unreliably automated systems on performance were first noted by Bainbridge (1983), who termed this in the "Ironies of Automation". Subsequent research has corroborated these findings across various human factors domains, including command-and-control environments (Rovira, 2002), military multitasking environments (Chen et al., 2011), aviation (Dixon and Wickens, 2004), air traffic control (ATC; Rovira and Parasuraman, 2007; Trapsilawati et al., 2015), and automated driving (Strand et al., 2014).

There is evidence suggesting that in some cases, manual task execution may outperform imperfect automation. For example, Metzger and Parasuraman (2017) found that participants using a medium-fidelity ATC simulator were more likely to detect a conflict when operating manually than when relying on unreliable automation. This decline in performance with automation failure can be partly explained by the "out-of-the-loop" phenomenon (Endsley and Kiris, 1995), where a loss of manual skills and situational awareness occurs, a problem not encountered during fully manual operation. However, it is also important to note that Xu et al. (2007) found some benefit to using imperfect automation, particularly when the system reliability exceeded 80%.

1.3. Adaptive automation

Out-of-the-loop experiences, where operators become disengaged from the automation process in a static task allocation between the human and the machine, can significantly impact performance in safety-critical systems by making it difficult for them to respond effectively to sudden changes in the environment or task demands. Adaptive assistance systems aim to mitigate these challenges by dynamically adjusting the level of automation based on the operator's needs, thereby maintaining engagement and awareness, which are essential for reliable human-machine interaction. Feighs et al. (2012) describe a framework of adaptive systems, in which the system monitors the operator's mental state, their tasks, as well as the environment, with sensors and information systems ("perceive"-stage). Based on that assessment, an adaptation manager selects context-appropriate adaptations ("selection"-stage), which are then transferred to the automation and the human-machine interface ("act"-stage). The benefits of adaptive assistance have been demonstrated in various contexts, such as preventing out-of-the-loop experiences in air traffic control (Di Flumeri et al., 2019) and enhancing naval officers' performance in a high-fidelity command-and-control environment compared to static automation (de Tjerk et al., 2010).

A key requirement for effective adaptive assistance is the accurate estimation of the operator's mental state during the "perceive" stage. Stasch and Mack (2024) have contributed to this area by demonstrating that control modes (stable and flexible) can be diagnosed with high accuracy using eye-tracking metrics in a virtual flight environment. However, their study was limited to manual task performance, leaving the interaction between control modes and varying levels of automation reliability unexplored.

1.4. Research question

Given the increasing complexity of multitasking scenarios in the modern workplace, it is crucial to investigate how cognitive control modes (flexible and stable) influence performance under varying levels of automation reliability. Understanding this relationship is essential not only for advancing the theoretical framework of adaptive systems but also for enhancing practical implementations that ensure safety and efficiency in human-machine interactions.

Thus, the research question posed is: How does cognitive control mode influence performance under conditions of reliable and unreliable automation in a multitasking environment? The subsequent study aims to investigate the interaction between cognitive control modes (stable and flexible) and automation reliability (reliable and unreliable) to improve safety and efficiency in human-machine interactions. Specifically, the study seeks to.

1. Provide insights on how the cognitive control mode of operators can be integrated into the perceive stage of adaptive automation systems respecting a tasks automation level.
2. Make recommendations regarding the interface design of adaptive assistance systems that respect the cognitive control modes of operators.

2. Methods

2.1. Experimental task

To address this research question, the open-source version of the MATB-II (Santiago-Espada et al., 2011), known as openMATB (Cegarra et al., 2020), was used as a virtual flight environment, similar to the approach of Stasch and Mack (2024). The MATB is a widely recognized tool for studying the cognitive demands associated with performing multiple flight-like tasks simultaneously. The environment comprises four distinct subtasks (see Fig. 1).

1. **Tracking Task:** Participants use a joystick to maintain a cursor at the center of a blue square, with the cursor moving in a sinusoidal pattern.

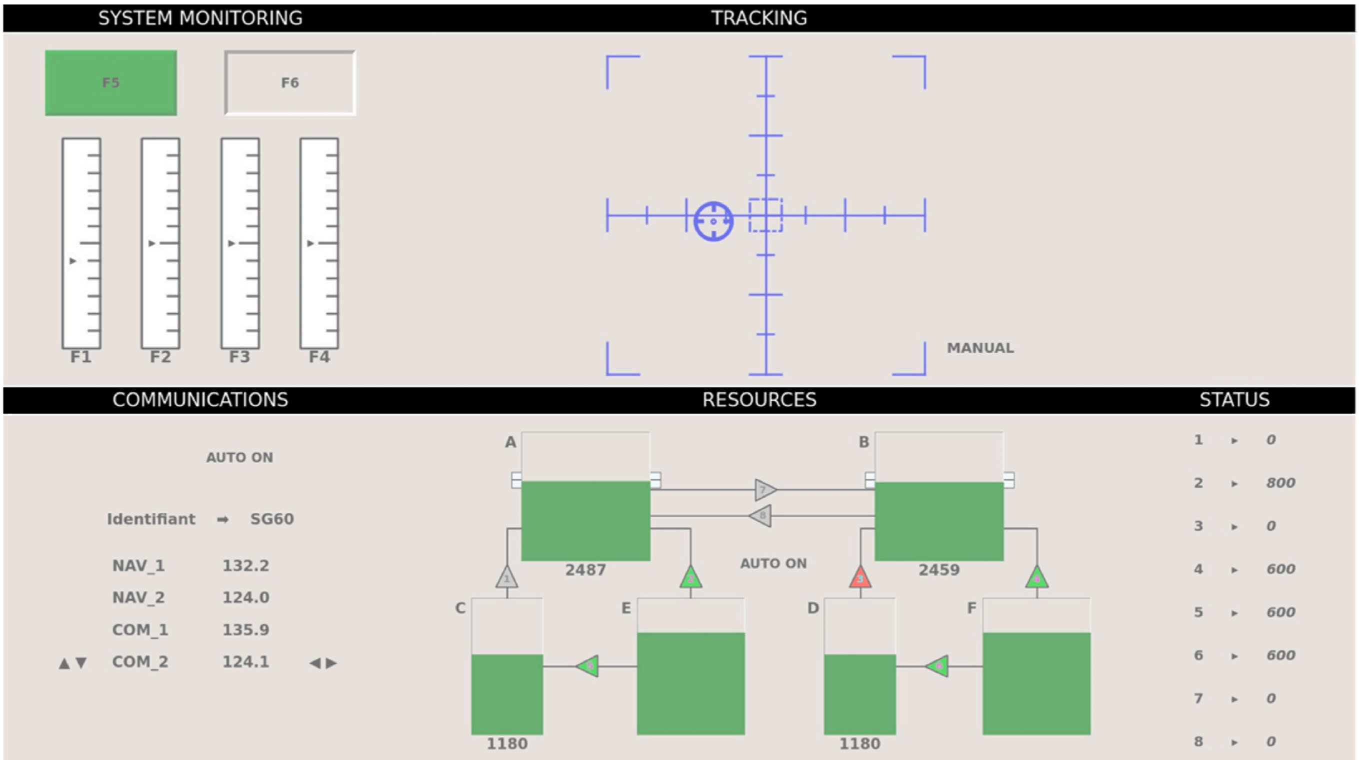

Fig. 1. The open-MATB.
Note. Top left: System-Monitoring task, top right: Tracking task, bottom left: Communication task, bottom right: Resource Management task.

2. **System Monitoring Task:** Participants monitor F1-F4 scale indicators for deviations, pressing the corresponding F-key on the keyboard to respond. Additionally, they react to color changes in the F5 and F6 buttons.
3. **Communication Task:** Participants respond to their designated call sign by adjusting the announced radio frequency on one of four radios (NAV1, NAV2, COM1, COM2).
4. **Resource Management Task:** The goal is to maintain optimal fuel levels in both main tanks (Tank A and Tank B). Participants manage fuel levels by toggling pumps 1–8 on or off using the corresponding number keys on the keyboard.

Hereafter, the tracking task will also be referred to as the main task, while the system-monitoring task, communication task, and resource management task will also be referred to as subtasks.

2.2. Participants

Fourty-eight subjects participated in the experiment, aged 18 to 59 (M = 23.46; SD = 6.48). All participants had normal or corrected-to-normal vision and no red-green color deficiency. The sample consisted of 15 participants identifying as female and 33 participants identifying as male. Additionally, 22.91% of participants has previous flight experience with an average of 364.36 flight hours (SD = 585.65). The study was approved by the local Ethics Committee of the University of the Bundeswehr Munich (EK UniBw M 23–50). All participants provided written informed consent before participating in the study and received a lab token as compensation.

2.3. Experimental design and procedure

A 2x3 within factorial design was employed with two factors: control mode (flexible, stable) and automation mode (reliable, unreliable, manual), resulting in six conditions (flexible reliable, flexible unreliable, flexible manual, stable reliable, stable unreliable, stable manual). Automation mode variations were applied to the tracking task. In the reliable condition, the tracking task was fully automated; in the unreliable condition, automation failed 50% of the time; and in the manual condition, participants had full manual control of the tracking task. Control modes were manipulated using the Stasch and Mack (2023b) gamification method. In this method, participants were instructed prior to one experimental condition on how to prioritize the MATB tasks based on simulated weather conditions. Specifically, participants were instructed to prioritize the tracking task during stormy weather conditions to induce a stable control mode. Conversely, a flexible control mode was induced by instructing participants to prioritize all tasks equally due to good weather conditions. One condition consisted of five trials.

After each trial, participants received a feedback score ranging from 0 to 100, based on the number of task switches performed, along with a brief prompt on how to prioritize the tasks. A task switch was registered when a fixation was assessed on a task different from the last task that received a fixation. The total number of task switches was assessed during each trial and normalized according to values obtained in a previous experiment using a similar scenario. The scoring system was designed to reinforce task prioritization according to the condition in which participants were placed. In the stable control mode, participants received higher scores for predominantly focusing on the tracking task by switching less between subtasks. In contrast, in the flexible control mode, participants were rewarded for distributing their attention equally across all subtasks by switching more frequently between tasks. This approach was chosen because a flexible control mode is associated with a higher number of task switches, while a stable control mode is associated with a lower number of task switches (Fröber and Dreisbach, 2017).

Participants were trained on all tasks before the experiment. The order of conditions was randomized to minimize order effects. Each condition involved five trials, followed by a brief questionnaire assessing mental workload (NASA-TLX; Hart and Staveland, 1988) and trust in automation (Jian et al., 2000). Eye movement data were captured at 1000 Hz using the EyeLink 1000 Plus in a head-fixed position. After completing the experiment, participants filled out a demographic questionnaire.

2.4. Data processing

For the tracking task, the root-mean-square error (RMSE) was calculated during instances of manual control in the unreliable condition and across the entire trial in the manual control condition. The RMSE was also computed for the first 2 s after transitioning from automatic to manual mode in the unreliable condition. Performance metrics for the system monitoring and communication tasks included the proportion of hits and reaction times. For the resource management task, the average deviation from optimal fuel levels in Tank A and Tank B was recorded. Performance data were averaged across trials for each condition and participant.

Total MATB performance was calculated by summing z-scores, with sign changes applied where necessary to ensure higher scores indicated better performance. For the system monitoring and communication tasks, mean reaction time and hit rate were combined. Eye movement data were processed using EyeLink Data Viewer, with all values standardized (z-scores).

Specifically, the AOI-unspecific metrics (see Section 3.2.5) were calculated as follows: A task switch was identified when a fixation on one task differed from the previous fixation on another task. The coefficient K is a measure derived from saccade amplitudes and fixation durations. Positive values of K are associated with ambient visual processing, while negative values are linked to focal visual processing (Krejtz et al., 2016). Entropy is a metric that reflects the predictability of transitions between tasks (AOIs). Stationary entropy indicates how equally attention is distributed across all AOIs, while transition entropy reflects the randomness and predictability of transitions between AOIs (Krejtz et al., 2014; Nahlik and Daubenmire, 2022).

Mental workload was calculated by summing and z-standardizing all NASA-TLX subscales. Similarly, values of trust and mistrust in automation were averaged and z-standardized, excluding the manual automation condition, as it does not apply.

2.5. Statistical analysis

Mixed linear effects models were computed using the lme4 package (Bates et al., 2015) in RStudio Version 4.3.3 (R Core Team, 2024). Predictors, including control mode (stable vs. flexible) and automation mode (reliable, unreliable, manual), were sequentially added to the null model using a stepwise procedure. A stepwise procedure was employed to evaluate the incremental improvement in model fit with the addition of each fixed effect. Building on prior findings by Stasch and Mack (2023a), which demonstrated a significant effect of control mode on MATB performance and various eye-tracking metrics, control mode was selected as the initial fixed effect in Model 1. Subsequently, automation level was added as a second fixed factor to determine whether it provided additional explanatory power. The best-fitting model was selected using a Chi-Square test, with participants included as a random effect to account for individual difference. Post hoc contrasts were performed using the emmeans package (Lenth, 2024) on the best fitting models that included an interaction term (model 3).

The models were built as follows:

$$\text{Model } 0_{ij} = \beta_0 + u_{0j} + \epsilon_{ij}$$

$$\text{Model } 1_{ij} = \beta_0 + \beta_1 \times \text{Control\_mode}_{ij} + u_{0j} + \epsilon_{ij}$$

$$\text{Model } 2_{ij} = \beta_0 + \beta_1 \times \text{Control\_mode}_{ij} + \beta_2 \times \text{Automation\_level}_{ij} + u_{0j} + \epsilon_{ij}$$

$$\text{Model } 3_{ij} = \beta_0 + \beta_1 \times \text{Control\_mode}_{ij} + \beta_2 \times \text{Automation\_level}_{ij} + \beta_3 \times (\text{Control\_mode}_{ij} \times \text{Automation\_level}_{ij}) + u_{0j} + \epsilon_{ij}$$

Hereby:
β_0 = Intercept
u_{0j} = Random effect (random intercept) for subject *j*
ϵ_{ij} = Residual error term for observation *i* within subject *j*
β_1 = Coefficient for the control mode
β_2 = Coefficient for the automation level
β_3 = Coefficient for interaction between the control mode and the automation level.

3. Results

Detailed results about the model comparisons can be found in the following repository (https://osf.io/tf9vk/). See Fig. 2 for a visualization of performance effects, Fig. 3 for visualization of effects on number of fixations and Fig. 4 for visualization of effects on mean fixation duration.

3.1. Performance

Tracking Task. The reliable tracking condition was excluded from analysis since RMSE cannot be calculated with task automation. Model 1, including the control mode, best explained the data, showing that the stable condition negatively affected the RMSE ($M_{flexible} = 0.07$; $SD_{flexible} = 0.02$; $M_{stable} = 0.06$; $SD_{stable} = 0.02$). Additionally, the RMSE in the 2-s interval after manual takeover with unreliable automation was lower in the stable condition ($M_{stable} = 0.07$; $SD_{stable} = 0.02$) than in the flexible one ($M_{flexible} = 0.07$; $SD_{flexible} = 0.02$).

System Monitoring Task. Model 2, which includes the control mode and automation reliability, best explained the *number of hits*, indicating that the stable condition reduced hits ($M_{flexible} = 5.37$; $SD_{flexible} = 1.15$; $M_{stable} = 4.66$; $SD_{stable} = 1.87$), while reliable automation increased them ($M_{manual} = 4.77$; $SD_{manual} = 1.80$; $M_{unreliable} = 5.06$; $SD_{unreliable} = 1.56$; $M_{reliable} = 5.22$; $SD_{reliable} = 1.35$). No significant effect was found for unreliable automation (see Table 1). For *reaction time*, Model 1, which includes control mode, fits best, showing that the stable condition increased reaction time ($M_{flexible} = 3.05$; $SD_{flexible} = 1.02$; $M_{stable} = 3.19$; $SD_{stable} = 1.14$).

Communication Task. The *number of hits* and the *reaction time* was best explained by Model 2, which includes control mode and automation reliability. The stable condition decreased the *number of hits* ($M_{flexible} = 1.65$; $SD_{flexible} = 0.62$; $M_{stable} = 1.39$; $SD_{stable} = 0.79$) and increased *reaction time* ($M_{flexible} = 16.66$; $SD_{flexible} = 2.52$; $M_{stable} = 17.23$; $SD_{stable} = 2.84$), while reliable automation increased the *number of hits* ($M_{manual} = 1.46$; $SD_{manual} = 0.78$; $M_{unreliable} = 1.50$; $SD_{unreliable} = 0.72$; $M_{reliable} = 1.60$; $SD_{reliable} = 0.66$) and decreased *reaction time* ($M_{manual} = 17.26$; $SD_{manual} = 2.87$; $M_{unreliable} = 17.21$; $SD_{unreliable} = 2.83$; $M_{reliable} = 16.36$; $SD_{reliable} = 2.25$). No significant effect was found for both outcomes with unreliable automation (see Table 1).

Resource Management Task. Deviation from optimal fuel levels was best explained by Model 2, which includes control mode and automation reliability. The stable condition increased deviation ($M_{flexible} = 420.45$; $SD_{flexible} = 209.35$; $M_{stable} = 554.64$; $SD_{stable} = 361.06$), while the reliable condition decreased it ($M_{manual} = 535.23$; $SD_{manual} = 330.00$; $M_{unreliable} = 485.22$; $SD_{unreliable} = 299.70$; $M_{reliable} = 442.19$; $SD_{reliable} = 268.21$). No effect was found for unreliable automation (see Table 1).

Total Performance. Total performance (in z-scores) on the MATB was best explained by Model 2, including control mode and automation reliability. The stable condition led to worse performance compared to the flexible condition ($M_{flexible} = 0.31$; $SD_{flexible} = 1.68$; $M_{stable} = -0.12$; $SD_{stable} = 1.90$), while the reliable condition improved performance compared to the manual condition ($M_{manual} = -0.84$; $SD_{manual} = 1.66$; $M_{unreliable} = -0.28$; $SD_{unreliable} = 1.60$; $M_{reliable} = 1.43$; $SD_{reliable} = 1.23$; see Table 1).

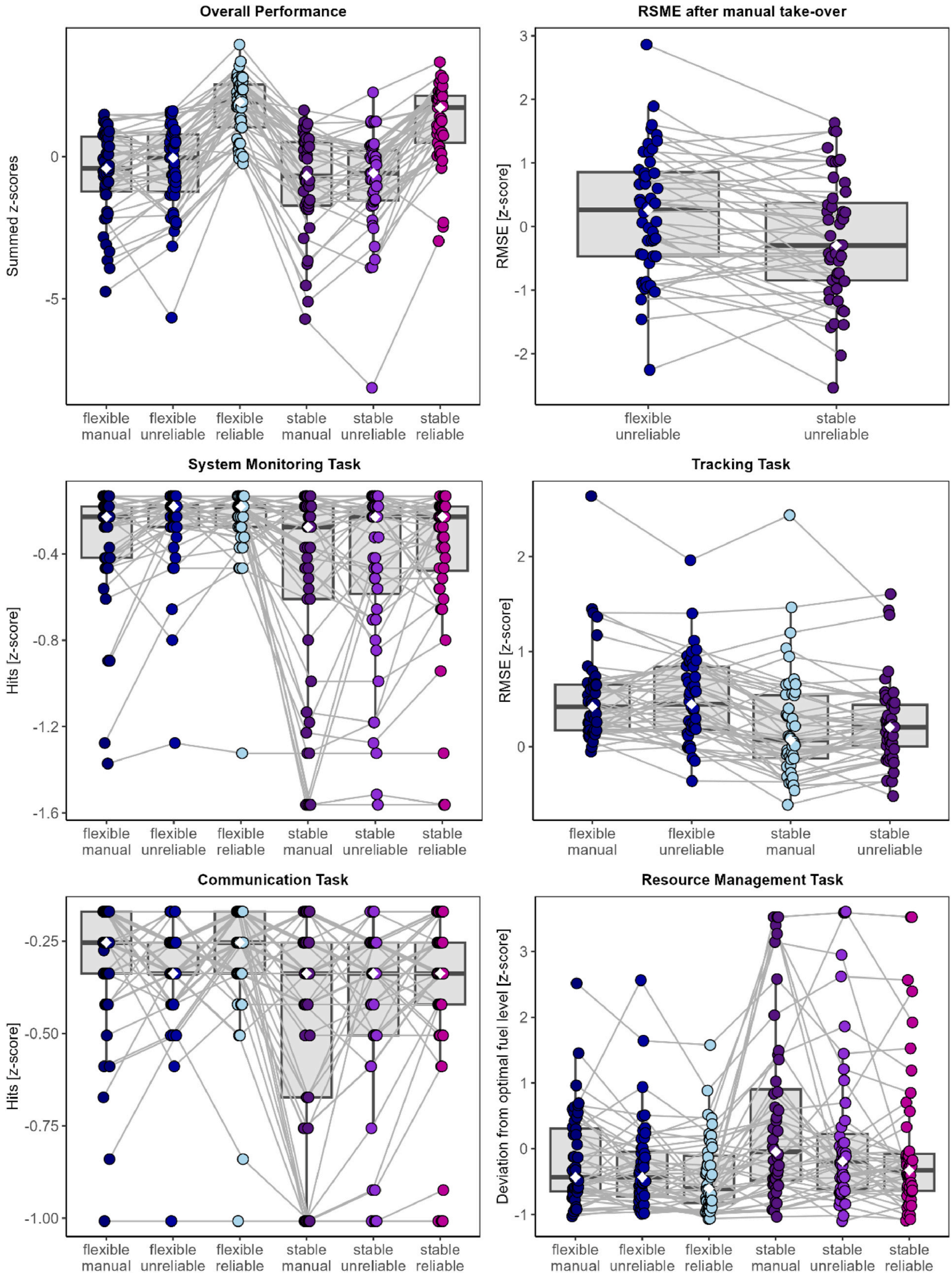

Fig. 2. Total MATB performance, subtask performance and performance after manual take-over.

Note. All values have been z-standardized in the plot. To enhance comprehensibility, performance of the overall MATB score required a sign change, so that higher z-values correspond to a better overall performance. RMSE after take-over has only been calculated for the unreliable condition in a 2-s time interval.

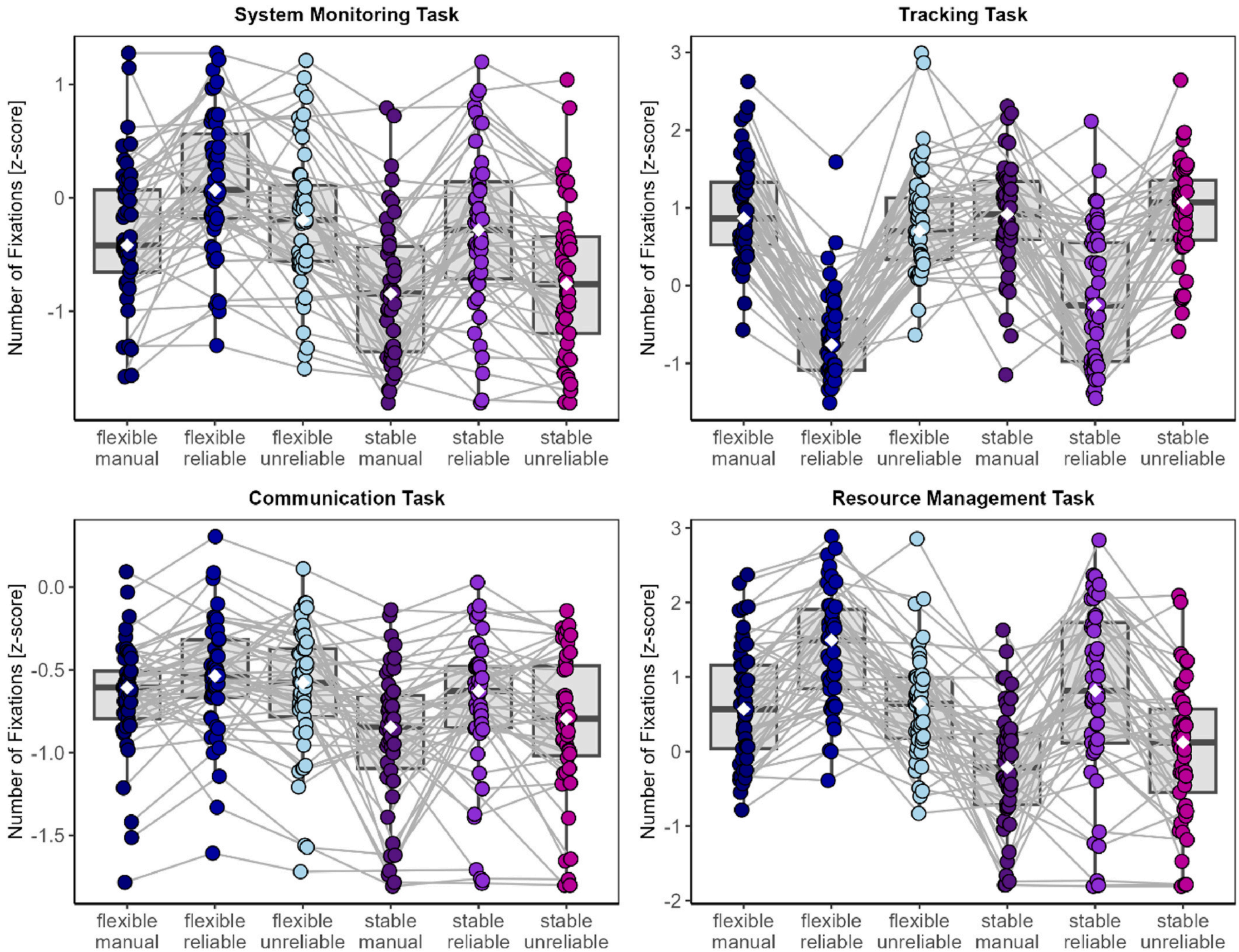

Fig. 3. Number of Fixations per MATB task.
Note. Plot displays z-transformed values.

3.2. Eye-tracking metrics

3.2.1. Number of fixations

Tracking Task. Model 3, which includes control mode, automation reliability, and their interaction, best explained the number of fixations. While a stable control mode ($M_{flexible}$ = 369.73; $SD_{flexible}$ = 169.75; M_{stable} = 413.71; SD_{stable} = 166.91) and unreliable automation (M_{manual} = 475.56; SD_{manual} = 122.20; $M_{unreliable}$ = 465.21; $SD_{unreliable}$ = 128.41; $M_{reliable}$ = 234.39; $SD_{reliable}$ = 133.34) had no effect, participants fixated the tracking task less often with reliable automation. However, a positive interaction indicated more fixations with reliable automation and a stable control mode ($M_{flexible\text{-}manual}$ = 473.35; $SD_{flexible\text{-}manual}$ = 113.34; $M_{stable\text{-}manual}$ = 477.77; $SD_{stable\text{-}manual}$ = 131.63; $M_{flexible\text{-}unreliable}$ = 447.67; $SD_{flexible\text{-}unreliable}$ = 121.29; $M_{stable\text{-}unreliable}$ = 482.75; $SD_{stable\text{-}unreliable}$ = 134.12; $M_{flexible\text{-}reliable}$ = 188.17; $SD_{flexible\text{-}reliable}$ = 96.34; $M_{stable\text{-}reliable}$ = 280.60; $SD_{stable\text{-}reliable}$ = 149.30; see Table 2). The post-hoc comparison of model 3 shows that participants made fewer fixations in the reliable automation condition compared to the manual and unreliable conditions (see Table 3). No significant difference in number of fixations was found between the manual and unreliable conditions in either control mode. However, in the reliable condition, participants made more fixations in the stable condition compared to the flexible condition. This interaction was not observed between the flexible and stable conditions for the manual and unreliable automation levels (see Table 4).

System Monitoring Task – Communication Task – Resource Management Task. Model 2, including control mode and automation reliability, best explained the number of fixations on the system monitoring task ($M_{flexible}$ = 286.66; $SD_{flexible}$ = 109.52; M_{stable} = 207.23; SD_{stable} = 119.18; M_{manual} = 212.98; SD_{manual} = 113.55; $M_{unreliable}$ = 232.56; $SD_{unreliable}$ = 117.17; $M_{reliable}$ = 295.42; $SD_{reliable}$ = 117.50), the communication task ($M_{flexible}$ = 203.92; $SD_{flexible}$ = 61.07; M_{stable} = 170.84; SD_{stable} = 73.22; M_{manual} = 175.19; SD_{manual} = 68.06; $M_{unreliable}$ = 185.97; $SD_{unreliable}$ = 69.80; $M_{reliable}$ = 201.39; $SD_{reliable}$ = 68), and the resource management task ($M_{flexible}$ = 461.40; $SD_{flexible}$ = 151.95; M_{stable} = 354.91; SD_{stable} = 185.80; M_{manual} = 345.15; SD_{manual} = 149.10; $M_{unreliable}$ = 366.06; $SD_{unreliable}$ = 143.50; $M_{reliable}$ = 513.15; $SD_{reliable}$ = 188.23). Participants fixated less on these tasks in the stable condition than in the flexible one, but more often with reliable automation (see Table 2).

3.2.2. Fixation duration

Tracking Task. Model 3, which includes control mode, automation reliability, and their interaction, best explained fixation duration. Participants had longer fixations in the stable condition than in the flexible one ($M_{flexible}$ = 294.75; $SD_{flexible}$ = 334.99; M_{stable} = 451.80; SD_{stable} = 798.53), with no differences across reliability levels (M_{manual} = 402.29; SD_{manual} = 711.80; $M_{unreliable}$ = 359.20; $SD_{unreliable}$ = 540.98; $M_{reliable}$ = 364.45; $SD_{reliable}$ = 611.27). However, the stable condition's effect on fixation duration was reduced with both reliable and unreliable

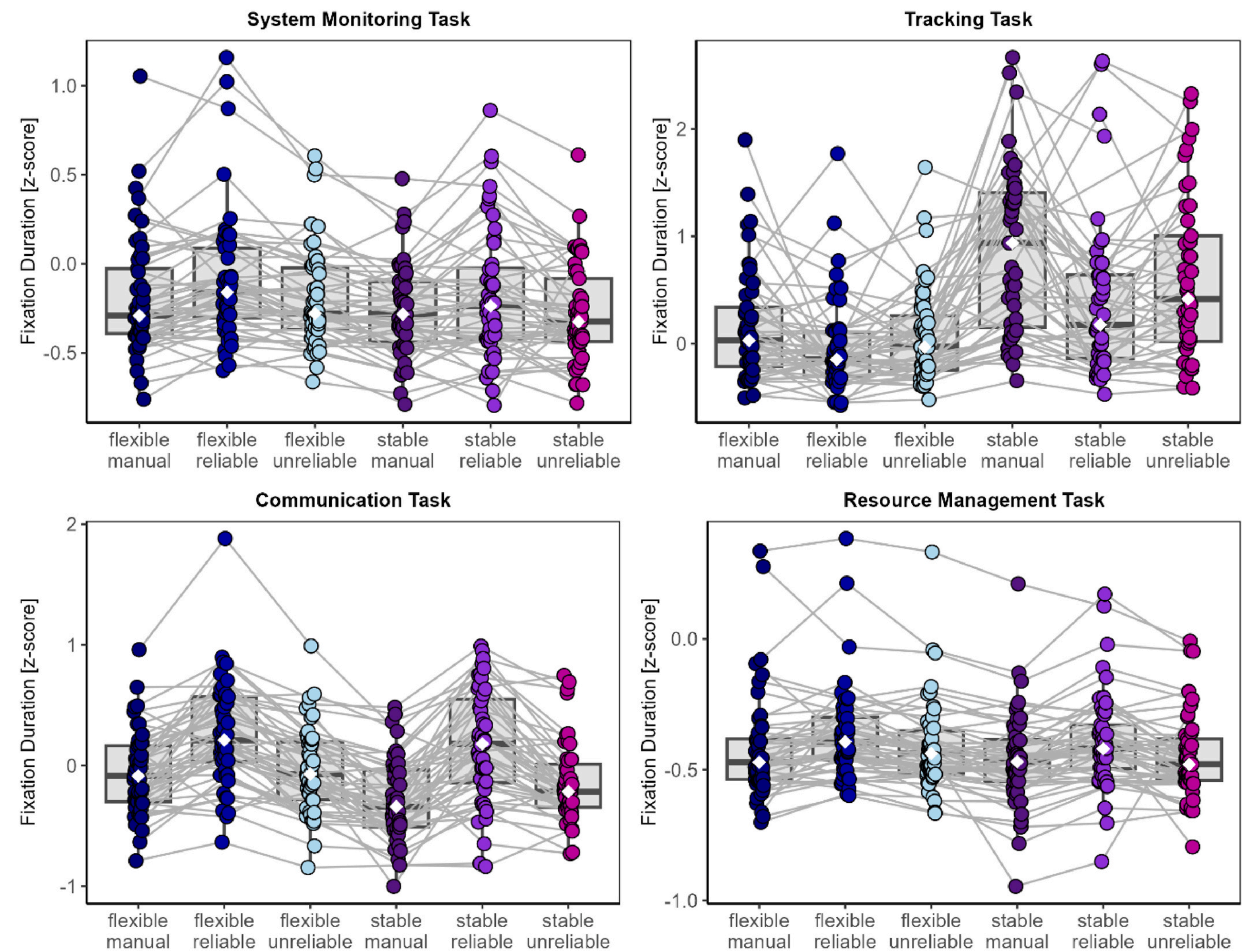

Fig. 4. Mean Fixation Duration per MATB subtask.
Note. Plot displays z-transformed values.

automation ($M_{flexible\text{-}manual}$ = 304.82; $SD_{flexible\text{-}manual}$ = 358.56; $M_{stable\text{-}manual}$ = 498.86; $SD_{stable\text{-}manual}$ = 928.73; $M_{flexible\text{-}unreliable}$ = 293.03; $SD_{flexible\text{-}unreliable}$ = 324.22; $M_{stable\text{-}unreliable}$ = 420.56; $SD_{stable\text{-}unreliable}$ = 677.32; $M_{flexible\text{-}reliable}$ = 273.54; $SD_{flexible\text{-}reliable}$ = 295.47; $M_{stable\text{-}reliable}$ = 425.41; $SD_{stable\text{-}reliable}$ = 745.94; see Table 5). Specifically, the post hoc comparison for the condition interaction effects of model 3indicates that in the stable condition, participants show significantly longer fixation durations in the manual automation level compared to both unreliable and reliable levels. This effect was not observed in the flexible condition (see Table 6). Across all automation levels (manual, unreliable, reliable), fixation durations are longer in the stable condition compared to the flexible condition (see Table 7).

Communication Task – Model 2, including control mode and automation reliability, best explained fixation duration. Fixation duration decreased in the stable condition ($M_{flexible}$ = 294.31; $SD_{flexible}$ = 403.18; M_{stable} = 278.23; SD_{stable} = 367.66) and increased with unreliable automation (M_{manual} = 256.35; SD_{manual} = 287.55; $M_{unreliable}$ = 268.62; $SD_{unreliable}$ = 315.45; $M_{reliable}$ = 329.98; $SD_{reliable}$ = 500.95). This increase in fixation duration was even stronger with reliable automation (see Table 5).

System Monitoring – Resource Management Task – Model 2, including control mode and automation reliability, best explained fixation duration. The stable control mode decreased fixation duration for the system monitoring task ($M_{flexible}$ = 248.37; $SD_{flexible}$ = 172.70; M_{stable} = 242.83; SD_{stable} = 166.19) and the resource management task ($M_{flexible}$ = 202.20; $SD_{flexible}$ = 122.43; M_{stable} = 200.07; SD_{stable} = 115.78), while reliable automation increased it, both for the system monitoring task (M_{manual} = 241.91; SD_{manual} = 144.59; $M_{unreliable}$ = 237.18; $SD_{unreliable}$ = 134.50; $M_{reliable}$ = 255.99; $SD_{reliable}$ = 207.16) and the resource management task (M_{manual} = 196.90; SD_{manual} = 113.14; $M_{unreliable}$ = 197.53; $SD_{unreliable}$ = 109.90; $M_{reliable}$ = 206.86; $SD_{reliable}$ = 129.77). No effect was found for unreliable automation (see Table 5).

3.2.3. AOI unspecific metrics

Task Switches – Model 3, including automation reliability, control mode, and their interaction, best explained the number of task switches. Participants switched tasks less frequently in the stable condition compared to the flexible one ($M_{flexible}$ = 478.12; $SD_{flexible}$ = 131.69; M_{stable} = 402.68; SD_{stable} = 154.86). They also switched tasks less often with reliable automation (M_{manual} = 485.23; SD_{manual} = 164.53; $M_{unreliable}$ = 469.36; $SD_{unreliable}$ = 139.90; $M_{reliable}$ = 366.60; $SD_{reliable}$ = 109.19), except when being in the stable condition. In that case, participants switched tasks more often, as indicated by a significant positive interaction between stable control and reliable automation ($M_{flexible\text{-}manual}$ = 543.13; $SD_{flexible\text{-}manual}$ = 127.09; $M_{stable\text{-}manual}$ = 427.33; $SD_{stable\text{-}manual}$ = 178.10; $M_{flexible\text{-}unreliable}$ = 514.29; $SD_{flexible\text{-}unreliable}$ = 109.48; $M_{stable\text{-}unreliable}$ = 424.44; $SD_{stable\text{-}unreliable}$ = 153.84; $M_{flexible\text{-}reliable}$ = 376.94; $SD_{flexible\text{-}reliable}$ = 376.94; $M_{stable\text{-}reliable}$ = 356.27; $SD_{stable\text{-}reliable}$ = 119.46; see

Table 1
Results of the best-fitting linear-mixed models explaining MATB performance.

Task	b	CI	t	p
Tracking (RMSE)				
stable	−0.27	[-0.34, −0.20]	−7.91	<0.001
Tracking (RMSE take-over)				
stable	−0.42	[-0.66, −0.19]	−3.58	<0.001
System Monitoring (RT)				
stable	0.22	[0.05, 0.40]	2.51	0.013
System Monitoring (Hits)				
stable	−0.17	[-0.23, −0.12]	−6.25	<0.001
unreliable	0.07	[-1.39e^{-03}, 0.13]	1.93	0.055
reliable	0.1	[0.04, 0.17]	3.07	0.002
Communication (RT)				
stable	0.35	[0.19, 0.51]	4.29	<0.001
unreliable	−0.04	[-0.23, 0.16]	−0.35	0.726
reliable	−0.50	[-0.69, −0.30]	−4.96	<0.001
Communication (Hits)				
stable	−0.11	[-0.14, −0.07]	−6.13	<0.001
unreliable	0.02	[-0.02, 0.06]	0.83	0.406
reliable	0.06	[0.02, 0.10]	2.89	0.004
Resource Management Task				
stable	0.51	[0.35, 0.66]	6.31	<0.001
unreliable	−0.19	[-0.38, 5.49e^{-03}]	−1.91	0.057
reliable	−0.35	[-0.54, −0.16]	−3.56	<0.001
Total performance				
stable	−0.49	[-0.72, −0.26]	−4.20	<0.001
unreliable	0.13	[-0.15, 0.41]	0.94	0.347
reliable	2.27	[1.99, 2.55]	15.97	<0.001

Note. Only estimates of the best fitting models are displayed. The RMSE for the take-over is calculated for a 2 s interval after transition from automatic to manual mode in the unreliable mode. Values have been z-transformed before analysis. See text for calculation of the total performance. b = beta, CI = 95% Confidence Interval, t = t-value, p = p-value.

Table 2
Results of the best-fitting linear-mixed models explaining the number of fixations.

Task	b	CI	t	p
Tracking Task				
Unreliable	−0.15	[-0.36, 0.06]	−1.41	0.159
Reliable	−1.68	[-1.89, −1.47]	−15.67	<0.001
Stable	0.03	[-0.19, 0.24]	0.24	0.808
Unreliable × Stable	0.18	[-0.12, 0.48]	1.19	0.235
Reliable × Stable	0.52	[0.22, 0.82]	3.42	<0.001
System Monitoring Task				
Stable	−0.48	[-0.59, −0.38]	−9.36	<0.001
Unreliable	0.13	[5.66e^{-03}, 0.25]	2.06	0.040
Reliable	0.50	[0.38, 0.62]	7.99	<0.001
Communication Task				
Stable	−0.22	[-0.28, −0.15]	−6.98	<0.001
Unreliable	0.06	[-0.01, 0.14]	1.68	0.094
Reliable	0.17	[0.10, 0.24]	4.54	<0.001
Resource Management Task				
Stable	−0.64	[-0.78, −0.49]	−8.72	<0.001
Unreliable	0.14	[-0.04, 0.31]	1.52	0.129
Reliable	1.00	[0.83, 1.18]	11.20	<0.001

Note. Estimates are based on z-transformed values extracted from the best fitting models according to a chi-square comparison. Detailed results of the model comparisons can be found in the supplementary material. b = beta, CI = 95% Confidence Interval, t = t-value, p = p-value.

Table 8). The post-hoc comparisons for model 3 show that participants made fewer task switches in the reliable automation condition compared to the manual and unreliable conditions, regardless of control mode (see Table 9). No significant difference in task switches was found between the manual and unreliable conditions in either control mode. Additionally, participants made more task switches in the flexible control condition than in the stable condition for both the manual and unreliable automation conditions. This effect was not observed in the reliable automation condition, where task switch frequency remained consistent across control modes (see Table 10).

Coefficient K – Model 2, which includes control mode and automation reliability, best explains the coefficient K. The stable condition increased the coefficient K ($M_{flexible}$ = −0.08; $SD_{flexible}$ = 0.19; M_{stable} = 0.29; SD_{stable} = 0.82). Unreliable automation decreased the coefficient K and reliable automation decreased it even more (M_{manual} = 0.27; SD_{manual} = 0.81; $M_{unreliable}$ = 0.08; $SD_{unreliable}$ = 0.48; $M_{reliable}$ = −0.04; $SD_{reliable}$ =

Table 3
Post Hoc Contrasts of Model 3: Interaction Effect of cognitive control on number of fixations on the tracking task.

Condition	Contrast	Estimate	SE	t-ratio	p
flexible	manual - unreliable	0.15	0.11	1.41	0.340
flexible	manual - reliable	1.68	0.11	15.67	<0.001
flexible	unreliable - reliable	1.53	0.11	14.25	<0.001
stable	manual - unreliable	−0.03	0.11	−0.27	0.959
stable	manual - reliable	1.16	0.11	10.83	<0.001
stable	unreliable - reliable	1.19	0.11	11.10	<0.001

Note. SE = standard error, p = p-value.

Table 4
Post Hoc Contrasts of Model 3: Interaction Effect of automation reliability on number of fixations on the tracking task.

Automation level	Contrast	Estimate	SE	t-ratio	p
manual	flexible - stable	−0.03	0.11	−0.24	0.808
unreliable	flexible - stable	−0.21	0.11	−1.93	0.055
reliable	flexible - stable	−0.55	0.11	−5.08	<0.001

Note. SE = standard error, p = p-value.

Table 5
Results of the best-fitting linear-mixed models explaining the fixation duration.

Task	b	CI	t	p
Tracking Task				
Unreliable	−0.11	[-0.61, 0.39]	−0.44	0.662
Reliable	−0.25	[-0.75, 0.24]	−1.00	0.317
Stable	1.70	[1.20, 2.20]	6.74	<0.001
Unreliable × Stable	−0.71	[-1.42, −0.01]	−2.00	0.047
Reliable × Stable	−0.91	[-1.62, −0.21]	−2.56	0.011
System Monitoring Task				
Unreliable	−2.75e^{-03}	[-0.06, 0.05]	−0.10	0.921
Reliable	0.13	[0.08, 0.19]	4.82	<0.001
Stable	−0.11	[-0.15, −0.06]	−4.74	<0.001
Communication Task				
Unreliable	0.12	[0.04, 0.20]	2.90	0.004
Reliable	0.53	[0.45, 0.61]	13.18	<0.001
Stable	−0.20	[-0.27, −0.14]	−6.16	<0.001
Resource Management Task				
Unreliable	3.69e^{-03}	[-0.02, 0.03]	0.28	0.781
Reliable	0.08	[0.06, 0.11]	6.18	<0.001
Stable	−0.04	[-0.06, −0.02]	−3.70	<0.001

Note. Estimates are based on z-transformed values extracted from the best fitting models according to a chi-square comparison. Detailed results of the model comparisons can be found in the supplementary material. b = beta, CI = 95% Confidence Interval, t = t-value, p = p-value.

Table 6
Post Hoc Contrasts of Model 3: Interaction Effect of cognitive control on fixation duration on the tracking task.

Condition	Contrast	Estimate	SE	t-ratio	p
flexible	manual - unreliable	15.80	36.1	0.44	0.900
flexible	manual - reliable	36.20	36.1	1.00	0.576
flexible	unreliable - reliable	20.40	36.1	0.57	0.839
stable	manual - unreliable	117.80	36.1	3.26	0.004
stable	manual - reliable	166.80	36.1	4.62	<0.001
stable	unreliable - reliable	49.00	36.1	1.36	0.365

Note. SE = standard error, p = p-value.

Table 7
Post Hoc Contrasts of Model 3: Interaction Effect of automation reliability on number of fixations on the tracking task.

Automation level	Contrast	Estimate	SE	t-ratio	p
manual	flexible - stable	−243.00	36.1	−6.74	<0.001
unreliable	flexible - stable	−141.00	36.1	−3.92	<0.001
reliable	flexible - stable	−113.00	36.1	−3.12	0.002

Note. SE = standard error, p = p-value.

Table 8
Results of the best-fitting linear-mixed models explaining AOI unspecific metrics.

Predictor	b	CI	t	p
Task Switches				
unreliable	−0.19	[-0.46, 0.07]	−1.46	0.146
reliable	−1.12	[-1.38, −0.86]	−8.40	<0.001
stable	−0.78	[-1.04, −0.52]	−5.86	<0.001
unreliable*stable	0.17	[-0.20, 0.55]	0.93	0.354
reliable*stable	0.64	[0.27, 1.01]	3.40	<0.001
Coefficient K				
stable	0.24	[0.18,0.30]	7.64	<0.001
unreliable	−0.10	[-0.18, −0.03]	−2.64	0.009
reliable	−0.23	[-0.31, −0.15]	−5.97	<0.001
Stationary Entropy				
stable	−0.06	[-0.09, −0.04]	−4.88	<0.001
Transition Entropy				
stable	−0.05	[-0.07, −0.03]	−4.53	<0.001
unreliable	$2.08e^{-03}$	[-0.02, 0.03]	0.16	0.869
reliable	−0.06	[-0.09, −0.04]	−4.92	<0.001

Note. Task switches have been z-standardized before analysis. Parameter estimates are extracted from the best fitting model. b = beta, CI = 95% Confidence Interval, t = t-value, p = p-value.

Table 9
Post Hoc Contrasts of Model 3: Interaction Effect of cognitive control on number of task switches.

Condition	Contrast	Estimate	SE	t-ratio	p
flexible	manual - unreliable	0.19	0.13	1.46	0.313
flexible	manual - reliable	1.12	0.13	8.40	<0.001
flexible	unreliable - reliable	0.93	0.13	6.95	<0.001
stable	manual - unreliable	0.02	0.13	0.15	0.988
stable	manual - reliable	0.48	0.13	3.59	0.001
stable	unreliable - reliable	0.46	0.13	3.45	0.002

Note. SE = standard error, p = p-value.

Table 10
Post Hoc Contrasts of Model 3: Interaction Effect of automation reliability on number of task switches.

Automation level	Contrast	Estimate	SE	t-ratio	p
manual	flexible - stable	0.78	0.13	5.86	<0.001
unreliable	flexible - stable	0.61	0.13	4.54	<0.001
reliable	flexible - stable	0.14	0.13	1.05	0.297

Note. SE = standard error, p = p-value.

0.48; see Table 8).

Stationary entropy – Model 1, including control mode, best explains stationary entropy. The stable condition had a negative effect on stationary entropy ($M_{flexible} = 0.92$; $SD_{flexible} = 0.07$; $M_{stable} = 0.87$; $SD_{stable} = 0.17$; see Table 8).

Transition entropy – Model 2, which includes control mode and automation reliability, best explained transition entropy. The stable condition decreased transition entropy ($M_{flexible} = 0.65$; $SD_{flexible} = 0.08$; $M_{stable} = 0.61$; $SD_{stable} = 0.15$). Additionally, reliable automation, but not unreliable automation, decreased transition entropy ($M_{manual} = 0.65$; $SD_{manual} = 0.13$; $M_{unreliable} = 0.65$; $SD_{unreliable} = 0.11$; $M_{reliable} = 0.59$; $SD_{reliable} = 0.11$; see Table 8).

3.3. Mental workload

Mental workload. Mental workload was best explained by Model 2, which includes control mode and automation reliability. The stable condition reduced mental workload ($M_{flexible} = 11.84$; $SD_{flexible} = 2.24$; $M_{stable} = 11.28$; $SD_{stable} = 2.74$), as well as the reliable automation condition ($M_{manual} = 11.86$; $SD_{manual} = 2.58$; $M_{unreliable} = 11.86$; $SD_{unreliable} = 2.39$; $M_{reliable} = 10.96$; $SD_{reliable} = 2.72$). No effect was found for unreliable automation (see Table 11).

3.4. Trust in automation

Trust in Automation. Trust in automation was measured using two scales: trust and mistrust in automation. Model 1, which includes automation reliability, best explains trust ($M_{unreliable} = 3.28$; $SD_{unreliable} = 1.33$; $M_{reliable} = 5.35$; $SD_{reliable} = 1.20$) and mistrust ($M_{unreliable} = 3.99$; $SD_{unreliable} = 1.23$; $M_{reliable} = 2.44$; $SD_{reliable} = 1.10$) in automation. Hereby, reliable automation positively influenced trust and reduced mistrust (see Table 11).

4. Discussion

This study explored the interaction between the cognitive control mode (flexible, stable; manipulated through task prioritization) and automation reliability (manual, unreliable, reliable) in the tracking task, examining their effects on overall performance, eye-tracking metrics, and mental workload. Considering the total MATB performance, our findings indicate that participants performed best with a flexible control mode and reliable automation, and worst with a stable control mode and no automation. No interaction effect was found on a performance level between the control mode and the automation reliability. Instead, performance on the tracking task was in general superior in the stable condition compared to the flexible one. Considering the initial 2 s after transitioning from automatic to manual mode in the unreliable condition, participants also performed better in the stable condition compared to the flexible one. This improvement is likely due to the benefits of cognitive stability, such as clearer goal focus and reduced susceptibility to distractions.

Contrary to previous studies (Metzger and Parasuraman, 2017; Rovira et al., 2002), our results did not show that unreliable automation leads to worse performance than manual control in the primary task. This discrepancy may stem from the continuous nature of the tracking task, which inherently demands significant attention from participants, who were aware that manual readjustment might be required unexpectedly at any time in both conditions.

For the system monitoring, communication, and resource management tasks, performance improved under flexible control relative to stable control, consistent with previous research (Stasch and Mack,

Table 11
Results of the best-fitting linear-mixed models explaining mental workload and trust in automation.

Outcome	b	CI	t	p
Mental Workload				
stable	−0.56	[-0.93, −0.20]	−3.07	0.002
unreliable	$-6.94e^{-03}$	[-0.45, 0.44]	−0.03	0.975
reliable	−0.90	[-1.34, −0.46]	−4.00	<0.001
Trust in automation				
reliable	2.07	[1.79, 2.35]	14.63	<0.001
Mistrust in automation				
reliable	−1.55	[-1.82, −1.28]	−11.37	<0.001

Note. For the calculation of trust and mistrust in automation, the manual condition has been excluded from the analysis. b = beta, CI = 95% Confidence Interval, t = t-value, p = p-value.

2023a; Stasch et al., 2024). This improvement likely results from a flexible task switching ability associated with cognitive flexibility, since the task switching costs were lower in the flexible condition compared to the stable condition. Evidence for that notion is provided by Siqi-Liu and Egner (2020), Braem and Egner (2018), as well as Dreisbach and Fröber (2018), indicating that greater cognitive flexibility is associated with smaller switching costs.

Reliable automation enhanced secondary task performance, whereas manual and unreliable automation conditions produced similar performance levels, aligning with findings from Chavaillaz et al. (2016), showing that low-reliability impairs secondary task performance. A possible reason for the improvement with reliable automation, but not between the manual and unreliable conditions, may be that it reduced the cognitive load associated with the primary tracking task. According to Wickens' 4-D multiple resource model (Wickens, 2002), the ability to perform multiple tasks simultaneously, depends not only on the amount of mental resources each task requires but also on the type of resources. Tasks that require different types of resources can be performed more efficiently together than tasks that rely on the same resource. In the present study, reliable automation of the tracking task likely reduced the mental load on visual-spatial processing and spatial-manual responses, freeing up these resources to be used for secondary tasks (e.g., system monitoring, resource management), which also required these resources. As a result, reliable automation led to improved secondary task performance. In contrast, unreliable automation continued to demand attention for the tracking task, even when no input was required (50%), occupying the same cognitive resources as the manual condition and leading to no improvement in secondary task performance.

Visual attention analysis via eye-tracking shows that participant in the stable control mode made longer, but not more frequent, fixations on the tracking task, replicating the result of Stasch and Mack (2023a). Furthermore, the stable control mode led to fewer, but longer, fixations on the system monitoring, communication, and resource management task, expect for the fixation duration on the system monitoring task, which is also in line with Stasch and Mack (2023a). The fact that the stable control mode demonstrated an effect on the fixation duration in this study might be attributed to the larger sample size offering greater test sensitivity. The task switching data indicates that operating under a stable control mode significantly reduced the frequency of task switches. These results support the notion that the task-switching ability is diminished, but accompanied by enhanced goal-shielding, under a stable control mode (Dreisbach and Fröber, 2019). Furthermore, low entropy values, indicating a more predictable scanning sequence, and a coefficient K greater than 0, which suggests a focus on focal processing (Krejtz et al., 2016) were found in the stable condition compared to the flexible condition.

Participants fixated the tracking task less frequently with reliable automation, which aligns with Sato et al. (2023), who found that higher trust in automation correlates with reduced fixation frequency. Reliable automation of the tracking task increased both fixation frequency and duration on the system-monitoring, communication and resource management task, suggesting it indeed released visual-spatial processing resources from the tracking task, which in turn could be utilized for more detailed task processing on the other tasks. Notably, also unreliable automation increased the number of fixations, but not the fixation duration, on the system monitoring task, while no effect on of unreliable automation was detected on the performance level. Together with the finding that the coefficient K was lower than 0 (indicating ambient visual processing) in the unreliable condition, this result might be explained by the fact that the stimuli in the system monitoring task, namely the moving scales and the color-change in the F5 and F6 buttons acted as bottom-up cues attracting visual attention. Since the fixation duration was not enhanced with unreliable automation, this result can simply be a result of distraction, since the scales were also moving during the entire trial.

Additionally, reliable automation further reduced task switching compared to manual control and unreliable automation. This reduction is likely because participants did not need to frequently check the status of the tracking task. However, when the tracking task was reliably automated and participants were in a stable control mode, task switching increased. This increase may be due to participants prioritizing the tracking task and frequently checking it with brief fixations, as intended by the experimental design. Generally, the reliable automation condition was associated with ambient visual processing (Coefficient K < 0). Additionally, reliable automation decreased mental workload, as reflected in lower NASA-TLX scores, which may have allowed participants to allocate more resources to monitoring secondary tasks and process these tasks with ambient visual attention, which would be explainable with the previously mentioned 4-D multiple resource model of Wickens (2002).

The present eye-tracking metrics provide insights not only into their relationship with cognitive control modes but also offer interpretations in terms of Situational Awareness (SA)—the ability to perceive elements in the environment over time and space (Level 1 SA), understand their significance (Level 2 SA), and anticipate their future status (Level 3 SA; Endsley, 1995). Van de Merwe et al. (2012) demonstrated that eye-tracking metrics can serve as indicators of different SA levels in the cockpit: higher fixation rates and dwell times correlate with information acquisition (Level 1 SA), while lower entropy values suggest an ability to efficiently guide information acquisition activities (Level 3 SA). Additionally, Lounis et al. (2020) found that high Coefficient K values may reflect over-focalization, potentially diminishing SA (Sarter and Woods, 1991).

In the stable condition, decreases in secondary task performance could be related to a reduction in SA. This is further supported by decreases in fixation count and duration in the stable condition, which, according to van de Merwe et al. (2012), might indicate low Level 1 SA. However, in the present study, entropy values were also lower in the stable condition, suggesting high Level 3 SA—a result that does not align with observed performance decrements. Generally, higher SA levels are associated with improved performance (Endsley, 1990; Endsley and Kiris, 1995; Ma and Kaber, 2007). However, a recent meta-analysis by Bakdash et al. (2022) suggests that the predictive validity of SA for performance is relatively weak, with significant variation among effect sizes, which could explain this discrepancy.

Conversely, a reduction in mental workload was not observed in the unreliable automation condition. In this case, participants likely anticipated potential failures and maintained an active task-set for the tracking task, preventing a reduction in workload. Finally, the results of the trust questionnaire indicate that the reliable condition was positively associated with trust in automation and negatively associated with mistrust, confirming that the reliability manipulation functioned as intended.

4.1. Implications for the interface design

The current findings have several implications for the interface design of human-machine systems. Firstly, the observed interaction effects between control mode and automation level, as reflected in various eye-tracking metrics, underscore the importance considering a tasks automation level for the accurate diagnosis of the cognitive control mode. Failing to correctly identify whether a system's task is operating in a manual, unreliable, or reliable mode can lead to misdiagnosis of the control mode, potentially resulting in inappropriate adaptations within the adaptive assistance system and a degraded human-machine performance.

Assuming an accurate diagnosis of the control mode and automation reliability level, how might an adaptive assistance system further support an operator? The results indicate that adapting to control mode and automation level is not a one-size-fits-all solution. This adaptation depends on whether the system is functioning reliably or unreliably. Four plausible scenarios can be considered, extending the modifications proposed by Stasch and Mack (2024): the system's automated subtask may be functioning reliably or unreliably, while the operator's control mode may be stable or flexible.

4.1.1. Reliable automation & stable control mode

In this scenario, overall task performance is high and mental workload rather low. Despite the good performance and low mental workload, task performance, particularly in subtasks, could be further improved if the operator were in a flexible control mode. An adaptive assistance system could address this by increasing the intensity of warning signals of other subtasks, potentially redirecting visual attention and allowing for deeper cognitive processing of those.

4.1.2. Reliable automation & flexible control mode

This combination yields the best performance among all scenarios, but it is associated with higher workload levels compared to when the system has reliable automation and the operator is in a stable control mode. Therefore, it is crucial to closely monitor the operator's workload, potentially through psycho-physiological measurements (Brookhuis and de Waard, 2010; Schwarz et al., 2014). In the event of workload overload, the system could be adapted to increase automation, potentially achieving full reliable automation of subtasks to reduce overall task load.

4.1.3. Unreliable automation & stable control mode

This scenario is characterized by improved take-over behavior compared to a flexible control mode. However, other subtasks may suffer in performance due to the operator's focus on responding to unpredictable system failures. To support this control mode during severe disruptions of the primary task—reflecting its urgency and importance in that instance—the adaptive assistance system could simplify the interface to minimize distractions from non-relevant details. In a team context, other subtasks might also be delegated to team members in a flexible control mode to prevent mental overload situations.

4.1.4. Unreliable automation & flexible control mode

Manual take-over behavior would benefit from a stable control mode, so adaptation should enhance the saliency of the primary task to improve take-over behavior. Results showed no significant performance differences between manual and unreliable operating modes, but switching entirely to a manual operation mode in that instance needs to be evaluated in future studies.

4.2. Limitations

The study effectively validates the use of gamification for manipulating cognitive control through task prioritization (Stasch and Mack, 2023b), demonstrates how the stability-flexibility dilemma impacts performance, and provides valuable insights into attention distribution via eye-tracking. However, several limitations should be noted. First, the MATB (Multi-Attribute Task Battery) simulates tasks that, while relevant, are abstracted from real-world multitasking scenarios faced by pilots. Additionally, the study predominantly involved local university students, which limits the generalizability of the findings to an expert sample. Moreover, the stability-flexibility dilemma was experimentally manipulated rather than observed in natural settings. Also the fact that the current study found no difference between a manual operation mode and unreliable automation on a performance level should be considered with caution, since this finding contradicts findings from previous research using a different experimental task. Therefore, future research should aim to replicate these results in high-fidelity multitasking environments with naturally varying task prioritization. Furthermore, the study used only two extremes of automation reliability (50% and 100%). Further research should examine how more nuanced levels of automation reliability, as explored by Avril et al. (2021), might affect performance in relation to the stability-flexibility dilemma. Lastly, the efficacy of the proposed recommendations for how an adaptive assistance system could support operators in a stable or flexible control state requires further evaluation in future studies.

5. Conclusion

Promoting a flexible control mode is advisable when the automated system is functioning reliably. However, when the system is unreliable, a stable control mode provides better primary task performance, though it may impact secondary task performance. Accurate user state diagnosis in the perceive stage, considering automation reliability, is crucial for effective adaptive assistance. Future research needs to explore additional interaction effects, such as varying levels of automation reliability and their impact on task performance.

CRediT authorship contribution statement

Sophie-Marie Stasch: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wolfgang Mack:** Resources, Methodology, Funding acquisition.

Funding

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr (project MissionLab). Dtec.bw is funded by the European Union – NextGenerationEU.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank Dr. Yannik Hilla for his valuable comments on the design of the experiment.

References

Avril, E., Valéry, B., Navarro, J., Wioland, L., Cegarra, J., 2021. Effect of imperfect information and action automation on attentional allocation. Int. J. Hum. Comput. Interact. 37 (11), 1063–1073.

Bainbridge, L., 1983. Ironies of automation. Automatica 19 (6), 775–779.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Braem, S., Egner, T., 2018. Getting a grip on cognitive flexibility. Curr. Dir. Psychol. Sci. 27 (6), 470–476. https://doi.org/10.1177/0963721418787475.

Brookhuis, K.A., De Waard, D., 2010. Monitoring drivers' mental workload in driving simulators using physiological measures. Accid. Anal. Prev. 42 (3), 898–903.

Cegarra, J., Valéry, B., Avril, E., Calmettes, C., Navarro, J., 2020. OpenMATB: a Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability. Behav. Res. Methods 52 (5), 1980–1990. https://doi.org/10.3758/s13428-020-01364-w.

Chavaillaz, A., Wastell, D., Sauer, J., 2016. System reliability, performance and trust in adaptable automation. Appl. Ergon. 52, 333–342. https://doi.org/10.1016/j.apergo.2015.07.012.

Chen, J.Y., Barnes, M.J., Kenny, C., 2011. Effects of unreliable automation and individual differences on supervisory control of multiple ground robots. In: Proceedings of the 6th International Conference on Human-Robot Interaction, pp. 371–378. https://doi.org/10.1145/1957656.1957793.

De Tjerk, E.G., Henryk, F.A., Neerincx, M.A., 2010. Adaptive automation based on an object-oriented task model: implementation and evaluation in a realistic c2 environment. J. Cogn. Eng. Decis. Mak. 4 (2), 152–182.

Di Flumeri, G., De Crescenzio, F., Berberian, B., Ohneiser, O., Kramer, J., Aricò, P., et al., 2019. Brain–computer interface-based adaptive automation to prevent out-of-the-loop phenomenon in air traffic controllers dealing with highly automated systems. Frontiers in human neuroscience 13, 296. https://doi.org/10.3389/fnhum.2019.00296.

Dixon, S.R., Wickens, C.D., 2004. Automation reliability in unmanned aerial vehicle flight control. In: Proceedings of Human Performance, Situational Awareness and Automation Conference, pp. 205–209.

Dreisbach, G., Fröber, K., 2019. On how to be flexible (or not): modulation of the stability-flexibility balance. Curr. Dir. Psychol. Sci. 28 (1), 3–9.

Egner, T., 2023. Principles of cognitive control over task focus and task switching. Nature Reviews Psychology 2 (11), 702–714. https://doi.org/10.1038/s44159-023-00234-4.

Endsley, M.R., 1990. Predictive utility of an objective measure of situation awareness. Proceedings of the Human Factors Society Annual Meeting 34 (1), 41–45. https://doi.org/10.1177/154193129003400110.

Endsley, M.R., Kiris, E.O., 1995. The out-of-the-loop performance problem and level of control in automation. Hum. Factors 37 (2), 381–394.

Eriksson, A., Stanton, N.A., 2017. Takeover time in highly automated vehicles: noncritical transitions to and from manual control. Hum. Factors 59 (4), 689–705. https://doi.org/10.1177/0018720816685832.

Feigh, K.M., Dorneich, M.C., Hayes, C.C., 2012. Toward a characterization of adaptive systems: a framework for researchers and system designers. Hum. Factors 54 (6), 1008–1024. https://doi.org/10.1177/0018720812443983.

Fischer, R., Plessow, F., 2015. Efficient multitasking: parallel versus serial processing of multiple tasks. Front. Psychol. 6, 1366.

Fröber, K., Dreisbach, G., 2017. Keep flexible – keep switching! The influence of forced task switching on voluntary task switching. Cognition 162, 48–53. https://doi.org/10.1016/j.cognition.2017.01.024.

Geddert, R., Egner, T., 2022. No need to choose: independent regulation of cognitive stability and flexibility challenges the stability-flexibility trade-off. J. Exp. Psychol. Gen. 151 (12), 3009.

Goschke, T., Bolte, A., 2014. Emotional modulation of control dilemmas: the role of positive affect, reward, and dopamine in cognitive stability and flexibility. Neuropsychologia 62, 403–423.

Gratton, G., Cooper, P., Fabiani, M., Carter, C.S., Karayanidis, F., 2018. Dynamics of cognitive control: theoretical bases, paradigms, and a view for the future. Psychophysiology 55 (6), e13016. https://doi.org/10.1111/psyp.13016.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology (Vol. 52, Pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9.

Hommel, B., 2015. Between persistence and flexibility: the Yin and Yang of action control. In: Advances in Motivation Science (Vol. 2, Pp. 33–67). Elsevier.

Hoover, A.L., 2008. Prioritizing tasks in the cockpit: a review of cognitive processing models, methods of dealing with cognitive limitations, and training strategies. The Collegiate Aviation Review International 26 (1), 36–45. https://doi.org/10.22488/OKSTATE.18.100371.

Jian, J.Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. Int. J. Cognit. Ergon. 4 (1), 53–71.

Koch, I., Poljac, E., Müller, H., Kiesel, A., 2018. Cognitive structure, flexibility, and plasticity in human multitasking—an integrative review of dual-task and task-switching research. Psychol. Bull. 144 (6), 557.

Krejtz, K., Szmidt, T., Duchowski, A.T., Krejtz, I., 2014. Entropy-based statistical analysis of eye movement transitions. Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14) 159–166. https://doi.org/10.1145/2578153.2578176.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., Kopacz, A., 2016. Discerning ambient/focal attention with coefficient K. Trans. Appl. Percept. 13 (3), 1–20. https://doi.org/10.1145/2896452.

Lenth, R.V., 2024. emmeans: estimated marginal means, aka least-squares means (R package version 1.10.5). https://CRAN.R-project.org/package=emmeans.

Lounis, C.A., Hassoumi, A., Lefrancois, O., Peysakhovich, V., Causse, M., 2020. Detecting ambient/focal visual attention in professional airline pilots with a modified Coefficient K: a full flight simulator study. In: In Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Adjunct. Association for Computing Machinery, pp. 1–6. https://doi.org/10.1145/3379157.3391412. Article 25.

Ma, R., Kaber, D.B., 2007. Situation awareness and driving performance in a simulated navigation task. Ergonomics 50 (8), 1351–1364. https://doi.org/10.1080/00140130701318913.

Mackie, M.A., Van Dam, N.T., Fan, J., 2013. Cognitive control and attentional functions. Brain Cognit. 82 (3), 301–312. https://doi.org/10.1016/j.bandc.2013.05.004.

Metzger, U., Parasuraman, R., 2017. Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. In: Decision Making in Aviation. Routledge, pp. 345–360.

Musslick, S., Jang, S.J., Shvartsman, M., Shenhav, A., Cohen, J.D., 2018. Constraints associated with cognitive control and the stability-flexibility dilemma. In: Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018. The Cognitive Science Society, pp. 804–809. Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018).

Nahlik, P., Daubenmire, P.L., 2022. Adapting gaze-transition entropy analysis to compare participants' problem solving approaches for chemistry word problems. Chem. Educ. Res. Pract. 23 (3), 714–724. https://doi.org/10.1039/d2rp00066k.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Syst. Hum. 30 (3), 286–297. https://doi.org/10.1109/3468.844354.

Rovira, E., Parasuraman, R., 2007. Effects of imperfect automation support on air traffic controller (ATCO) performance, mental workload, and attention allocation: miss vs. False alarm prone automation. In: 2007 International Symposium on Aviation Psychology, p. 611.

Rovira, E., McGarry, K., Parasuraman, R., 2002. Effects of unreliable automation on decision making in command and control. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 46, No. 3, Pp. 428-432). SAGE Publications.

Salvucci, D.D., Taatgen, N.A., 2008. Threaded cognition: an integrated theory of concurrent multitasking. Psychol. Rev. 115 (1), 101.

Santiago-Espada, Y., Myer, R.R., Latorella, K.A., Comstock, J.R., 2011. *The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide* (NASA/TM-2011-217164). National Aeronautics and Space Administration, Langley Research Center.

Sarter, N.B., Woods, D.D., 1991. Situation awareness: a critical but ill-defined phenomenon. Int. J. Aviat. Psychol. 1 (1), 45–57.

Sato, T., Inman, J., Politowicz, M.S., Chancey, E.T., Yamani, Y., 2023. A meta-analytic approach to investigating the relationship between human-automation trust and attention allocation. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 67 (1), 959–964. https://doi.org/10.1177/21695067231194333.

Schwarz, J., Fuchs, S., Flemisch, F., 2014. Towards a more holistic view on user state assessment in adaptive human-computer interaction. 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1228–1234. https://doi.org/10.1109/SMC.2014.6974082.

Siqi-Liu, A., Egner, T., 2020. Contextual adaptation of cognitive flexibility is driven by task- and item-level learning. Cognit. Affect Behav. Neurosci. 20 (4), 757–782. https://doi.org/10.3758/s13415-020-00801-9.

Stasch, S.M., Mack, W., 2023a. Why the stability-flexibility-dilemma should be taken into consideration when studying pilots' multitasking behaviour. In: Praetorius, G., Sellberg, C., Patriarca, R. (Eds.), Human Factors in Transportation. AHFE (2023) International Conference, AHFE Open Access, vol. 95. AHFE International, USA. https://doi.org/10.54941/ahfe1003846.

Stasch, S.M., Mack, W., 2023b. A new experimental method to investigate multitasking strategies in flight environments via the use of gamification. In: de Waard, D., Hagemann, V., Onnasch, L., Toffetti, A., Coelho, D., Botzer, A., et al. (Eds.), Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2023, Annual Conference. Human Factors and Ergonomics Society Europe Chapter, pp. 23–33.

Stasch, S.M., Mack, W., 2024. Diagnosing cognitive control with eye-tracking metrics in a multitasking environment. In: Harris, D., Li, W.-C. (Eds.), Engineering Psychology and Cognitive Ergonomics. Springer, Nature Switzerland, pp. 89–102.

Stasch, S.M., Mack, W., Hilla, Y., 2024. Always in (partner) action? Working in teams may improve simulated flight performance – but only in the apt cognitive control state. https://osf.io/preprints/psyarxiv/2kajt.

Strand, N., Nilsson, J., Karlsson, I.M., Nilsson, L., 2014. Semi-automated versus highly automated driving in critical situations caused by automation failures. Transport. Res. F Traffic Psychol. Behav. 27 (Part B), 218–228. https://doi.org/10.1016/j.trf.2014.04.005.

Trapsilawati, F., Qu, X., Wickens, C.D., Chen, C.H., 2015. Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control. Ergonomics 58 (6), 897–908. https://doi.org/10.1080/00140139.2014.997301.

van de Merwe, K., van Dijk, H., Zon, R., 2012. Eye movements as an indicator of situation awareness in a flight simulator experiment. Int. J. Aviat. Psychol. 22 (1), 78–95. https://doi.org/10.1080/10508414.2012.635129.

Vogelpohl, T., Kühn, M., Hummel, T., Gehlert, T., Vollrath, M., 2018. Transitioning to manual driving requires additional time after automation deactivation. Transport. Res. F Traffic Psychol. Behav. 55, 464–482.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theor. Issues Ergon. Sci. 3 (2), 159–177. https://doi.org/10.1080/14639220210123806.

Xu, X., Wickens, C.D., Rantanen, E.M., 2007. Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. Ergonomics 50 (1), 112–130. https://doi.org/10.1080/00140130601002658.