Assessing Human-AI Teaming in Single-Pilot Operations: Effects of AI Support on Cognitive Workload and Performance

Michelle Fini (michelle.fini@ils.uni-stuttgart.de)[1], **Zamira Daw (zamira.daw@ils.uni-stuttgart.de)**[1] **&**
Maria Wirzberger (maria.wirzberger@iris.uni-stuttgart.de)[2]
[1]University of Stuttgart, Institute of Aircraft Systems, Stuttgart, Germany
[2]University of Stuttgart, Interchange Forum for Reflecting on Intelligent Systems, Stuttgart, Germany

Abstract

As AI-driven concepts like Single-Pilot Operations challenge traditional two-pilot operations, debates among manufacturers, pilot unions, and regulators highlight the need for empirical evidence on how AI uncertainty affects human cognition and performance. This research compares human-human and human-AI cockpit teams, focusing on workload and performance when the pilot monitoring role is assumed by AI. 34 pilots completed simulated takeoff scenarios under three conditions: a human co-pilot, a reliable AI co-pilot, and a faulty AI co-pilot. Results indicate that collaboration with reliable AI can sustain workload levels and team performance comparable to human crews, whereas faulty AI behavior corresponds to increased workload and degraded team performance. Further analyses suggest that performance degradation is linked to undetected AI errors and AI limitations. Although pilots were often able to compensate for AI shortcomings, team collaboration was adversely affected, underscoring the importance of AI reliability and transparency for safe human-AI teaming.

Keywords: Human-AI Teaming; Workload; Performance; Taskload; AI; Aviation; Cognition; Single-Pilot Operations

Introduction

In aviation, a well-established task-sharing concept divides responsibilities between the Pilot Flying (PF) and Pilot Monitoring (PM) to reduce failure probability and enhance efficiency through redundancy, error detection, and balanced workload. However, Single-Pilot Operations (SiPO) have become a hardly discussed topic in the past years. While manufacturers promote these concepts and pilot unions raise safety concerns (Puca & Guglieri, 2025), the European Union Aviation Safety Agency (EASA) has halted further pursuit due to the lack of demonstrated safety equivalence compared to two-pilot operations with current technologies (Zon & Roelen, 2025). This underscores the need for validated advanced automation and AI to support a sole operator for a safety level comparable to two-pilot operations. EASA distinguishes three AI levels by autonomy and adaptability under human oversight: assisting functions (Level 1), human-AI teaming (Level 2), and more autonomous systems (Level 3) (Soudain, 2024). Advances in AI could enable SiPO to achieve comparable safety levels to or even exceed those in conventional two-pilot operations by leveraging effective human-AI task-sharing (Korentsides et al., 2024; Ziakkas et al., 2025). But potential efficiency gains may also cover safety-critical inefficiencies. Due to the absence of the PM role in SiPO, the pilot's individual performance becomes highly critical for flight safety and introduces major challenges in workload management (Ziakkas et al., 2025). The strong link between performance, task demands and workload underscores the importance of balanced workload for safe flight operations. Performance is optimal when task demands match operator capacity as reflected in perceived workload, whereas both underload and overload degrade performance. When demands exceed human capacity, performance declines sharply, unless task demands are reduced or tasks are delegated (de Waard, 1996). While SiPO flight-simulator studies have examined workload and performance effects during regular and irregular events, they lack the involvement of human-AI teaming (HAIT) (Faulhaber, 2019; Miranda, 2025). Prior work explored AI assistants in SiPO to support situational awareness (SA) and decision-making, addressing key HAIT aspects such as workload and system transparency, while highlighting both potential and challenges (Duchevet et al., 2022; Minaskan et al., 2022). However, the effects of AI uncertain outcomes induced by machine-learning models on workload and team performance remain an unresolved research gap. This black-box behavior of AI amplifies new hazards in novel human-machine teaming concepts, resulting in "unknown unknowns" (Graydon et al., 2025). Preventing AI-induced accidents in HAIT requires frameworks and real-world validation that address uncertain AI behavior (Graydon et al., 2025; Kirwan, 2025; Korentsides et al., 2024). Consequently, as part of their research agenda for effective HAIT, the National Academies of Science, Engineering and Medicine (2022) call for models and metrics to better assess team effectiveness, thereby highlighting challenges in training via simulation-based evaluation, and the need for integrated human factors and AI engineering approaches.

The research reported in this paper contributes insights on HAIT in SiPO by systematically examining the impact of an AI serving as PM on team performance and workload as safety-critical factors, relative to conventional two-human-pilot crews in an aviation specific scenario. Thereby, it advances prior work by explicitly incorporating uncertain AI behavior that can induce irregular conditions, enabling a more realistic assessment of performance and workload effects in safety-critical operations.

Related Work

Survey-based studies have examined AI's role in mitigating SiPO risks. Ziakkas et al. (2025) highlight AI's potential to

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

improve decision-making and workload management, while identifying HAIT, trust and regulation as key challenges. Complementing this, Zinn et al. (2023) addressed high perceived risks of SiPO related to pilot overload, incapacitation, automation, and responsibility. Although an introduced AI- and remote-pilot-supported concept reduced perceived severity of some workload- and automation-related risks, willingness to operate SiPO remained low, underscoring the need for task-level analyses across flight phases that account for AI failures and actively involve pilots to support trust and operational feasibility.

Faulhaber (2019) found that pilot workload in SiPO was comparable to two-pilot operations under regular conditions but increased during turbulence and irregular events, accompanied by performance issues like missed checklist items. These findings are reinforced by Miranda (2025), which showed stable performance and manageable workload under regular conditions with aural automation support, whereas unanticipated aircraft system failures led to elevated workload, reduced SA, and performance degradation characterized by attentional tunneling. Together, these studies highlight that unexpected failures can induce cognitive overload, impairing even basic task execution. Minaskan et al. (2022) showed that proactive AI interfaces improved SA in SiPO, while workload remained largely comparable to two-pilot operations, with slight increases likely due to system novelty. Duchevet et al. (2022) demonstrated that AI support for aborted-landing decisions was generally appreciated but raised trust and explainability concerns when alerts were unclear or irrelevant.

Kirwan (2025) defined requirements for efficient HAIT in aviation, emphasizing workload management and human performance as central success factors. The framework highlights the importance of shared SA, clear human authority, and AI explainability, particularly under irregular conditions, and translates these principles into concrete evaluation criteria for different conditions. Kirwan (2025) calls for empirical studies to refine these requirements. Lorrig et al. (2025) have already developed an evaluation framework for assessing HAIT in SiPO that addresses AI uncertainty. The framework proved feasible for structured HAIT evaluation of workload and performance but revealed shortcomings in authority definition and the reference scenario.

Research Objectives and Hypotheses

This paper empirically evaluates HAIT in a SiPO-relevant, aviation-specific flight simulator study, extending prior survey-based and conceptual work through task-level analyses across varying AI behaviors (Ziakkas et al., 2025; Zinn et al., 2023). Building on Faulhaber (2019) and Miranda (2025), AI is introduced as a team-mate, with a focus on irregular conditions caused by faulty AI behavior rather than aircraft or environmental failures. While prior aviation HAIT studies primarily examined AI decision-support under regular conditions (Duchevet et al., 2022; Minaskan et al., 2022), this research investigates the effects of AI uncertainty on workload and team performance, addressing a critical gap in safety-relevant SiPO evaluation. The study responds to recommendations to refine HAIT requirements (Kirwan, 2025) by providing empirical evidence on workload and performance, extending and refining the framework of Lorrig et al. (2025) for a larger sample study, and following the National Academies of Science, Engineering and Medicine (2022) agenda through an integrated human factors and AI engineering approach. Building on prior research, the following hypotheses are formulated:

H1a Interacting with a non-faulty AI-PM results in a workload similar to interacting with a human-PM.

H1b Interacting with a faulty AI-PM results in a higher workload than interacting with a human-PM.

H2a Interacting with a non-faulty AI-PM results in a level of team performance similar to interacting with a human-PM.

H2b Interacting with a faulty AI-PM results in a lower team performance than interacting with a human-PM.

Method

Participants

In total, 34 pilots ($range_{Age}$ = 18-79, M_{Age} = 38.95, SD_{Age} = 14.09) participated; 97.06% identified themselves as male and 2.94% as female. Most were commercial or former commercial pilots (61.76%), followed by private pilot license holders (35.29%) and students with extensive knowledge of piloting in Microsoft Flight Simulator 2020 (MSFS2020) (2.94%). The majority were Germans (94.12%). The participants further held different experience levels in piloting A320-family aircraft measured in flight hours ($range_{hours}$ = 0-11000, M_{hours} = 1556.74, SD_{hours} = 2658.77). Participants received a monetary compensation of 45 Euro to account for their time. The data collection and analysis was approved by the ethics committee at the University of Stuttgart (approval number: 25-037).

Study Design

The study employed a within-subject design to examine the effects of HAIT in SiPO on workload and team performance (see Fig. 1). Using a refined evaluation framework based on Lorrig et al. (2025), pilot workload and team performance were assessed during the high-workload A320neo takeoff phase. Uncertainty in AI behavior was emulated through reliable and faulty AI-PM conditions, resulting in three experimental conditions: a two-pilot reference condition with a human PM, a SiPO condition with a non-faulty AI-PM, and a SiPO risk condition with a faulty AI-PM.

Scenarios

Cockpit tasks are normally divided between the PF, who actively controlles the aircraft, and the PM, who supports and monitors the operation, while command authority remains with the Pilot in Command (PiC). Participants completed five A320neo takeoff scenarios, acting as both PF and PiC, with full responsibility for flight safety and leading the multi-crew cooperation.

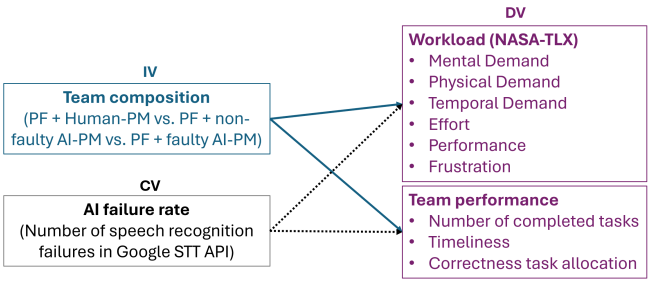

Figure 1: Study design diagram showing the interrelations between dependent variables (DVs): Workload and team performance, the independent variable (IV): Team composition and the covariate (CV): AI failure rate

Reference Scenario (PF + human-PM): Participants performed the standardized A320 takeoff procedure in a two-pilot configuration, following airline standard operating procedures (SOPs). After the takeoff call, the thrust was set, flight mode annunciations (FMAs) were monitored and called, and critical callouts (e.g., "Thrust Set", "100 knots", "V1", "Rotate", "Positive Climb") were executed by the PM and cross-checked by the PF. The PF conducted aircraft handling actions (e.g., rotation, thrust reduction), while the PM monitored system states and executed configuration changes (landing gear and flap retraction) upon command. The autopilot could be engaged at any time. The scenario ended once the aircraft reached clean configuration after flap retraction to position zero.

Non-faulty AI scenarios (PF + AI-PM1, PF + AI-PM2): The procedure remained fundamentally the same. However, the human co-pilot was substituted by the AI-PM system. So, the AI assumed all PM tasks specified by the SOPs. The participant in the role of PF communicated with the AI-PM via voice commands and remained responsible for monitoring AI responses and verifying correct task execution. In case of missing AI responses, the PF was required to repeat commands or manually perform the task.

Faulty AI scenarios (PF + faulty AI-PM1, PF + faulty AI-PM2): The procedure remained the same, but intended AI errors were included. The AI either announced “V1” prematurely and failed to retract the landing gear despite confirming the action (faulty AI-PM1), or acknowledged a command to set flaps to position 1 but incorrectly set them to position 3 (faulty AI-PM2). In both cases, participants were responsible for detecting and correcting the AI errors to ensure successful task completion.

Surrogate AI

A deterministic, scenario-driven surrogate AI was employed to manipulate simulator variables and execute event-based actions. For more immersive human-AI interaction, the system was extended with speech interaction using the Google Speech-to-Text API (Google Cloud, n.d.), enabling continuous voice command recognition. The AI-PM corresponds to EASA Level 2a automation due to its deterministic, pilot-directed behavior, while simulating selected Level 2b collaborative aspects through voice interaction and a shared operational goal.

Experimental Setup

A desktop-based MSFS2020 setup was used to collect participants’ behavioral responses. For immersive purposes and handling quality, a curved display and several flight control hardware was used as shown in Fig. 2. The setup further comprised headsets for voice interaction and an eye-tracking system was set up as PF monitor in front of the participant to cover relevant indicators for task completion (see Fig. 3).

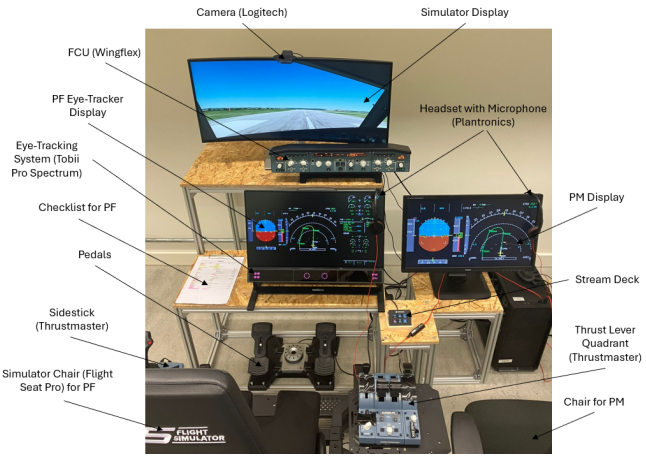

Figure 2: Hardware Setup used for recording participants’ task-related responses.

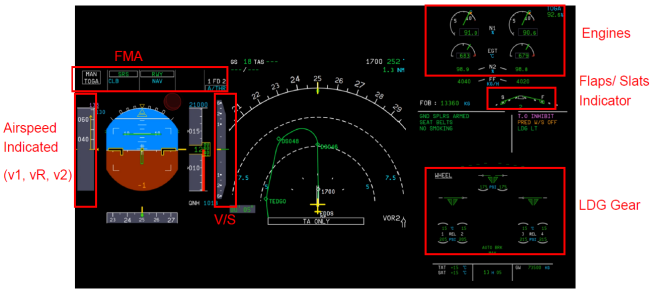

Figure 3: Eye-tracking display, including Airspeed Indicator, Flight Mode Annunciator (FMA), Vertical Speed Scale (V/S), Flap and Slat Indicator and Wheel State (LDG Gear).

Workload

Participants’ workload was assessed after each scenario using the NASA Task Load Index (NASA-TLX; (Hart & Staveland, 1988)) on a 20-point Likert scale ranging from very low to very high demand. In line with the advantages described by Casner and Gore (2010), the repeated measurement ensured that the perceived task experience was still present. Generally,

the NASA-TLX enables the assessment of workload in six different dimensions and seeks to account for biases arising from participants' beliefs about their own performance (Casner & Gore, 2010). It has been widely validated (Braarud, 2021; Flägel et al., 2019; Longo, 2018; Xiao et al., 2005), with reported scale reliabilities of $.75 < \alpha < .86$ across domains.

Performance

Performance was defined as the successful execution of tasks during the takeoff phase in an A320neo aircraft. The overall goal of a successful takeoff was systematically decomposed into 26 sub-tasks, defined by SOPs. These covered critical takeoff phases, including pre-takeoff procedures, runway roll, rotation and initial climb phase and configuration management. Beyond task completion and timeliness, performance included adherence to prescribed role allocation and effective communication between the PF and PM, as incorrect role assignment negatively affects monitoring and decision-making (Becker & Ayton, 2024; Behrend & Dehais, 2020). SOPs define task and authority allocation between PF and PM and are followed to ensure comparability between conventional two-pilot and AI-supported configurations. Communication via standardized commands was treated as a critical performance component, aligning with Kirwan (2025). In line with best practices for simulator-based training by Salas et al. (2009), two complementary performance metrics were defined:

- **Team Performance** considers the correct execution of tasks and compliance with the role allocation and responsibilities of each team member as defined by the SOPs.
- **Overall Task Performance** assesses whether all tasks were completed successfully, regardless of the correct role distribution. This score reflects stronger the individual performance of the participants as PiC.

Procedure

After being welcomed, all participants provided informed consent and their demographic data. Subsequently, they were briefed on their role as PF and PiC. To familiarize with the simulator and the required procedures, they further completed a short flight training session. After achieving two failure-free takeoffs, training proceeded with the AI-PM. Thereby, participants received information about the AI role and the available commands. Each participant performed at least one takeoff with the AI-PM during training. After eye-tracker calibration, participants flew the reference scenario followed by four AI-PM scenarios in randomized order. Workload was measured after each scenario using the NASA-TLX. Finally, they received their monetary compensation and were debriefed.

Scoring

Workload The NASA-TLX scoring procedure was modified by cutting the individual weighting of the six dimensions to avoid fatigue effects of the participants. This approach corresponds to the often applied RAW-TLX, in which dimension scores are averaged directly into an overall workload score, without the paired comparison step (Hart, 2006). The aggregated NASA-TLX total score showed good internal consistency (α=.84), comparable to prior research (Braarud, 2021; Flägel et al., 2019; Longo, 2018; Xiao et al., 2005).

Performance Performance was evaluated using an event-based scoring framework. It comprised standardized criteria for each of the 26 tasks defined, accounting for correct task execution, timing tolerances, procedural requirements, and selected human-observed behaviors. The final performance score was derived using binary scoring. Taken together, the approach for computing the overall task performance can be described mathematically as indicated by Lorrig et al. (2025):

$$PS(\delta, \tau) = \sum_{i=1}^{N} \delta_i \cdot \tau_i \tag{1}$$

$$\delta_i = \begin{cases} 1 & \text{if task } i \text{ was completed} \\ 0 & \text{otherwise} \end{cases} \qquad \tau_i = \begin{cases} 1 & \text{if } t_i \leq T_{\max}^{(i)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- $PS(\delta, \tau)$: Performance Score as a function of task completion (δ) and timeliness (τ)
- N: Total number of defined tasks
- δ_i: Binary indicator for whether task i was completed
- τ_i: Binary indicator for whether task i was completed within the allowed time
- t_i: Actual time taken to complete task i
- $T_{\max}^{(i)}$: Maximum allowed time to complete task i

Team performance additionally considered role allocation, represented by the subvariable ρ. Accordingly, the formula was extended to include the evaluation of correct task allocation in accordance with the SOPs as follows:

$$PS(\delta, \tau, \rho)_{Team} = \sum_{i=1}^{N} \delta_i \cdot \tau_i \; - \; \sum_{i=1}^{N} (1 - \rho_i) \tag{3}$$

$$\rho_i = \begin{cases} 1 & \text{if task } i \text{ was executed by the correct role (SOPs)} \\ 0 & \text{otherwise } (-1 \text{ point penalty}) \end{cases} \tag{4}$$

- ρ_i: Binary indicator for whether task i was executed by the correct role (1 = yes, 0 = no)

To ensure comparability, performance scores in faulty AI scenarios were normalized to scenario specific maximum scores that accounted for intentional AI errors and altered role allocations. Overall normalized task performance was assessed descriptively, whereas normalized team performance comprised the focus of the subsequent analyses, as it captured task fulfillment, timeliness, communication, and role allocation.

Results

Effects of team composition on workload and team performance were analyzed using linear mixed models (LMMs) with dummy coding, including the different team compositions as

fixed effects, workload and team performance as dependent variables, and the number of unforeseen failures of an implemented speech recognition API as covariate. Participant-specific variance was captured by a random intercept. Models were estimated in Python using statsmodels (Seabold & Perktold, 2010) with Restricted Maximum Likelihood (REML). Planned contrasts compared each AI-based condition with the PF + human-PM reference ($N_{Observation}$ = 170, n_{groups} = 34) in a balanced design. Fixed effects were assessed via asymptotic z-tests and β-weights were calculated following Cohen (1988) using the standard deviations of the dummy-coded predictor and the dependent variable.

Workload

Workload remained comparable between the human-PM and non-faulty AI-PM scenarios, with minor increases in mental demand and effort. In contrast, faulty AI-PM behavior increased perceived workload, particularly mental and temporal demand, effort, and frustration, with the faulty AI-PM2 condition showing the largest effect as shown in Fig. 4.

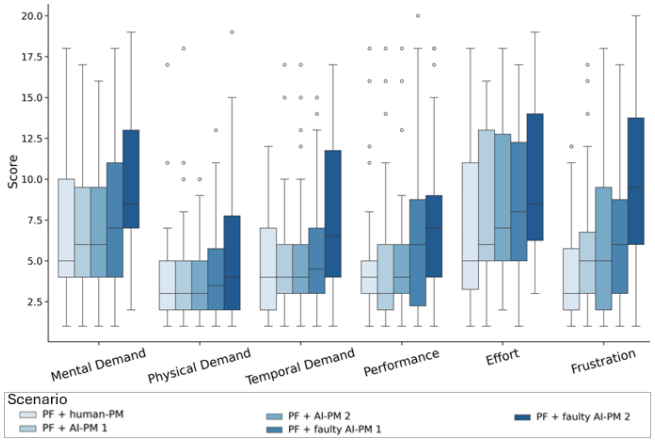

Figure 4: NASA-TLX scores per scenario and TLX dimension. Whiskers represent the range of values within 1.5 times the IQR, excluding outliers.

LMM analyses confirmed the absence of significant differences between scenarios with a non-faulty AI and the reference sencario, but small to medium significant workload increases for faulty AI compared to the reference scenario, supporting H1a and H1b (see Table 1) for the tested sample. An ICC of 0.66 revealed high between-participant variability.

Table 1: LMM results for workload: Fixed effect β-weights, the standard errors SE, p-values and 95% confidence interval.

Condition	β	SE	z	p	95%CI
AI-PM1	0.05	0.46	0.84	.403	[-0.52, 1.29]
AI-PM2	0.08	0.47	1.28	.201	[-0.32, 1.51]
fAI-PM1	0.13	0.46	2.44	.015	[0.22, 2.03]
fAI-PM2	0.35	0.47	6.46	<.001	[2.10, 3.94]

Performance

Team and overall task performance were highest in the reference condition, remained largely unchanged with AI-PM1 and AI-PM2, but decreased in faulty AI-PM scenarios, especially for interacting with the faulty AI-PM1. Overall task performance shows larger variability in the faulty AI-PM1 scenario, while team performance does in the remaining scenarios, as shown in Fig. 5.

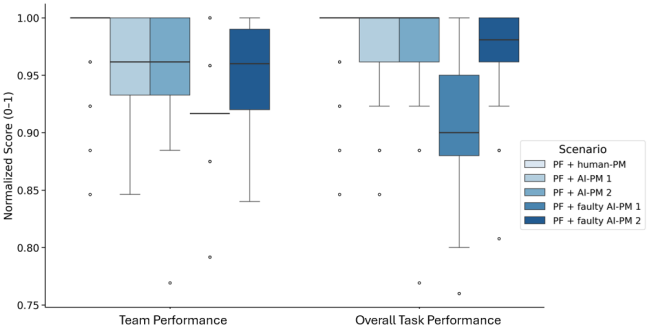

Figure 5: Normalized Team and Overall Task Performance. The whiskers represent the range of values within 1.5 times the IQR, excluding outliers.

LMM analyses confirmed the absence of a significant performance loss with reliable AI, but significant reductions in team performance with faulty AI, supporting H2a and H2b (see Table 2). Results further showed that team performance was significantly negatively impacted by speech recognition failures ($b = -0.03$, $SE = 0.01$, $z = -5.75$, $p < .001$).

Table 2: LMM results for normalized team performance: Fixed effect β-weights, the standard errors SE, p-values and 95% confidence interval.

Condition	β	SE	z	p	95%CI
AI-PM1	-0.09	0.01	-1.16	.248	[-0.02, 0.01]
AI-PM2	-0.11	0.01	-1.81	.071	[-0.03, 0.00]
fAI-PM1	-0.46	0.01	-7.29	<.001	[-0.07, -0.04]
fAI-PM2	-0.17	0.01	-2.22	.027	[-0.03, -0.00]

Discussion

Within the current debate around prerequisites for safe AI-supported SiPO, particularly the cognitive and performance-related impacts of AI-induced errors and uncertainty raises open questions. Building on recommendations by Zinn et al. (2023), we conducted task-level analyses across takeoff scenarios, explicitly considering AI failures, and actively involving pilots. Comparing both faulty and non-faulty AI-PM support to a conventional two-pilot setting in a simulated environment, we inspected both human workload and takeoff-specific team performance. Overall, the observed effects were consistent with the expected outcomes: In line with H1a and H2a, perceived workload and team performance remained on

a comparable level when interacting with the non-faulty AI-PMs. By contrast, workload was significantly increased and team performance significantly decreased when collaborating with a faulty AI co-pilot, confirming both H1b and H2b. Furthermore, the failures of the speech-recognition AI were found to have a significant negative impact on team performance and workload was shaped by strong inter-individual differences.

Extending prior work (Faulhaber, 2019; Minaskan et al., 2022; Miranda, 2025), workload remained comparable to the two-pilot scenario in SiPO with a non-faulty AI-PM, but increased significantly with faulty AI, markedly faulty AI-PM2. This pattern matches findings indicating that malfunctions and irregular events elevate workload (Faulhaber, 2019; Miranda, 2025), even when this malfunction is embodied by a faulty AI system. In line with Faulhaber (2019), descriptive results indicate that workload was driven by frustration and temporal demand. Moreover, in faulty scenarios, elevated perceived effort and mental demand may have resulted from increased monitoring and corrective task demands. The high between-participant variability likely reflects differences in experience and training, with varying ability to compensate for AI errors, consistent with de Waard (1996).

Team performance declined under irregular conditions, likely driven by increased workload and less task accuracy (Faulhaber, 2019; Miranda, 2025). The results show that workload was higher in the faulty AI-PM2 condition than in reference and faulty AI-PM1 conditions, despite stronger performance degradation when collaborating with the faulty AI-PM1. Moreover, sub-task analyses suggest that undetected AI errors were a primary contributor to performance losses observed in the faulty AI-PM1 condition. Such pattern implies that workload was elevated when AI errors were detected and actively managed, whereas undetected AI errors resulted in lower subjective workload and poorer team performance, posing a safety risk. Increased variability in overall task performance when collaborating with the faulty AI-PM1 indicates substantial inter-individual differences in how participants responded to faulty AI behavior. While some participants were able to detect and compensate for errors effectively, others showed marked performance decrements. Consequently, performance degradation may further reflect the link between SA and performance (Endsley, 2020). Overconfidence in AI-PM information or insufficient time or cues to fully comprehend the situation may have led to bad outcomes (Endsley, 2020), consistent with findings that unexpected errors reduce SA (Miranda, 2025). Pilot training differences may have further contributed to performance variability, as our sample comprised both private and commercial pilots. However, performance degradation was not solely driven by individual performance. Descriptive-level comparisons imply a dissociation between overall task success and HAIT quality. While participants as PiC largely ensured successful takeoffs, team performance was degraded under AI conditions. Speech recognition errors directly impaired communication and caused SOP-violating task misallocations, contributing directly to reduced team performance. Overall, the findings indicate that potentially unnoticed AI malfunctions, frequent speech recognition errors, and the resulting SOP violations were the main drivers of the observed decline in team performance. However, overall task success was largely maintained when AI errors were detected.

The findings underscore the benefit of simulator studies to include faulty AI behavior to capture AI uncertainty and pilot recovery strategies, informing effective HAIT in SiPO and requirements for safe AI systems in aviation (Kirwan, 2025; National Academies of Science, Engineering and Medicine, 2022; Zinn et al., 2023). Unnoticed or misinterpreted AI failures emerged as a key risk, degrading SA and team performance, highlighting the need for transparent AI and effective human-in-the-loop interaction (Endsley, 2020, 2023), while maintaining balanced workload. The observed unpredictability of the speech-recognition AI further illustrates current black-box limitations and reinforces the demand for reliable AI (Graydon et al., 2025; Kirwan, 2025). Given the high safety standards in aviation, the results require critical scrutiny. While pilots compensated for AI failures under controlled conditions, this may not hold in complex real-world scenarios where aircraft system failures could amplify these effects. Flight deck AI should be introduced incrementally, while ensuring sufficient explainability, adjusted pilot training, well-defined fallback procedures and understanding of HAIT long-term effects to sustain performance and mitigate new hazards (Kirwan, 2025; Van Droogenbroeck et al., 2025).

Key limitations include equal task weighting in the performance score and a deterministic, non-proactive AI design. Methodological constraints arose from speech-recognition reliability and limited ecological validity of the desktop-based simulator. The short study duration and limited sample diversity restrict long-term insights and generalizability. Moreover, no comparison with an unreliable human PM was included. Future work could use the recorded eye-tracking data for in-depth analyses of mechanisms underlying workload, performance, trust and SA. The multi-level performance approach proved valuable and may be transferable to other SOP-driven domains with adapted scenarios and task-sharing schemes.

Conclusions

This research contributed to empirically analyzing effects of AI on human cognition and performance in SiPO. Results show that workload and team performance remain comparable to two-pilot teams when AI functions reliably, but degrade significantly with erroneous AI. Performance was affected by unnoticed AI failures and shortcomings of the implemented AI system, while human operators were usually capable of leading the team to success. Overall, the findings indicate that AI supports performance in SiPO when it is highly reliable. With unreliable AI, team success depends strongly on AI transparency and individual human performance. Increased workload and unnoticed errors emerge as central challenges for efficient HAIT and should be tackled to embed AI as a reliable colleague in future cockpits.

Acknowledgements

Funding

We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, under the reference number UP 31/1) for the Stuttgart Research Focus Interchange Forum for Reflecting on Intelligent Systems (IRIS).

The research conducted by Michelle Fini, Zamira Daw and Maria Wirzberger was funded by the State Ministry of Science, Research, and the Arts Baden-Wuerttemberg (grant number: Az. MWK11-0430.0-1/7/5).

Additional Contributions

The authors gratefully acknowledge Alina Schmitz-Hübsch and Nadine Koch for their support with developing the research design.

Furthermore, the authors thank Lennart Lux for his contribution to the experimental study as the human co-pilot.

References

Becker, T., & Ayton, P. (2024). Effects of flight crew role assignment on aviation accidents and incidents: Evidence of a systemic safety issue. *Safety Science*, *170*. https://doi.org/10.1016/j.ssci.2023.106352

Behrend, J., & Dehais, F. (2020). How role assignment impacts decision-making in high-risk environments: Evidence from eye-tracking in aviation. *Safety Science*, *127*. https://doi.org/10.1016/j.ssci.2020.104738

Braarud, P. Ø. (2021). Investigating the validity of subjective workload rating (nasa tlx) and subjective situation awareness rating (sart) for cognitively complex human–machine work. *International Journal of Industrial Ergonomics*, *86*, 103233. https://doi.org/10.1016/j.ergon.2021.103233

Casner, S. M., & Gore, B. F. (2010). *Measuring and evaluating workload: A primer* (Report No. NASA/TM–2010-216395). NASA Ames Research Center.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

de Waard, D. (1996). *The measurement of drivers' mental workload* [Doctoral dissertation, University of Groningen]. s.n. https://research.rug.nl/en/publications/the-measurement-of-drivers-mental-workload/

Duchevet, A., Imbert, J.-P., Hogue, T. D. L., Ferreira, A., Moens, L., Colomer, A., Cantero, J., Bejarano, C., & Vázquez, A. L. R. (2022). Harvis: A digital assistant based on cognitive computing for non-stabilized approaches in single pilot operations. *Transportation Research Procedia*, *66*, 253–261. https://doi.org/10.1016/j.trpro.2022.12.025

Endsley, M. R. (2020). The divergence of objective and subjective situation awareness: A meta-analysis. *Journal of Cognitive Engineering and Decision Making*, *14*(1), 34–53. https://doi.org/10.1177/1555343419874248

Endsley, M. R. (2023). Supporting human-ai teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*. https://doi.org/10.1016/j.chb.2022.107574

Faulhaber, A. K. (2019). From crewed to single-pilot operations: Pilot performance and workload management. *20th International Symposium on Aviation Psychology*, 283–288. https://corescholar.libraries.wright.edu/isap_2019/48

Flägel, K., Galler, B., Steinhäuser, J., & Götz, K. (2019). The "national aeronautics and space administration-task load index" (nasa-tlx) - an instrument for measuring consultation workload within general practice: Evaluation of psychometric properties. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, *147*, 90–96. https://doi.org/10.1016/j.zefq.2019.10.003

Google Cloud. (n.d.). *Speech-to-text: Sprache mit der ki von google in text umwandeln* [Accessed: 10/10/2025]. https://cloud.google.com/speech-to-text?hl=de

Graydon, M. S., Holbrook, J. B., Neogi, N. A., Maddalon, J. M., & McCormick, G. F. (2025). *Challenges, research, and opportunities for human–machine teaming in aviation* (Report No. NASA/TM–2025-0002888). NASA Langley Research Center.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). North-Holland.

Kirwan, B. (2025). Human factors requirements for human-ai teaming in aviation. *Future Transportation*, *5*(2), 200–215. https://doi.org/10.3390/futuretransp5020042

Korentsides, J., Keebler, J. R., Fausett, C. M., Patel, S. M., & Lazzara, E. H. (2024). Human-ai teams in aviation: Considerations from human factors and team science. *Journal of Aviation/Aerospace Education & Research*, *33*(4). https://doi.org/10.58940/2329-258X.2046

Longo, L. (2018). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. *Proceedings of 10th Int. Conf. Computer Supported Education (CSEDU)*, 166–178. https://doi.org/10.5220/0006801801660178

Lorrig, P., Fini, M., Lux, L., Daw, Z., & Wirzberger, M. (2025). Assessing an evaluation framework for human-ai-teaming in flight deck applications. *Proceedings of the 44. IEEE/AIAA Digital Avionics Systems Conference (DASC)*. https://doi.org/10.1109/DASC66011.2025.11257385

Minaskan, N., Alban-Dromoy, C., Pagani, A., Andre, J.-M., & Stricker, D. (2022). Human intelligent machine teaming in single pilot operation: A case study. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cog-*

nition (pp. 348–360). Springer International Publishing. https://doi.org/10.1007/978-3-031-05457-0_27.

Miranda, N. (2025). Single pilot operations: Performance in normal and abnormal scenarios. *Transportation Research Procedia*, *88*, 185–192. https://doi.org/10.1016/j.trpro.2025.05.023

National Academies of Science, Engineering and Medicine. (2022). *Human-ai teaming: State-of-the-art and research needs*. The National Academies Press. https://doi.org/10.17226/26355

Puca, N., & Guglieri, G. (2025). Enabling civil single–pilot operations: A state-of-the-art review. *Aerotecnica Missili & Spazio*, *104*, 187–212. https://doi.org/10.1007/s42496-024-00223-7

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming*, *40*(3), 328–376. https://doi.org/10.1177/1046878108326734

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*.

Soudain, G. (2024). *Easa concept paper: Guidance for level 1 and 2 machine learning applications* (Report No. Issue 02). EASA.

Van Droogenbroeck, C., Rankova, E., Papenfuss, A., Bos, T., & Zon, R. (2025). Human-ai teaming - challenges from a practitioner's perspective. In D. Harris & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics* (pp. 337–354). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93718-7_22.

Xiao, Y., Wang, Z., Wang, M., & Lan, Y. (2005). The appraisal of reliability and validity of subjective workload assessment technique and nasa-task load index. *Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi*, *23*(3), 178–181.

Ziakkas, D., Pechlivanis, K., & Plioutsias, A. (2025). Enhancing aviation risk assessment through artificial intelligence: The single pilot operations case study. *Transportation Research Procedia*, *88*, 305–314. https://doi.org/10.1016/j.trpro.2025.05.037

Zinn, F., della Guardia, J., & Albers, F. (2023). Pilot's perspective on single pilot operation: Challenges or showstoppers. In D. Harris & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics* (pp. 216–232). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35389-5_16.

Zon, G., & Roelen, A. (2025). *Emco-sipo d-9 final report on risk assessments for emcos and sipos* (Report No. EASA.2022.C17). EASA.