Don't blame me: How Intelligent Support Affects Moral Responsibility in Human Oversight

Cedric Faas[1,*], Richard Uth[2], Sarah Sterz[1], Markus Langer[2] and Anna Maria Feit[1]

[1]Saarland Informatics Campus, Saarland University, Campus E 1 7, Saarbrücken, 66123, Germany
[2]Department of Psychology, University of Freiburg, Engelbergerstraße 41, Freiburg im Breisgau, 79085, Germany

Abstract

AI-based systems can increasingly perform work tasks autonomously. In safety-critical tasks, human oversight of these systems is required to mitigate risks and to ensure responsibility in case something goes wrong. Since people often struggle to stay focused and perform good oversight, intelligent support systems are used to assist them, giving decision recommendations, alerting users, or restricting them from dangerous actions. However, in cases where recommendations are wrong, decision support might undermine the very reason why human oversight was employed – genuine moral responsibility. The goal of our study was to investigate how a decision support system that restricted available interventions would affect overseer's perceived moral responsibility, in particular in cases where the support errs. In a simulated oversight experiment, participants (N=274) monitored an autonomous drone that faced ten critical situations, choosing from six possible actions to resolve each situation. An AI system constrained participants' choices to either six, four, two, or only one option (between-subject study). Results showed that participants, who were restricted to choosing from a single action, felt less morally responsible if a crash occurred. At the same time, participants' judgments about the responsibility of other stakeholders (the AI; the developer of the AI) did not change between conditions. Our findings provide important insights for user interface design and oversight architectures: they should prevent users from attributing moral agency to AI, help them understand how moral responsibility is distributed, and, when oversight aims to prevent ethically undesirable outcomes, be designed to support the epistemic and causal conditions required for moral responsibility.

Keywords

Human Oversight, Moral Responsibility, AI-supported Decision-making, Moral Agency, Risk Mitigation, High-Risk Decision-Making, Safety

1. Introduction

AI-based systems increasingly support people at work or perform work tasks autonomously. Consider, for example, drones that can autonomously deliver packages. In such safety-critical tasks, human oversight of the system is required to mitigate risk by ensuring responsibility and preserving human judgment in decisions affecting people's lives [1, 2]. If the drone is about to crash or harm others, manual control would be handed over to the overseer [3]. However, for human oversight to achieve its goals, it must be designed to be *effective*, that is overseers can detect inaccurate or inadequate system outputs, malfunctions, and intervene appropriately to prevent negative effects [4]. At the same time, Sterz et al. [5] argue that *effectiveness* also depends on the sociotechnical conditions under which oversight occurs and is closely linked to moral responsibility: to be morally responsible, overseers must satisfy both an epistemic condition (understanding why the drone crashes and how it could have been prevented) and a causal condition (having caused or contributed to the crash) [6, 7].

A potential approach to enhance overseers' epistemic access, decision accuracy, and overall safety [3, 8] is to provide additional information or specific decision recommendations [9, 10, 11, 12, 13]. These enhancements may indirectly support responsibility by helping the overseer better understand their

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
faas@cs.uni-saarland.de (C. Faas); richard.bergs@psychologie.uni-freiburg.de (R. Uth); sterz@depend.uni-saarland.de (S. Sterz); markus.langer@psychologie.uni-freiburg.de (M. Langer); feit@cs.uni-saarland.de (A. M. Feit)
0009-0000-7918-4233 (C. Faas); 0000-0001-9697-2311 (R. Uth); 0000-0002-5365-2198 (S. Sterz); 0000-0002-8165-1803 (M. Langer); 0000-0003-4168-6099 (A. M. Feit)

decision situation and prevent harm more effectively. However, in high-stakes scenarios, improving cognitive factors might not be enough. If the overseer gets bored or distracted while monitoring the drone, they will struggle to maintain situational awareness [14], and there might be strong reasons to ensure safety by restricting the overseer's freedom of choice in addition to increasing their epistemic access. In the drone example, safety could be increased by restricting the overseer from selecting particularly dangerous control options. However, such restrictions also reduce human control, diminish the overseer's moral responsibility by weakening the causal connection between their actions and potential crashes, and, in extreme cases, relegate them to the role of a scapegoat if something goes wrong [15, 16]. Empirical research raises doubts about whether human oversight achieves the desired performance [15, 17, 18], which raises normative concerns about whether humans can genuinely fulfill their assigned responsibility under certain constraints. This leaves us with a dilemma: restricting the overseer's freedom of action can improve safety and performance, yet this decision support may undermine the very reason why human oversight was employed – genuine moral responsibility.

The goal of this paper is to shed light on this dilemma and to investigate how restrictive decision support that ensures safety affects the overseer's judgments of moral responsibility. Therefore, we present results from a between-participant study where participants performed an oversight task, monitoring an autonomous delivery drone. During the study, the drone faced different critical situations that required the overseer to choose an action to resolve the critical situation (e.g., bad weather conditions) safely. Participants received decision support from a (simulated and perfectly accurate) AI system, either by showing *Six* possible actions to resolve the situation or by making only *Four* or *Two*, or *One* of these actions selectable while the other options were grayed out (i.e., not selectable, due to the AI categorizing them as unsafe). This restriction increased the overall safety of the process, since the correct action was always recommended. We hypothesized that restricting the number of *Selectable Actions* would affect participants' responsibility judgments.

2. Method

2.1. Participants

A-priori power analysis (G*Power; [19]) determined a required sample size of N = 271 to achieve a power of 1-β = .95 at α = .05, assuming a medium effect size (f = .25). We recruited 280 participants via Prolific (151 female, 129 male; age: 18-71, M = 33.25, SD = 9.81). After data cleaning in accordance with our preregistration, the final sample consisted of N = 274. Our study was approved by the first author's institution's ethical review board and preregistered via the Open Science Foundation, where we also published the collected data.

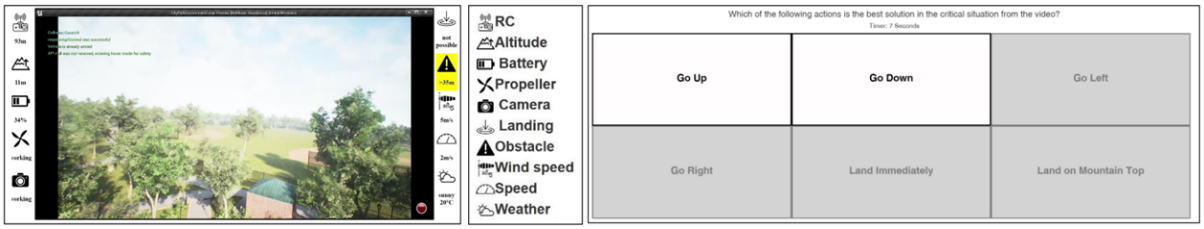


Figure 1: Screenshots of the drone monitoring interface. Left: Screenshot of the demo task, where participants saw a video recorded by the drone and information about the current status. After thirty seconds, the drone entered a critical situation, indicated by an auditory signal and highlighted icons for critical values. Middle: Legend of all icons used in the interface and introduced during the training phase; it was shown to the participants during the entire video. Right: After ten seconds, participants saw six possible actions, and depending on the experimental condition, some of them were grayed out by an AI decision support indicating they were not available (here: *Two Selectable Actions* condition).

2.2. Procedure

At the start, participants received instructions on the task, sensor values, critical situations, and possible actions, followed by a trial task. In the main task, they watched ten 40-second drone flight videos. When it came to a critical situation, an auditory signal indicated a handover to the human overseer. After ten seconds, the video stopped, and six actions were displayed, with some grayed out depending on the condition. Participants had up to seven seconds to select an action or reject all. This task provided a realistic yet relatable decision-making scenario under time constraints while maintaining experimental control to test our hypotheses. The interface and study setup ensured no prior drone expertise was needed. See Figure 1 for screenshots of the interface and selectable actions. After the last trial, participants filled out a questionnaire with Likert-scale items assessing their judgments of moral responsibility. In the final trial, all actions, even though some actions seemed plausible, would lead to a crash. This was done to ensure that participants experienced at least one undesired outcome in all conditions. See Faas et al. [20], for more details on the procedure and findings on additional questions.

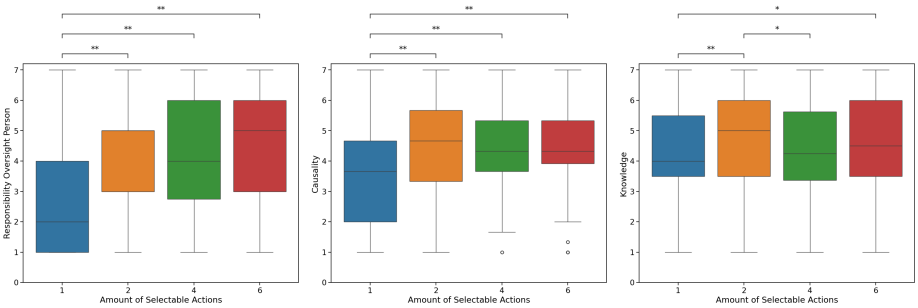
2.3. Measures

We used self-report items to measure perceived *moral responsibility of the oversight person*, *moral responsibility of the system*, *moral responsibility of the developer of the system*, *causality*, and *knowledge* on a 7-point Likert scale (1 - strongly disagree to 7 - strongly agree). All items and further exploratory variables are presented in Appendix A along with references to the literature from which we sourced these items.

We additionally captured performance-related variables: *decision accuracy* and *decision time*. Decision accuracy was calculated for each participant as the percentage of selected actions that led to a successful outcome (i.e., drone not crashing) in the first nine rounds. Decision time was measured as the time from when participants first saw the available actions until they chose a specific action.

3. Results

We report the overall effects of *Selectable Actions* on the participants' judgements about *moral responsibility of the oversight person*, *moral responsibility of the system*, *moral responsibility of the developer of the system*, *causality* and *knowledge*. To test for significant differences, we chose a Kruskal-Wallis test combined with the Dunn test with Bonferroni correction for the post-hoc comparisons of all experimental conditions. Figures with boxplots can be found in Figure 2 and Appendix B.



(a) Moral Responsibility (b) Causality (c) Knowledge

Note: *p<=0.05, **p<=0.01

Figure 2: Differences in participants' own responsibility, causality, and knowledge judgments on a 7-point Likert Scale between the experimental conditions

Selectable Actions significantly affected participants' perceived *moral responsibility of the oversight person* ($H = 98.87$, $p < .001$) and *moral responsibility of the system* ($H = 11.35$, $p < .01$). With *One Selectable Action*, participants felt less responsible than having *Two* ($p = .01$), *Four* ($p = .01$), or *Six* ($p = .01$)

Selectable Actions (Figure 2a). Participants' judgment about the system's responsibility varied more when having *One Selectable Action* compared to having *Six* ($p < .02$) *Selectable Actions*. In contrast, there were no significant differences for *moral responsibility of the developer of the system* across the conditions. Surprisingly, for all stakeholders there was no significant variance in the responsibility judgments between the *Two*, *Four*, or *Six Selectable Actions* conditions. Furthermore, *Selectable Actions* significantly affected the participants perceived *causality* (Figure 2b) ($H = 58.58$, $p < .001$) and *knowledge* (Figure 2c) ($H = 18.56$, $p < .001$).

The number of *Selectable Actions* also significantly affected the decision accuracy ($H = 128.88$, $p < .001$). Pairwise post-hoc comparisons showed, as expected, participants having *One Selectable Action* to make better decisions than participants having *Two* ($p < .001$), *Four* ($p < .001$), or *Six* ($p < .001$) *Selectable Actions*. Surprisingly, no significant differences were found when comparing conditions with more than *One Selectable Actions*, thus decision accuracy did not decrease with an increasing number of selectable actions. Notably in all conditions, a small number of decisions (*One*: 22, *Two*: 11, *Four*: 6, *Six*: 11, out of 700 per Condition) reflected that participants deliberately *rejected* all actions. We also found a significant difference in the *decision time* between the experimental conditions ($H = 28.79$, $p < .001$). Participants who had *One Selectable Action* decided faster than those with *Two* ($p < .001$), *Four* ($p < .001$), or *Six* ($p < .001$) *Selectable Actions*. No significant differences were found when comparing conditions with more than *One Selectable Actions*.

After each decision, the participants were asked three open-ended questions about the reasoning behind their decisions (see Appendix A). Their answers showed that the participants stayed engaged during the entire study and that they did not select actions at random. The answers also showed that participants understood the criteria to base their decisions on and that they were able to provide reasons for their decisions, e.g., "The camera broke, I couldn't continue the flight path" [Four selectable actions; Trial 6], even though the decision accuracy results showed that the task was challenging for them.

4. Discussion & Conclusion

We see three main limitations of our study. First, participants only experienced a simulated drone oversight task without any real-world consequences. Nevertheless, we tried to simulate a realistic situation by showing participants videos of drone flights and by simulating time pressure. Second, participants presumably had limited experience in flying drones and were not experts in human oversight tasks. This lack of expertise might have affected their overall perceived responsibility. Still, we believe that the current study included a task where our participants were able to immerse themselves and quickly judge their responsibilities. Lastly, we only measured the perceived responsibility of the participants, the system, and the developer. There might be further stakeholders that can be held morally responsible for undesirable outcomes (e.g., the deployer of the system or the overseer's organization).

Overall, our results highlight a central trade-off: restricting choices can boost safety and accuracy [21, 22, 23], but risks undermining the overseer's moral responsibility. Although potentially conflicting with design guidelines [24, 25, 1], this underscores that removing ineffective options can be a valid system design option. While limiting choice to approving or ignoring system recommendations may only be appropriate for specific oversight contexts, it might align with emerging regulatory requirements, such as those in the EU AI Act, that mandate human involvement for high-risk AI systems [5, 15].

An important challenge is to balance removing ineffective choices and preserving a meaningful decision. Our findings suggest that restricting ineffective options does not necessarily diminish perceived responsibility, as long as there is some meaningful choice left. This is promising for safety-critical domains where minimizing errors is essential [21, 3]. Thus, designers may limit user choices to improve performance, as long as users retain a meaningful decision. However, this conclusion comes with ethical risks: designers may be tempted to leave users some choice by providing pseudo-choices. As Sterz et al. [5] argue, oversight based on pseudo-choices is neither effective nor morally responsible and should not satisfy regulatory requirements, such as the EU AI Act. Our findings support this view: participants felt morally responsible only when they had a meaningful choice, and deliberately choosing inaction

was rare (44 of 2,800 decisions), likely because it constituted a pseudo-choice that could not resolve the situation. Moreover, our simulated AI always removed the most unsuitable options. If an AI were to restrict options to one suitable and one unsuitable action, oversight would become ineffective, making it unreasonable to hold the human morally responsible. On the other hand, in our study, performance improvements were only detected when no meaningful choice was left. This raises the question of whether human oversight should be required when safety-oriented interaction designs undermine overseers' responsibility without meaningfully benefiting from human involvement. This concern aligns with emerging criticism of legal requirements for human oversight [15] and underscores the need to define what constitutes effective oversight in ways that connect to the moral responsibility [5].

Responsibility is typically assigned to enable praise or blame. In the case of drone oversight, the overseer may be praised for mitigating harm or blamed for causing it. Still, other stakeholders may also bear moral responsibility. Our findings indicate that participants who had a choice attributed a shared responsibility between themselves, the AI system, and its developer. Across all conditions, participants held the AI and Developers equally responsible, but those who did not have a choice felt less responsible. This suggests that participants treated the AI as a moral agent, which conflicts with most theories of moral agency. Although increasing autonomy and capability may encourage people to view AI systems as moral agents, dominant philosophical accounts hold that moral agency is exclusive to humans who possess specific powers and capabilities, which are typically attributed to adult humans, but not to very young children, non-human animals, or computer systems [6, 7]. If oversight personnel attributed moral agency to the overseen AI, they might falsely assume moral considerations in the AI's recommendations. This perception of the AI's moral agency could influence their decision-making, which is already heavily influenced by AI behavior in morally challenging situations [26]. Therefore, oversight design should align with philosophical accounts of moral agency and responsibility [27].

Regardless of the debates about who qualifies as a moral agent, moral responsibility typically requires both a causal and an epistemic condition [6, 7]. An agent can be held responsible for an undesirable outcome only if their actions causally lead to it and if they possess sufficient understanding of the decision context. This framework aligns with our findings: participants without a meaningful choice felt less knowledgeable and perceived a weaker causal link to the drone crash compared to those with a choice. This insight can have different implications depending on the goal of human oversight or the reasons why it is employed. If the main goal of Human Oversight is to prevent ethically undesirable outcomes, interface design should prioritize enhancing the overseer's epistemic access and ensuring that they can establish a sufficient causal connection to the outcomes of the system. Overseers need support in understanding the decision situation, the AI's capabilities, and its strengths and limitations [28, 29, 24, 27]. Although decision support can increase epistemic access [30, 14, 31], designers should ensure meaningful choice and the possibility to prevent undesirable outcomes [5]. When Human Oversight is used primarily for performance improvements, limited epistemic access and causal contribution may be less critical. Even then, however, clearly communicating responsibilities remains essential, both so overseers understand the boundaries of their own responsibility if something goes wrong and to prevent them from being positioned as scapegoats [27, 15, 16].

Acknowledgments

This work was funded by DFG grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science. We thank Emilia Ellsiepen for all the support regarding the study design and execution.

Declaration on Generative AI

During the preparation of this work, the authors used Chat-GPT and Grammarly in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. URL: https://www.nature.com/articles/s42256-019-0088-2. doi:10.1038/s42256-019-0088-2, publisher: Nature Publishing Group.

[2] R. Crootof, M. E. Kaminski, W. N. I. Price, Humans in the Loop, Vanderbilt Law Review 76 (2023) 429. URL: https://heinonline.org/HOL/Page?handle=hein.journals/vanlr76&id=447&div=&collection=.

[3] J. Lundberg, M. Arvola, K. L. Palmerius, Human Autonomy in Future Drone Traffic: Joint Human–AI Control in Temporal Cognitive Work, Frontiers in Artificial Intelligence 4 (2021). URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.704082/full. doi:10.3389/frai.2021.704082.

[4] M. Langer, V. Lazar, K. Baum, On the Complexities of Testing for Compliance with Human Oversight Requirements in AI Regulation, 2025. URL: http://arxiv.org/abs/2504.03300. doi:10.48550/arXiv.2504.03300, arXiv:2504.03300 [cs].

[5] S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, On the quest for effectiveness in human oversight: Interdisciplinary perspectives, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, volume 15 of *FAccT '24*, ACM, 2024, p. 2495–2507. URL: http://dx.doi.org/10.1145/3630106.3659051. doi:10.1145/3630106.3659051.

[6] M. Talbert, Moral Responsibility, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, fall 2024 ed., Metaphysics Research Lab, Stanford University, 2024. URL: https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/.

[7] M. Noorman, Computing and Moral Responsibility, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, spring 2023 ed., Metaphysics Research Lab, Stanford University, 2023. URL: https://plato.stanford.edu/archives/spr2023/entries/computing-responsibility/.

[8] V. Uraikul, C. W. Chan, P. Tontiwachwuthikul, Artificial intelligence for monitoring and supervisory control of process systems, Engineering Applications of Artificial Intelligence 20 (2007) 115–131. URL: https://www.sciencedirect.com/science/article/pii/S0952197606001229. doi:https://doi.org/10.1016/j.engappai.2006.07.002, special Issue on Applications of Artificial Intelligence in Process Systems Engineering.

[9] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3411764.3445717. doi:10.1145/3411764.3445717.

[10] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, ACM, Cagliari Italy, 2020, pp. 454–464. URL: https://dl.acm.org/doi/10.1145/3377325.3377498. doi:10.1145/3377325.3377498.

[11] B. Green, Y. Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making, Proc. ACM Hum.-Comput. Interact. 3 (2019) 50:1–50:24. URL: https://dl.acm.org/doi/10.1145/3359152. doi:10.1145/3359152.

[12] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, H. Wallach, Manipulating and Measuring Model Interpretability, 2021. URL: http://arxiv.org/abs/1802.07810, arXiv:1802.07810 [cs].

[13] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, Barcelona Spain, 2020, pp. 295–305. URL: https://dl.acm.org/doi/10.1145/3351095.3372852. doi:10.1145/3351095.3372852.

[14] R. Parasuraman, D. H. Manzey, Complacency and bias in human use of automation: An attentional integration, Human Factors: The Journal of the Human Factors and Ergonomics Society 52 (2010) 381–410. URL: http://dx.doi.org/10.1177/0018720810376055. doi:10.1177/0018720810376055.

[15] B. Green, The flaws of policies requiring human oversight of government algorithms, Computer

Law & Security Review 45 (2022) 105681. URL: http://dx.doi.org/10.1016/j.clsr.2022.105681. doi:10.1016/j.clsr.2022.105681.

[16] M. C. Elish, Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, Engaging Science, Technology, and Society 5 (2019) 40–60. URL: http://estsjournal.org/index.php/ests/article/view/260. doi:10.17351/ests2019.260.

[17] S. Afroogh, K. R. Varshney, J. D'Cruz, A Task-Driven Human-AI Collaboration: When to Automate, When to Collaborate, When to Challenge, 2025. URL: http://arxiv.org/abs/2505.18422. doi:10.48550/arXiv.2505.18422, arXiv:2505.18422 [cs].

[18] M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and ai are useful: A systematic review and meta-analysis, Nature Human Behaviour (2024). URL: http://dx.doi.org/10.1038/s41562-024-02024-1. doi:10.1038/s41562-024-02024-1.

[19] F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behavior Research Methods 39 (2007) 175–191. URL: https://doi.org/10.3758/BF03193146. doi:10.3758/BF03193146.

[20] C. Faas, R. Bergs, S. Sterz, M. Langer, A. M. Feit, Give Me a Choice: The Consequences of Restricting Choices Through AI-Support for Perceived Autonomy, Motivational Variables, and Decision Performance, 2024. URL: http://arxiv.org/abs/2410.07728. doi:10.48550/arXiv.2410.07728, arXiv:2410.07728 [cs].

[21] G. Grote, Safety and autonomy: A contradiction forever?, Safety Science 127 (2020) 104709. URL: http://dx.doi.org/10.1016/j.ssci.2020.104709. doi:10.1016/j.ssci.2020.104709.

[22] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, C. Tan, Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, ACM, 2023. URL: http://dx.doi.org/10.1145/3593013.3594087. doi:10.1145/3593013.3594087.

[23] S. Eisbach, M. Langer, G. Hertel, Optimizing human-AI collaboration: Effects of motivation and accuracy information in AI-supported decision-making, Computers in Human Behavior: Artificial Humans 1 (2023) 100015. URL: https://www.sciencedirect.com/science/article/pii/S2949882123000154. doi:10.1016/j.chbah.2023.100015.

[24] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for Human-AI Interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3290605.3300233. doi:10.1145/3290605.3300233.

[25] T. Li, M. Vorvoreanu, D. Debellis, S. Amershi, Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction, ACM Trans. Comput.-Hum. Interact. 30 (2023) 69:1–69:45. URL: https://dl.acm.org/doi/10.1145/3511605. doi:10.1145/3511605.

[26] A. Salatino, A. Prével, E. Caspar, S. Lo Bue, Influence of AI behavior on human moral decisions, agency, and responsibility, Scientific Reports 15 (2025) 12329. URL: https://www.nature.com/articles/s41598-025-95587-6. doi:10.1038/s41598-025-95587-6.

[27] C. Faas, S. Kerstan, R. Uth, M. Langer, A. M. Feit, Design Considerations for Human Oversight of AI: Insights from Co-Design Workshops and Work Design Theory, 2025. URL: http://arxiv.org/abs/2510.19512. doi:10.48550/arXiv.2510.19512, arXiv:2510.19512 [cs].

[28] Apple, Human Interface Guidelines for Machine Learning., 2023. URL: https://developer.apple.com/design/human-interface-guidelines/machine-learning.

[29] Google, Google PAIR. People + AI Guidebook., 2019. URL: https://pair.withgoogle.com/guidebook.

[30] A. Das, Z. Wu, I. Skrjanec, A. M. Feit, Shifting focus with hceye: Exploring the dynamics of visual highlighting and cognitive load on user attention and saliency prediction, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–18. URL: http://dx.doi.org/10.1145/3655610. doi:10.1145/3655610.

[31] L.-V. Herm, IMPACT OF EXPLAINABLE AI ON COGNITIVE LOAD: INSIGHTS FROM AN EMPIRICAL STUDY, ECIS 2023 Research Papers (2023).

[32] P. Hinds, T. Roberts, H. Jones, Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task, Human-Computer Interaction 19 (2004) 151–181.

[33] R. Van Dick, C. Schnitger, C. Schwartzmann-Buchelt, U. Wagner, Der Job Diagnostic Survey im Bildungsbereich: Eine Überprüfung der Gültigkeit des Job Characteristics Model bei Lehrerinnen und Lehrern, Hochschulangehörigen und Erzieherinnen mit berufsspezifischen Weiterentwicklungen des JDS, Zeitschrift für Arbeits- und Organisationspsychologie A&O 45 (2001) 74–92. URL: https://econtent.hogrefe.com/doi/10.1026//0932-4089.45.2.74. doi:10.1026//0932-4089.45.2.74.

[34] G. M. Spreitzer, An empirical test of a comprehensive model of intrapersonal empowerment in the workplace, American Journal of Community Psychology 23 (1995) 601–629. URL: https://onlinelibrary.wiley.com/doi/10.1007/BF02506984. doi:10.1007/BF02506984.

[35] G. B. Parker, M. P. Hyett, Measurement of Well-Being in the Workplace: The Development of the Work Well-Being Questionnaire, The Journal of Nervous and Mental Disease 199 (2011) 394.

[36] M. Gagné, J. Forest, M.-H. Gilbert, C. Aubé, E. Morin, A. Malorni, The Motivation at Work Scale: Validation Evidence in Two Languages, Educational and Psychological Measurement 70 (2010) 628–646. URL: http://journals.sagepub.com/doi/10.1177/0013164409355698. doi:10.1177/0013164409355698.

[37] J. A. Gailey, R. F. Falk, Attribution of Responsibility as a Multidimensional Concept, Sociological Spectrum 28 (2008) 659–680. URL: https://doi.org/10.1080/02732170802342958. doi:10.1080/02732170802342958, publisher: Routledge _eprint: https://doi.org/10.1080/02732170802342958.

[38] K. Nolan, N. Carter, D. Dalal, Threat of Technological Unemployment: Are Hiring Managers Discounted for Using Standardized Employee Selection Practices?, Personnel Assessment and Decisions 2 (2016). URL: https://scholarworks.bgsu.edu/pad/vol2/iss1/4. doi:https://doi.org/10.25035/pad.2016.004.

[39] M. Körber, Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation, in: S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, Y. Fujita (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Springer International Publishing, Cham, 2019, pp. 13–30.

[40] E. Demerouti, A. B. Bakker, The oldenburg burnout inventory: A good alternative to measure burnout and engagement, Handbook of stress and burnout in health care 65 (2008) 1–25.

[41] S. J. Motowidlo, J. S. Packard, M. R. Manning, Occupational stress: Its causes and consequences for job performance., Journal of Applied Psychology 71 (1986) 618–629. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.71.4.618. doi:10.1037/0021-9010.71.4.618.

[42] D. L. Ames, S. T. Fiske, Intentional Harms Are Worse, Even When They're Not, Psychological Science 24 (2013) 1755–1762. URL: https://doi.org/10.1177/0956797613480507. doi:10.1177/0956797613480507, publisher: SAGE Publications Inc.

[43] E. Yao, J. T. Siegel, The influence of perceptions of intentionality and controllability on perceived responsibility: Applying attribution theory to people's responses to social transgression in the COVID-19 pandemic., Motivation Science 7 (2021) 199–206. URL: https://research.ebsco.com/linkprocessor/plink?id=3d442377-a318-3c66-9d78-d82f9abb8a4c. doi:10.1037/mot0000220, publisher: Educational Publishing Foundation.

[44] S. G. Hart, L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: P. A. Hancock, N. Meshkati (Eds.), Advances in Psychology, volume 52 of *Human Mental Workload*, North-Holland, 1988, pp. 139–183. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869. doi:10.1016/S0166-4115(08)62386-9.

A. Measures

A.1. Questionnaire Items

Table 1
Questionnaire items and measured concepts

Concept	Question(s)	Ref
Moral Responsibility of the Oversight Person	'I was responsible for the crash of the drone.'	[32]
Moral Responsibility of the System	'The system was responsible for the crash of the drone.'	[32]
Moral Responsibility of the Developer of the System	'The developer of the system was responsible for the crash of the drone.'	[32]
Additional exploratory questionnaire items and measured concepts		
Perceived Autonomy	'I am so restricted by the available actions that I can hardly make my own decisions.' *(inversely coded)*	[33]
	'I can make decisions about how to prevent crashes independently.'	
Perceived Meaningfulness	'The work I do is important.'	[34]
	'The work I do is meaningful.'	
Satisfaction	'I feel fulfilled performing the task.'	[35]
Motivation	'I have fun performing the task.'	[36]
Causality	'I caused the drone to crash.'	[37]
	'I could have avoided crashing the drone.'	
	'I could have prevented the drone from crashing.'	
Knowledge	'I could have foreseen the crash of the drone.'	[37]
	'I understand why the drone crashed.'	
Intention	'I had the intention of crashing the drone.'	[37]
Control	'I felt like I was in control over the crash of the drone.'	[38]
Trust	'I can trust the system.'	[39]
	'I can rely on the available options.'	
Disengagement	'I feel that I have become disconnected from my work task. '	[40]
Exhaustion	'I feel tired performing the task. '	[40]
Stress	'I am currently stressed by the task.'	[41]
Blame	'To what extent do you think you are to blame for the drone crash?	[42]
Punishment	'To what extent should you be punished for the crash of the drone?'	[43]
Task Load (repeated subset)	'The task was mentally demanding.'	[44]
	'The pace of the task was hurried or rushed.'	
Performance	'I was successful in accomplishing the task.'	[44]
Physical Demand	'The task was physically demanding.'	
Effort	'I had to work hard to accomplish my level of performance.'	
Frustration	'I was irritated, stressed, and annoyed during the task.'	

A.2. Open Questions

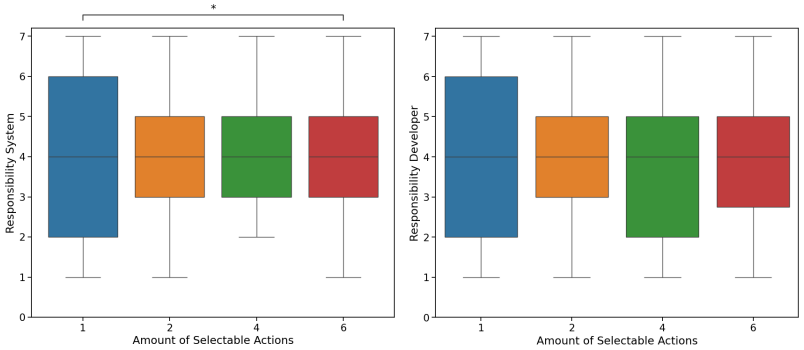
Open ended questions, the participants were asked about their reasoning after each decision:

Q1. Which information did you consider when selecting this option and why?

Q2. What about the other options? Are there any reasons why they are not as suitable?

Q3. Would you select any of the other actions and why?

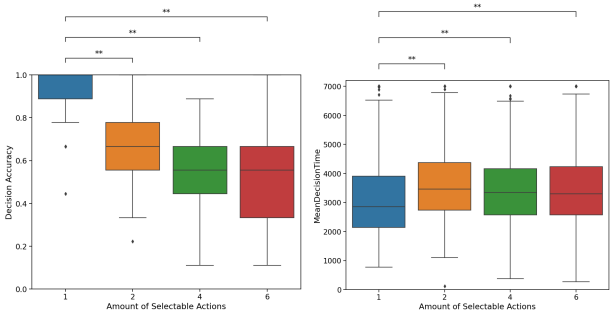
B. Visualization of Findings



(a) Moral Responsibility System (b) Moral Responsibility Developer

Note: *p<=0.05, **p<=0.01

Figure 3: Differences in responsibility judgments on a 7-point Likert Scale between the experimental conditions



(a) Decision Accuracy (b) Decision Time in ms

Note: *p<=0.05, **p<=0.01

Figure 4: Differences in decision accuracy in percentage and decision time in milliseconds between the experimental conditions