Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P8179-cd

Balancing Automation and Human Oversight: Design Implications for Safety-Critical Systems

Mina Saghafian
1) The Institute of Transport Economics, Norway.Mina.Saghafian@toi.no
2) Department of Design, Norwegian University of Science and Technology, Norway.

Ole Andreas Alsos
Department of Design, Norwegian University of Science and Technology, Norway. oleanda@ntnu.no

Jooyoung Park
Department of Design, Norwegian University of Science and Technology, Norway. Jooyoung.Park@ntnu.no

Stine Thordarson Moltubakk
University Library, Norwegian University of Science and Technology, Norway. Stine.moltubakk@ntnu.no

Lene Elisabeth Bertheussen
University Library, Norwegian University of Science and Technology, Norway. Lene.bertheussen@ntnu.no

Stig Ole Johnsen
SINTEF Digital, Norway. Stig.o.johnsen@sintef.no

As advanced autonomous technologies and artificial intelligence (AI) proliferate across safety-critical sectors, they bring both unprecedented opportunities and significant challenges, often described as automation's double-edged sword. Recent literature highlights the shift from a technocentric to a human-centric focus in designing human-automation interactive systems, aligning with the EU AI regulation's emphasis on Human Oversight. However, as Levels of Automation (LoA) and system complexity increase, maintaining human involvement, control, and the ability to intervene becomes increasingly difficult. Ensuring observability, predictability, and directability of autonomous agents is crucial to achieving transparency in design as a step towards meaningful human oversight. This paper examines the concept of human oversight, its implications for design, and its role in balancing automation's advancements with the need for human control. Drawing from the MAS (Meaningful Human Control) project, we reviewed twelve articles that explicitly reference oversight, analyzing their contributions to human oversight design principles. Our findings reveal gaps and underscore the need for stronger integration of human oversight to ensure the safety and sustainability of advanced autonomous systems.

Keywords: Human oversight, Human-Automation Interaction, Transparency, Design, Level of Automation.

1. Introduction

Human oversight in human-automation interactive systems has become increasingly critical as autonomous technologies evolve across various domains. As systems become more reliable and capable of performing complex tasks, human oversight must balance automation's capabilities with human judgment to mitigate risks and ensure safety. This paper delves into the concept of human oversight, its scope, and its design implications in high-risk autonomous systems. One key area of concern is ensuring that automated systems are designed in a way that allows for effective human oversight. The European Union's Artificial Intelligence Act (EU AI Act) emphasizes the need for high-risk AI systems to be designed in a way that allows humans to effectively oversee their operations, ensuring that human operators can intervene when necessary. According to the act, oversight measures should match the risks and the context in which the AI system is used. The ultimate goal is to minimize risks to health, safety, or

fundamental rights arising from the use of these systems (AI Act, 2024). Effective oversight requires that the operator has access to the system's capabilities and limitations, can detect and address issues, avoid over-reliance on automation, and be able to stop or modify the system's operation if needed

However, although automation is meant to relieve humans of some duties, the integration of automation comes with its own set of challenges. One significant challenge, known as the "automation conundrum," emerges when the increased reliability of automation reduces human vigilance and awareness of the system's status, thus impairing the operator's ability to intervene effectively when required. As automation takes over more functions, human operators may become less attentive, lowering their situational awareness (SA). This reduced attention can lead to errors, especially during unexpected transitions or failures in automation (Endsley, 2017). The concept of transparency in automation plays a crucial role in overcoming this conundrum. Transparency in system design, such as providing clear and understandable feedback from the system to the human operator, is critical for maintaining SA and eventually enabling timely intervention. We need to understand how transparency and oversight are related and how they will be understood and designed for in advanced AI systems.

A framework for understanding human oversight can be found in the Human-Autonomy System Oversight (HASO) model by Endsley, which depicts the interaction between system design, human capabilities, and the level of autonomy in the system. The HASO model suggests that operator performance in overseeing and intervening in automated systems is influenced by their level of SA and workload. The model also highlights that as automation becomes more reliable and robust, the operator's attention to automation-related information may decrease, increasing the likelihood of oversight failure (Endsley, 2017). In such systems, it is crucial for the overseer to have both epistemic access (sufficient knowledge to understand the system's operations) and causal power (the ability to intervene when necessary).

Effective human oversight is not just about the design of the system but also about the individual factors influencing the overseer. Traits such as vigilance, cognitive abilities, and the ability to maintain motivation are crucial to ensuring that oversight is performed effectively. For instance, individuals with greater domain expertise are better positioned to understand system errors and to intervene effectively (Sterz et al., 2024). This is beyond transparency in interface design. On the other hand, factors such as automation bias and cognitive fatigue can impede oversight effectiveness, leading to errors in judgment (Parasuraman et al., 2000).

In sum, effective human oversight in human-automation systems is a complex and multifaceted issue that requires careful consideration of both system design and human factors. As autonomous systems continue to evolve, human oversight must adapt to address the challenges posed by increasing automation, ensuring that oversight responsible personnel are equipped to manage risks effectively and maintain safety in high-risk environments. The next sections will explore the technical design features, individual factors, and system architectures that influence human oversight in these dynamic systems.

2. Method

This paper aims to explore the concept of human oversight and identify the design implications discussed in recent literature on Human-Automation Interactive (HAI) systems. To achieve this, we utilized an existing database compiled through a systematic literature review that investigates digitalization and automation design applications in safety-critical industries. The search was conducted in November 2022 and focused on publications from the past ten years. Preprint articles available at the time were also included. For detailed information on the review process, see Saghafian et al. (2025).

In this paper, we specifically filtered for the terms "human oversight" and "oversight" within the context of HAI systems to identify factors that contribute to successful system design, performance, and meaningful human control. It is important to note that this is not a systematic review solely dedicated to the term "oversight."

Rather, we aimed to understand how oversight is conceptualized and incorporated into the design of HAI systems, particularly as these systems evolve with advanced autonomy and applications of artificial intelligence. A total of 15 articles containing the specified terms were initially identified. However, two articles were excluded due to lack of full-text availability, and one was deemed irrelevant after full-text screening. The remaining 12 articles (see Table 1) were analyzed to determine how oversight is defined and implemented in the design of HAI systems. This paper is a lens into how the term oversight has been used in the aforementioned literature base. It is intended to serve as a basis to contribute to the ongoing dialogue on the concept of human oversight and its developing implications.

Table 1. The overview of reviewed articles.

Author	Date
Alonso,V., and de la Puente, P.	2018
Barnes et al.	2015
Biondi et al.	2019
Chiossi et al.	2022
F et al.	2020
He et al.	2021
Man et al.	2018
Patel et al.	2020
Tien et al.	2016
van Aken et al.	2021
van de Merwe et al.	2024
Veitch et al.	2021

3. Results

The findings show that oversight is understood very similarly to the term transparency and the implications for maintaining situational awareness, workload, and trust in the system. However, while transparency focuses on the design of the interface in HAI systems, oversight seems to be the next step that enables the transition of control to human operators to restore the situation back into a safe state as well as maintain control and monitoring. We present how oversight was referred to and what it entailed and present the design implications retrieved from the articles that were analyzed.

3.1 *Oversight in HAI Systems*

Oversight refers to the role of human operators in supervising, monitoring, and intervening in automated systems to ensure safety, performance, and alignment with human goals. The following insights summarize oversight based on the articles considered.

3.1.1 *Transparency and Explainability for Oversight*

There is an overlap between the terms transparency, explainability, and oversight in this field of literature. Although these terms are very close in their design implications and their final goal, there are subtle differences. Transparency is critical for oversight as it enhances understanding, predictability, and trust in human-automation interactions. In shared autonomy, transparency improves interface design and helps address issues like failure detection and complacency effect, and control recovery after automation failures. According to Alonso & De La Puente, (2018) transparency frameworks should enable users to ask:

- Why did the system perform action A instead of B?
- When does the system fail or succeed?
- When is the system trustable?
- How can the operator correct errors?

This approach aligns with oversight by bridging explainability to human intervention.

To oversee autonomy is multifaceted. In maritime autonomous systems, "autonomy operators" are responsible for bridging autonomous AI tasks and human responsibilities. Their oversight includes monitoring mission-critical data, such as vessel conditions and marine traffic, and enabling backup interventions when necessary (Veitch et al., 2021). However, situational awareness and experience play a role in how well humans can oversee autonomy when they are removed from the site and moved to a remote operation room. It was found that experienced operators perform better in high-stakes oversight tasks, emphasizing the need for experience in forming mental models that help prioritize tasks and attend to system failure alarms (Man et al., 2018).

Furthermore, remote oversight introduces challenges due to the ecological shift in work domains, as seen in industries like autonomous maritime shipping. Therefore, in addition to experience and individual differences, designing

the new work domains and accounting for ecological factors is necessary.

Furthermore, remote oversight in the maritime sector still requires filling the regulatory vacuum of international regulatory standards. In this new work domain it is more difficult for overseeing person to form an accurate situational awareness and adopt the safest course of action. This new work domain requires regulatory update for maintaining oversightand having a legal base for the intervention (Man et al., 2018).

In the context of automated vehicles, oversight challenges include regaining control during emergencies while multitasking. Studies highlight that human operators often need approximately 27 seconds to recover from secondary tasks before taking over control, emphasizing the complexity of transition periods in high-level automation. (Biondi et al., 2019). This implies that not only is the human operator expected to maintain situational awareness, but they are also expected to form an accurate situational awareness, choose a 'correct' course of action and intervene, while engaged in other tasks that are now part of their obligations.

The advancing technologies, such as automation, put added pressure on overseeing human operators to do all the above, while in addition to that, making a political and ethical judgement of the consequences of their actions and that of the automation and taking the best decision (Barnes et al., 2015). Researchers emphasize ethical oversight in AI systems to address concerns like safety, bias, and transparency. For instance, ethical oversight is needed to enable human override of machine decisions and ensure system trustworthiness through reliability and safe performance (He et al., 2021).

3.2 *Design Implications for Oversight*

Effective system design can enhance oversight capabilities by addressing key challenges related to transparency, workload, and human-automation interaction. The following are core design implications.

3.2.1 *Transparency and Interface Design*

Systems should be designed to improve transparency by making automation processes observable ("seeing-into"). This helps human operators understand system dynamics, anticipate outcomes, and make informed interventions. Transparent designs should also account for specific task allocations and information needs based on the function distribution between humans and agents. (van de Merwe et al., 2024).

3.2.2 *Mixed-Initiative Architectures*

According to Barnes et al. (2015), effective oversight involves balancing human supervision and system autonomy. For systems with multiple agents, adaptive, adaptable, and mixed-initiative designs reduce cognitive load by supporting task delegation and oversight. Systems with partial autonomy should incorporate adaptive designs sensitive to environmental and human states. Mixed-initiative systems allow collaborative decision-making and ensure human oversight of critical actions. These models reduce cognitive load and improve performance in dynamic contexts. (Barnes et al., 2015).

3.2.3 *Tactile Feedback for Situation Awareness*

Tactile displays can enhance situational awareness in semi-automated systems by directing operator attention more effectively than auditory or visual stimuli. Such designs are particularly useful in maintaining oversight during time-sensitive maneuvers or automation failures (Chiossi et al., 2022).

3.2.4 *Standardization and Consistency*

Standardizing system alerts and threat estimation across manufacturers can reduce operator confusion and enhance trust. For example, variations in vehicle collision avoidance systems' sensitivity can impact driver vigilance, necessitating standardized approaches. (Fu et al., 2020).

3.2.5 *Human-Centered Design for Automation*

Oversight in automated systems should follow a human-centered design approach by:

- Identifying the appropriate levels of automation for tasks like information acquisition, decision-making, and action implementation.
- Evaluating the impact of automation on workload, trust, and situational awareness.

This ensures operators are not overwhelmed by multitasking and can effectively intervene when needed (Biondi et al., 2019).

3.2.6 *Training and Expertise*

Oversight roles demand operators with experience and specialized training that are better in situation assessment. Skilled operators are better at delaying unnecessary interventions and managing emergencies even when multiple alarms are activated. This highlights the importance of matching oversight tasks to operator expertise (Man et al., 2018).

3.2.7 *Ecological Considerations for Remote Systems*

Transitioning to remote supervision (e.g., in maritime operations) requires designs that account for new work domains and operator challenges. Interfaces must be intuitive, provide actionable feedback, and align with international regulatory standards (Man et al., 2018).

3.2.8 *Balancing Sensitivity and Trust*

Designing systems that avoid excessive caution (e.g., false alarms in emergency braking) can prevent operator complacency and maintain vigilance. Systems should strike a balance between sensitivity and reliability to optimize oversight (Fu et al., 2020).

4. Discussion and Conclusion

Although oversight and transparency seem very close in definition and design implications, one could argue that transparency must be designed in the interface, and an additional step for a smooth and safe transition to the human operator must be added to allow oversight. On the one hand, with advanced automation and widening applications of artificial intelligence and deep machine learning, we would expect the human operator role to be merely supervisory, thus making it seem less demanding and reduced to a 'fall-back' role. However, the literature shows that the opposite might occur in mixed human-agent systems where a human is expected to engage in secondary tasks but must take over primary system tasks, regain situational awareness, conduct a risk assessment, and decide on the correct course of action to avoid disaster. Ultimately, the responsibility for the consequences will be with the human operator even though their involvement and understanding of the system's doing is reduced. Furthermore, the technological push asks for increased trust in a system that can 'think' for itself. It is important to remember that increased automation is meant to facilitate human performance and not demand even more cognitive load and competition between the human and the autonomous agent in the HAI systems.

The question that must be asked is if every system and every application must be fully or highly autonomous? Where is the safety margin and where is the efficiency margin in task allocation in HAI systems? Is this a trend that we must all follow and apply in our respective systems, or can we engage in a sense-making process whereby a thorough analysis of task delegation and risk assessment should provide reasonable grounds for automation? If the concept of human oversight emphasizes finite accountability of the human overseer, what does that imply for the industry-wide design standards and practices?

Automation has the potential to create both trust and complacency in human operators. The more reliable the system, the more likely the operator is to trust it and disengage from actively monitoring the system. However, this trust may lead to over-reliance on automation, making it harder to notice when intervention is required (Sterz et al., 2024). Thus, a balance must be maintained between trusting the system's capabilities and remaining vigilant. Human oversight, therefore, is not a passive supervisory role but an active managerial role, especially in systems with a high level of autonomy. In this context, human oversight extends beyond just transparency. While transparency ensures that the operator can understand the system's operations, oversight involves the authority and capability to intervene, reverse faulty decisions, and adjust the system's parameters to improve outcomes. In this regard, oversight can be seen as a step beyond transparency, requiring operators to not only understand the system but also possess the ability to take control and manage the system effectively.

Acknowledgement

This research is funded by Meaningful Human Control of digitalization in safety-critical systems (MAS) project, Norwegian Research Council, RCN 326676.

References

Alonso, V., and De La Puente, P. (2018). System transparency in shared autonomy: A mini review. *Frontiers in neurorobotics, 12*, 83.

Artificialintelligenceact.eu. (2024). *AI Act: Article 14*. Retrieved from Artificial Intelligence Act.

Barnes, M. J., Chen, J. Y., and Jentsch, F. (2015, October). Designing for mixed-initiative interactions between human and autonomous systems in complex environments. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1386-1390). IEEE.

Biondi, F., Alvarez, I., and Jeong, K. A. (2019). Human–vehicle cooperation in automated driving: A multidisciplinary review and appraisal. *International Journal of Human–Computer Interaction, 35*(11), 932-946.

Chiossi, F., Villa, S., Hauser, M., Welsch, R., and Chuang, L. (2022, June). Design of on-body tactile displays to enhance situation awareness in automated vehicles. In *2022 IEEE 9th International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1-6). IEEE.

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human Factors*, 59(1), 5-27.

Fu, E., Johns, M., Hyde, D. A., Sibi, S., Fischer, M., and Sirkin, D. (2020, April). Is too much system caution counterproductive? Effects of varying sensitivity and automation levels in vehicle collision avoidance systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., and Mehnen, J. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems, 14*(4), 1398-1412.

Man, Y., Weber, R., Cimbritz, J., Lundh, M., and MacKinnon, S. N. (2018). Human factor issues during remote ship monitoring tasks: An ecological lesson for system design in a distributed context. *International Journal of Industrial Ergonomics, 68*, 231-244.

Patel, H., Kalghatgi, S., Feijo, L., and Scott, R. (2020, May). Unmanned/Minimally Manned Floating Deepwater Installations: Design and Safety Considerations. In *Offshore Technology Conference* (p. D031S032R006). OTC.

Saghafian, M., Vatn, D. M. K., Moltubakk, S. T., Bertheussen, L. E., Petermann, F. M., Johnsen, S. O., and Alsos, O. A. (2025). Understanding automation transparency and its adaptive design implications in safety–critical systems. *Safety Science*, *184*, 106730.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495-2507).

Tien, S. L. A., DeArmon, J., Bateman, H., Freer, D., and Somersall, P. (2016, September). Developing a real-time monitoring and alerting capability for traffic flow management. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1-10). IEEE.

van Aken, D., Janisch, D., and Borst, C. (2021). Development and Testing of a Collaborative Display for UAV Traffic Management and Tower Control. In *Fourteenth USA/Europe Air Traffic Management Research and Development Seminar* (pp. 1-10).

van de Merwe, K., Mallam, S., and Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review.

Veitch, E. A., Kaland, T., and Alsos, O. A. (2021). Design for resilient human-system interaction in autonomy: The case of a shore control centre for unmanned ships. *Proceedings of the Design Society, 1*, 1023-1032.