Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27810-cd

Beyond the Ethical Compass: Reliability Challenges in Adaptive Human-AI Teaming

Dirk Söffker
Chair of Dynamics and Control, University of Duisburg-Essen, Germany. E-mail: soeffker@uni-due.de

Olena Shyshova
Chair of Dynamics and Control, University of Duisburg-Essen, Germany. E-mail: olena.shyshova@uni-due.de

Traditionally, the ethical compass guides human decisions, often in contradiction to technically best solutions (e.g., AI-only decisions with low error probability). Studies show that human-AI combinations do not automatically generate synergies; on average, they perform worse than humans or AI alone. The decision of when humans should trust superior algorithms is particularly critical. Human augmentation and synergies in generating tasks are positive. Meta-analyses (106 studies) confirm that performance and reliability depend on tasks and the relative strengths of both. Advantages only arise in certain situations or combinations; wrong decisions occur when AI outperforms humans. Effective cooperation requires situational adaptation and understanding of the error structure. When ethical criteria are integrated into AI systems, the focus shifts to functional reliability related to the constellations: human-only, AI-only, and human-AI teaming (HAT). Dynamic workflows and flexible teaming structures enable adaptive collaboration. A key question is which methods or combination of methods will describe adequately the merged hardware reliability, functional reliability of AI, and the reliability of supervising humans. The paper conceptually outlines the fundamental challenges associated with the following research areas, which become relevant in this context: 1. Reliability-Aware Adaptive HAT: Situational AI adaptation (goals, KBI, emotions) and human interpretation 2. Situated Workflow Control: Integration of knowledge and situational control 3. Error and Knowledge Bases: Workflow knowledge, individual errors, and fundamental human and AI error structures to improve reliability 4. Reliability-oriented Metrics and Evaluation: Reliability of humans/AI and task type (decision vs. generation-oriented) are decisive; XAI/confidence alone does not provide significant synergy 5. New Reliability Descriptions: Methods for combined reliability of humans, AI, and hardware are required. Based on existing findings, this paper provides firstly a related structured approach to a reliability-oriented human-AI teaming task and presents the methodology both conceptually and through examples. In conclusion, evidence-based human-AI teaming systems require dynamic adaptation, situational workflow control, and robust reliability metrics. Reliability becomes a central design goal, extending beyond ethical control.

Keywords: Human-AI Teaming, Adaptive Human-AI Teaming, Resilient Human-AI Collaboration, HAT Reliability, Human-AI Synergy, Situated Workflow Control.

1. Introduction and initial indications of a fundamental ethical-functional conflict

The demand for responsibility and the associated necessity for morality in human actions, often reinforced or established by explicit stipulations in specific fields of application (e.g., the Geneva Conventions in the military sector), lead to a final decision being made by humans or human operators of technical systems (Söffker and Shyshova, 2025). This also applies, for example, to applications in the consumer sector, where the same conclusion (humans make the final decision) is drawn for liability reasons. Regardless of such motivation, however, there are other levels of consideration that raise a critical new conflict in the case of ethical or moral requirements or specifications. In the case study on object recognition in road traffic, for example, the purely technical AI decision (AI-only) achieved the highest reliability with the lowest error probability (11.3 %) (cf. Shyshova et al. (2025)), surpassing all decisions involving humans. While the ethical compass dictates the final human decision, this conflicts with the best technical solution in terms of the reliability of the correct decision and, consequently, the functional decision.

2. The reality of Human-AI cooperation

This (partial) finding is supported by recent meta-analyzes on human-AI collaboration (Vaccaro et al., 2024), which challenge the traditional and current paradigm (Berretta et al., 2023) that the combination of humans and AI automatically leads to synergies. The hope of solving current problems with the help of AI or with the support of AI can not be justified at present. A systematic review as meta-study of 106 experimental studies states that Human-AI combinations performed significantly worse on average than the best results achieved by humans alone or AI alone (human-AI synergy: Hedges' g = -0.23). Artificial Intelligence only conviences only in selected fields: in many task constellations, however, collaboration leads to a deterioration in performance parameters and the reliability of human-machine systems. According to Vaccaro et al. (2024), this applies in particular to decision-making tasks, where the combination of humans and AI led typically to performance losses on average (g = -0.27). It can be seen as critical that performance losses occurred when AI alone outperformed humans (g = -0.54). This suggests that humans are often not performing at deciding in which specific situations they should trust the (then superior) algorithms. The results are consistent with the statistically validated results of the study on action-relevant recognition tasks in humans (Shyshova et al. (2025)). However, the results also show that, on average, AI augments humans in combination (human augmentation: g = 0.64). It is important to note, for example, that a positive synergy was found in creation tasks (g = 0.19), which points to possible research approaches and design ideas in this area.

3. From Ethical Constraints to Design and evaluation Requirements: Implications for System Design

When moral arguments and requirements for formal or legal accountability lose importance because, for example, ethical criteria are merged into AI systems (cf. Defense Advanced Research Projects Agency (DARPA) (2024)), the purely technical, functional perspective and thus also the reliability of the overall system become relevant. The current basic structure (fundamental final decision made by humans) is then essentially dissolved. This allows for new design options, which essentially boil down to three basic constellations: human only, AI only, human-AI teaming in all variants and with complete flexibility at the beginning, end, and variable workflow design. Artificial Intelligence is then gaining importance in fields that did not previously allow this, and new constellations, including especially variable configurations, are possible for HAT issues. The focus is thus shifting toward mutual adaptation, which involves situational and presumably highly dynamic collaboration. In the context of this considerations/work, the focus is on the requirements for the design and evaluation of such systems. In the context of this work this is focused to the resulting questions regarding the reliability of human-machine/human-AI systems and human management and monitoring of autonomous systems require both a redesign of interaction protocols and appropriately adapted evaluation standards for reliability engineering. The following research directions can be justified (generically) for this purpose:

3.1. *Future challenges and design approaches (adaptive human-AI teaming)*

3.1.1. *Adaptive and reliability-aware HAT frameworks*

Following the above considerations, the correct execution of action sequences, correct decision-making, and correct recognition and recording of content and details in complex human-machine/human-AI/decision-making processes are fundamentally more important than other criteria. For the reasons mentioned above, this is easy to understand from a technical and content-related perspective, but is not necessarily always the case. As a measure of the correctness of assignments or correct recognitions, etc., "reliability" is defined here as "Reliability describes the probability that a system performs its intended function without failure over a defined period of time," in contrast to "resilience characterizes the

system's ability to withstand and recover from disturbances." In AI-based systems, reliability must therefore be interpreted as consistent functional behavior under operational conditions, so that the focus here is also on correct functional performance regardless of time. Although this contradicts the pure definition, it does not blur the core distinction from terms such as 'resilience' and 'security'.

The logical first step in this development focuses on the appropriately defined reliability and emphasizes the flexibility of the workflow as "Reliability-oriented adaptive Human-AI teaming" or, more succinctly, "Resilient Human-AI collaboration." This requires the development of a suitable framework that combines the aforementioned goals of 'reliability-oriented' and 'adaptive'. Central to this – and in contrast to the classic definition of reliability – is the time reference: the framework/adaptation/reference must be situational (and not lifetime-based) and take into account AI knowledge references (knowledge-based), action goals, and, if applicable, human performance as well as emotions on the one hand, and provide explanations/recommendations and explain and take into account the ongoing activities of the AI on the part of the acting human on the other. Here, situated refers to the moment and the current problem constellation (cf. Söffker (2008)) as well as the contextual framework of situational awareness as developed by Endsley (1995).

3.1.2. *Situated workflow control*

The definition of tasks themselves, the assignment of roles, control over the course of action, decisions regarding the success of the outcome of the action (with the need to revise if necessary) and thus over the role assignments of humans and machines/AI automatons as a team, or as a fluid whole, with the aim of performing individual tasks in specific situations (time-oriented, state-oriented) as well as possible. In this case, the definition of everything is unknown from the outset, even though fixed behavior patterns will certainly emerge in practise. The proposed idea/framework (cf. Fig. 1) is similar to the concepts of model predictive control, whereby situational system understanding and behavior are exchanged and updated in line with the restrictions or boundary conditions (in this case, essentially reliability-oriented).

For situational decisions, (in)variant interaction logic decisions must be made that largely correspond to the aforementioned decision logic: human decides, AI decides, or fluidic-interactive, whereby fluidic-interactive is the general case adressed here for the first time and means that priorities are only set during an interaction scenario, as they can only become decisive in a logical and rational manner (sufficient to a given criterion) when a decision is made. It can be assumed that LLMs can also be used for standardized situations. In principle, interactions can be designed to be fluidic, but regardless of this, invariant or current knowledge will be used.

3.1.3. *Knowledge base and human error structures*

If the team consisting of humans and machines or humans and AI is to function optimally (in terms of the criteria defined below), knowledge is of central importance. In the context of human-machine/human-AI interaction, in addition to factual knowledge about the system or process, this includes fundamental knowledge that is invariant across applications regarding i) the interaction workflow (as known from action organization [references]), ii) individual human errors and their probabilities (as also known from human reliability research [references]), and iii) the fundamental structure of human errors in recognition, cognitive processes such as planning and decision-making, and action control (cf. Fig. 2)(Söffker (2004a,b); Dörner (1990)).

3.1.4. *Forms of synergistic interaction between adaptive, situated human-AI interaction*

The meta-analysis by Vaccaro et al. (2024) shows that neither the presence of an explanation (XAI) nor the indication of AI confidence/confidence bound had and has a significant impact on the synergy of the human-AU team. This is a surprising result because it proves that previous research efforts, while valuable in themselves, have not led

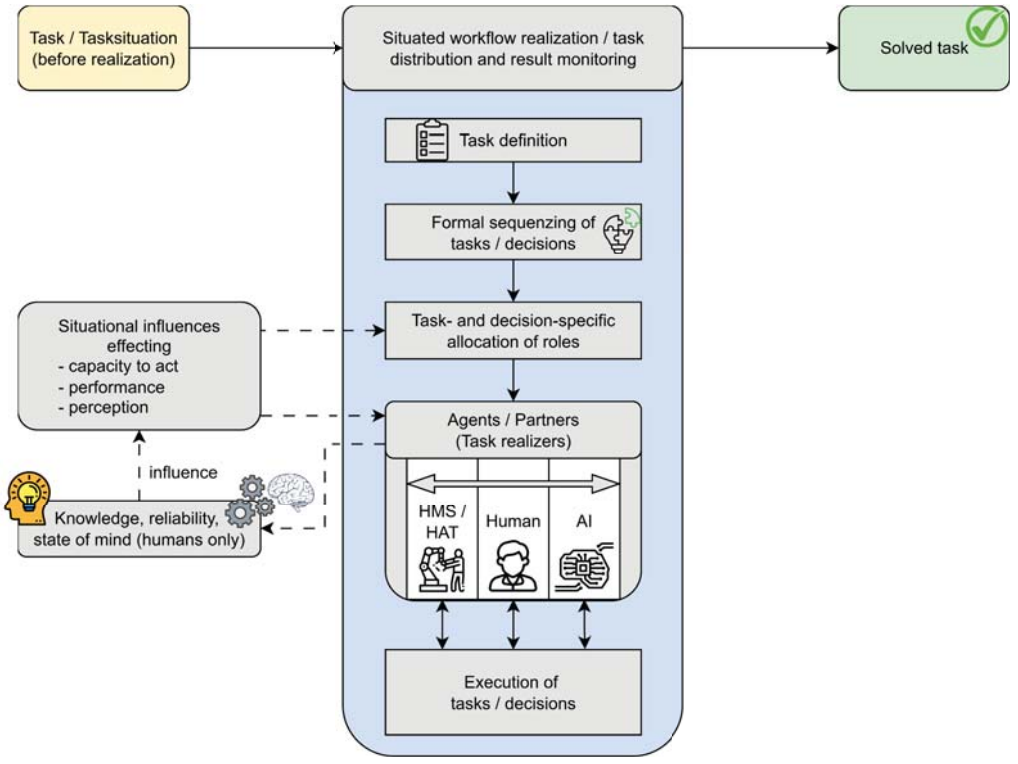

Fig. 1. Overview of the proposed situated workflow control

to the qualitative and content-related improvement of human-machine or human/AI systems that was the intended goal of previous research efforts of the last years. However, the same study also shows how advantages can be derived from the symbiosis of humans and AI.

Designing collaboration for ”creative” tasks: Relevant factors include the relative performance of humans and AI and, in particular, the type of task (decision-making (no advantages) vs. innovative new generation (denoted as ”creative”) (advantages)). These tasks are highlighted positively in Vaccaro et al. (2024). Creation tasks are those in which (in general) team members generate open-response content. In contrast to decision-making tasks, which were clearly associated with performance losses, the pooled effect size for Human-AI synergy in generation/creation tasks showed a positive value ($g = 0.19$). The difference between the losses in decision-making tasks and the gains in creation tasks is described as statistically significant in Vaccaro et al. (2024). The authors in Vaccaro et al. (2024) do not cite specific experimental examples from the meta-analysis, but describe the underlying mechanism using generic example scenarios:

- Text creation: Generating many types of text documents requires human knowledge or insight, but also filling in boilerplate or routine parts of the text. Here, AI can do the routine part better or faster, while humans contribute the creative or knowledge-based parts.
- Artistic image generation: Generating a good artistic image requires creative and adaptive inspiration (human) and routine characterization of details (AI).

What does this mean for the distribution of

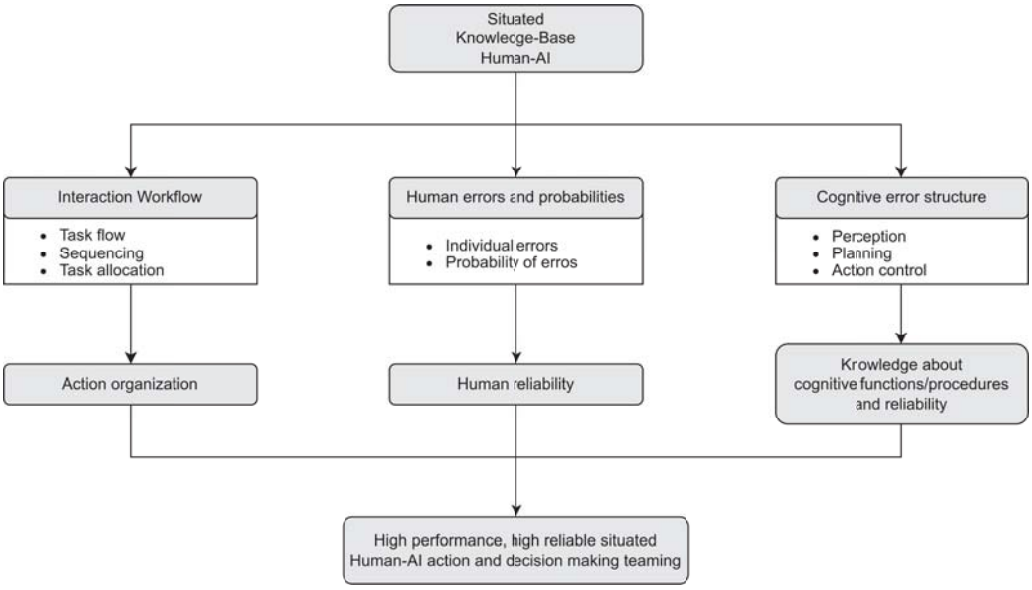

Fig. 2. Knowledge and human error structures in Human–AI interaction

work in a concrete, super-flexible HAT workflow in which decision-making and evaluation tasks (which are disadvantageous on their own) also occur? Based on the above examples, the implicitly suggested distribution of work is as follows: routine/standard tasks are performed better and faster by AI, while creative, knowledge-based tasks are performed by humans. This means that, for example, in evaluation tasks, the rough evaluation/recognition is specified and formulated by humans, e.g., as a question or hypothesis, possibly with alternatives, while machines automatically check and output these hypotheses/alternatives, possibly with alternatives and evaluations (routine consideration). Humans record the evaluation or alternatives and narrow them down, formulate new hypotheses, reject old statements, and in this sense: The AI evaluates humans as a sparring partner.

Cooperation through clear differences in relative initial performance: According to Vaccaro et al. (2024), a positive synergy is significant when humans alone outperformed AI alone. When humans were superior in terms of initial performance, the combined human-AI system significantly outperformed both individual partners. According to Vaccaro et al. (2024), the pooled effect size for human-AI synergy in this case was $g = 0.46$, which is rated as a medium-strength effect. The authors in Vaccaro et al. (2024) cite a specific experimental example: bird image classification. In this case, humans alone achieved a recognition performance of 81 %; the AI algorithm alone achieved 73 %, and the human-AI system achieved an accuracy of 90 %. According to Vaccaro et al. (2024), the synergy in this combination is explained by the fact that when humans are more reliable than algorithms in specific contexts, they are also better at deciding when to trust their own judgment and when to trust the algorithm. The crucial point here is that the ability (trust) only exists when humans know that they are good to very good (and therefore really know how to adequately evaluate the innovation of the control AI).

What does this mean for the distribution of work in a concrete, super-flexible HAT workflow in which decision-making and evaluation tasks (which are disadvantageous on their own) also occur? Based on the above examples, the implicitly suggested distribution of work can

be as follows: Whenever it is generally and/or situationally known that the human reliability for selected/upcoming actions is higher than that of machines, these actions must be carried out by humans. Machines/AI automata then inevitably only play the role of companions/sparring partners, who are also allowed to make judgments/suggestions, but these are only 'presented' if they deviate. In this case, they are presented/visualized with the note/request to review the human judgment or similar once again, but not with the advice/authority to change it. In this way, the team character remains in the decision-making process, but the human being is still primarily responsible and decisive in this situation/task/state due to their known higher competence/reliability, while at the same time being placed in the role of deciding on the innovation of the AI. There is no third party to make an assessment/weighting/decision.

3.1.5. *Metrics and evaluation*

If ethical/moral standards for structuring coexistence are left out of the consideration, what remains are technical evaluation standards that are well-defined and known. Reliability engineering provides the failure rate over the lifetime of a system, which addresses the (typically one-time) failure of the system and its functionality (at the end of its useful life) (case A); in computer science, the confusion matrix is used to compare the correct statements with the claimed statements (case B). For both perspectives, there are a large number of derived metrics for evaluating the performance of systems in correctly executing tasks and/or making correct statements. In the context of human-machine systems (HMS), human error probability (HEP) is a comparable measure of the percentage of correctly executed activities (Reason (1990); Kirwan (1994); Boring (2007); Groth and Mosleh (2012).

As in classical reliability engineering, such figures are used as a statistical measure for a set of comparable activities (correctness of execution in the context of situational execution), comparable to the combination of cases A and B. For HEP, adjustment parameters are parameterized to take into account further personal or environmental effects (so-called performance shaping factors (PSF)) (Xing (2021); D.I. Gertman (2021); Groth and Swiler (2013)), all of the above metrics are static and therefore not suitable for describing the situational existence of a human-machine system (or autonomous system), as this always involves a very specific constellation of influences that can come together at short notice and can be resolved or influenced just as quickly.

The approach 'Dynamic HRA' Boring (2007): "Dynamic Human Reliability Analysis: Benefits and Challenges of Simulating Human Performance" is adressing some kind of situated reliability performance metric, considering dynamically (here: changing quasi stationary) changing situations. Also He et al. (2022), are proposing a reliability-oriented metric for continously changing dynamical situations.

However, if it is possible to formulate (and therefore calculate) such a measure and dynamically anticipate situations, it is also possible to increase the reliability of the MMS.

3.1.6. *New reliability description for relative and situated performance*

For the new perspective proposed in this paper, it is necessary to determine appropriate forms of description for the reliability of humans and human-AI systems in dynamic decision-making processes. Therefore the question remains as to which standards or combinations of standards will be used to describe the new forms of interaction that adequately describe the combination of i) hardware reliability of the machine/automated system, ii) functional reliability of the AI (e.g., in recognition tasks or binary conclusions), iii) the fundamental reliability of supervising/decision-making humans (HEP with PSF), and iv) the individual, situational ability to function correctly. The procedures for i) and iii) are known through proven methods, e.g., (Boring (2007); Groth and Swiler (2013); Det Norske Veritas group (DNV) (2024); Naval Sea Systems Command (1991)). With regard to functional reliability for specific tasks under specific boundary conditions, probabilities (e.g., in tabular form) can also be determined

that can be applied. This applies in particular to technically easily determinable variables such as procedures, tasks, object sizes, environmental conditions, etc., which can also vary greatly due to their high variability. A major challenge lies in point iv. Similarly to the approach taken in He et al. (2022); Shyshova et al. (2025), an online measure influenced by updated environmental variables can be determined, which expresses an unscaled, thus relative, time-variant measure that describes the current reliability. With knowledge of the individual variables and the relationships in the different workflows as series/parallel connections, the reliable combinations can be selected online or visualized (as recommendation system etc.).

4. Dynamic and situated HAT, a concept: Summary and Conclusion

Building on the ethical dilemma identified earlier, this article concludes from literature knowledge in combination with own results that human-AI teaming systems are proving to be unreliable constellations in important application contexts, e.g., object recognition and decision-making, and that new considerations are needed here. Building on the findings of a meta-study on the reliability of human-AI combinations, as reported in the literature, this article examines and explores in greater depth and firstly the question of in which contexts human-AI teams are more reliable than the two original systems (human, AI). The article shows that the predictable development of HAT systems requires a shift away from a sole focus on ethical control in favor of evidence-based design based, for example, on dynamic adaptation, situation-specific workflow control, and robust performance measurement, with reliability improvements and the development of suitable metrics playing a decisive role. The developed approach (as concept) is based on the fact (well known in the literature), that 'creative' task design allows the increase of the reliability of human-AI teaming systems. Based on that this article develops in detail a theoretical concept that pursues the reliability of the overall interaction concept, which consists of best practices and methods, as well as new dynamic tools developed in other contexts, and therefore allows for dynamic and evidence-based reliability assessment and thus online optimization of human-AI teaming processes and interactions for future HAT systems. The framework presented thus avoids the known weaknesses of specific combinations of human-AI teaming. The strict emphasis on reliability as a design parameter is also new and firstly stated. Future work will involve experimentally validating the proposed approach through specific human-AI team tasks.

References

Berretta, S., A. Tausch, G. Ontrup, B. Gilles, C. Peifer, and A. Kluge (2023). Defining human–ai teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence 6*.

Boring, R. (2007, 06). Dynamic human reliability analysis: Benefits and challenges of simulating human performance. Volume 2.

Defense Advanced Research Projects Agency (DARPA) (2024). Autonomy standards and ideals with military operational values (asimov). `https://www.darpa.mil/research/programs`. Accessed: 2026-01-29.

Det Norske Veritas group (DNV) (2024). Offshore and onshore reliability data (oreda). `https://www.dnv.com/`. Accessed: 2026-02-06.

D.I. Gertman, H.S. Blackman, J. M. J. B. C. S. (2005, Update 2021). The spar-h human reliability analysis method" (nureg/cr-6883). *U.S. Nuclear Regulatory Commission*.

Dörner, D. (1990, 05). The logic of failure. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences 327*, 463–73.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors 37*(1), 32–64.

Groth, K. M. and A. Mosleh (2012). A data-informed pif hierarchy for model-based human reliability analysis. *Reliability Engineering System Safety 108*, 154–174.

Groth, K. M. and L. P. Swiler (2013). Bridging the gap between hra research and hra practice: A

bayesian network version of spar-h. *Reliability Engineering System Safety 115*, 33–42.

He, C., A. Bejaoui, and D. Söffker (2022, 09). Situated and personalized monitoring of human operators during complex situations. *European Conference on Safety and Reliability (ESREL)*.

Kirwan, B. (1994). A guide to practical human reliability assessment (1st ed.). *CRC Press*.

Naval Sea Systems Command (1991). *MIL-HDBK-217F: Military Handbook – Reliability Prediction of Electronic Equipment*. U.S. Department of Defense. Notice 2.

Reason, J. (1990). Human error. *Cambridge University Press*.

Shyshova, O., A. K. Mayouche, M. Seesunkur, and D. Söffker (2025). Performing best: Organizing suitable human-ai teaming for improved performance: an experimental study on human recognition processes under time pressure. *IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*.

Söffker, D. (2004a). Understanding mmi from a system-theoretic point of view — part i: Formalization of mmi introducing a modified situation–operator concept. In *Proceedings of the 9th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human–Machine Systems*, Atlanta, GA, USA.

Söffker, D. (2004b). Understanding mmi from a system-theoretic point of view — part ii: Concepts for supervision of human and machine. In *Proceedings of the 9th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human–Machine Systems*, Atlanta, GA, USA.

Söffker, D. (2008). Interaction of intelligent and autonomous systems - part i: Qualitative structuring of interactions. *MCMDS-Mathematical and Computer Modelling of Dynamical Systems 14*(4), 303–318.

Söffker, D. and O. Shyshova (2025). Reliability or ethics: Why should the human decision be initial or final? *Proceedings of the 35th European Safety and Reliability Conference (ESREL) & 33rd Society for Risk Analysis Europe Conference*.

Vaccaro, M., A. Almaatouq, and T. Malone (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour 8*, 2293–2303.

Xing, J. e. a. (2021). Idheas-data: A human reliability analysis data collection tool and database (nureg-2256). *U.S. Nuclear Regulatory Commission*.