*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

doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P2453-cd

# Reliability or ethics: Why should the human decision be initial or final?

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*

Decision support systems or, more recently, human-AI teaming systems vary in one way or another: Who makes the final decision when it comes to important matters or life or death? For fundamental legal, safety, moral, or human persuasion reasons, the final decision in cooperative systems is practically often assigned to humans. The article addresses the complexity of the discussion and discusses the pros and cons of a technological and of a human final decision. Especially in contexts where human behavior is known to be unreliable, the question arises whether it is appropriate to subordinate more reliable decision-makers to human decision-making. The question also arises as to how AI-based assistance and suggestion systems influence the quality and quantity of the decision favorably or unfavorably. The first results of a study on this question are presented, which replaces the ethical and moral discussion with a reliability-oriented one. This is particularly important as our everyday lives are influenced in the same way by the same issues and we have already adapted to technological solutions or, on the contrary, no longer make final decisions on our own because we also trust technological solutions.

*Keywords*: Human-AI teaming, decision-making, assistance systems, human machine systems, ethics vs. reliability, study.

## 1. Introduction

### 1.1. *Motivation and Research Questions*

In recent years, the integration of decision-making systems, particularly those utilizing Artificial Intelligence (AI), has significantly increased in safety-critical applications. In medicine, autonomous driving, or disaster management, these systems' ability to make real-time, precise, and reliable decisions can make the difference between life and death. While technological progress continues, questions surrounding system reliability and ethical accountability become central, especially in safety-critical domains. A key point of discussion is the relation between the technical reliability of AI systems and the ethical standards governing their decisions. Human decision-making often relies on intuition, experience, and moral values, whereas AI systems operate primarily based on data-driven logic, which may lack transparency or comprehensibility. This raises critical questions: Who is responsible for erroneous decisions? How can ethical principles be integrated into automated systems? Can algorithms replace or complement human intuition in critical situations? How should a final erroneous decision be legally evaluated? The answers to these questions are not merely theoretical but have profound implications for the trust placed in these technologies by both operators and society. A comprehensive understanding of reliability and ethics in decision-making systems is therefore essential for responsibly implementing technological innovations in safety-critical fields. The growing intersection of human and machine decision-making requires a deeper discussion tailored to specific scenarios.

While AI systems have the potential to optimize decision-making processes in life-critical scenarios, human involvement becomes a central factor in their development, implementation, and oversight. Studies indicate that human errors are increasing in safety-critical contexts, and this trend extends to AI-supported decision systems. The errors described below can occur at multiple levels and have significant consequences for system reliability and safety:

1) Programming Errors: Algorithms, as the core of

AI systems, are only as reliable as their programming. Faulty logic, inadequate consideration of edge cases, or poorly designed safety mechanisms can cause failures in critical situations.
2) Data Selection and Quality: Biases, incomplete datasets, or flawed training data can result in erroneous patterns being learned by AI systems, compromising decision accuracy. Human judgment is critical in data selection and pre-processing.
3) Validation of the results: Validating AI decisions requires a deep understanding of underlying models and their limitations. Misinterpreting or uncritically accepting AI outputs can result in technical errors propagating through the system.

Errors can occur at every stage of the life cycle of an AI system, from conceptualization and design to deployment and real-world application. Identifying these vulnerabilities and establishing targeted error prevention measures are crucial to building trust in AI-supported decision systems and ensuring their safety in life-critical scenarios.

Over the past three decades, the understanding of human influence on technical systems, including computer-based and AI-driven technologies, has evolved significantly. Early approaches often reduced human factors to simple operator errors, but with the development of Human Reliability Assessment (HRA) methods, a more precise perspective has been established as described here:
First-Generation HRA: Focused on identifying and quantifying operator errors in well-defined and predictable situations, using mechanistic models to analyze human failures.
Second-Generation HRA: Expanded to consider cognitive processes, recognizing that human decision-making often involves complex perception, interpretation, and planning/reasoning. This generation also integrates organizational factors into the analysis.
Third-Generation HRA: Addressed dynamic interactions between humans, machines, and their environments, analyzing individualized and situational error conditions related to stress, fatigue, and cognitive overload. These advances underscore that humans are not just potential sources of error, but crucial actors in ensuring successful interaction with technical systems. This knowledge is fundamental for designing human-AI interfaces that are not only functional but also reliable and safe in real-world applications.

Situational awareness and subsequent decision-making processes become particularly challenging in complex dynamic scenarios. These situations require the simultaneous evaluation of numerous factors under significant time pressure, often in the seconds range. Under such constraints, the time available for information gathering and processing is drastically reduced, forcing reliance on heuristics, experience, or automated reaction patterns. Competing or interdependent objectives further complicate situational assessments. The quality and timeliness of available information, as well as the ability to distinguish between relevant and irrelevant data, become decisive factors. Errors in such scenarios often arise not from a lack of knowledge but from the inability to process all relevant information within the given time frame.

Examples of time-critical decision-making challenges in complex situations are in the fields Maritime Navigation in dense Waterways (captains must make rapid course corrections to avoid collisions while accounting for changing currents and vessel positions),aviation (pilots must react within seconds to technical malfunctions or unexpected obstacles), emergency medical services (doctors must make life-saving decisions under incomplete information and time constraints), autonomous vehicles (systems must calculate optimal responses to sudden obstacles or unpredictable driver behaviors in milliseconds), or disaster response management (emergency teams must prioritize actions to maximize resource efficiency and minimize damage.). In these examples, situational awareness and decision making under time pressure require seamless human-AI collaboration. Effective human-machine teaming is essential to address reliable and efficient time-critical challenges.

Human-AI Teaming (HAT) is an interdisciplinary field focused on optimizing collaboration between humans and AI systems. The goal is to generate effective partnerships where human creativity, contextual understanding, and ethical reasoning complement AI's computational power and pattern recognition, aligning with Fitt's List (Fitts (Ed.), 1951). Key HAT research emphasizes trust, transparency, and explainability. Methods include cognitive modeling, user-centered design, real-time adaptive interfaces, and explainable AI (XAI). Techniques like eye tracking and real-time feedback loops help to understand human cognitive states during interaction with AI systems. Simulation environments often serve as testbeds for system refinement. Significant progress has been made in trust calibration, where AI systems predict when human operators might lose confidence in automation. This has shown improvements in aviation, healthcare, and defense, where AI assists in rapid decision-making. DARPA (Defense Advanced Research Projects Agency) has driven HAT research since the early 2000s through programs like "Explainable Artificial Intelligence (XAI)" and "AI Next Campaign," focusing on integrating AI into human workflows, especially in defense contexts where human-AI interaction can determine mission success. The growing reliance on AI in defense emphasizes maintaining human control while leveraging AI capabilities for reliability and ethical responsibility, especially in dynamic environments. In weapon systems, particularly autonomous weapons (LAWS – Lethal Autonomous Weapon Systems), the principle of "meaningful human control" ensures human decision-making remains central, with accountability rooted in international humanitarian law (IHL). However, the increasing capabilities of AI, processing data faster than humans, raise ethical and legal concerns. Proponents argue that autonomous systems could reduce human error and bias, while critics highlight issues of accountability, unintended consequences, and the inability of AI to fully grasp complex ethical contexts. This dilemma extends beyond military applications to autonomous vehicles, medical AI systems, and industrial safety mechanisms, where the final responsibility, especially for life-critical decisions, must remain with humans. Although AI can assist and optimize decisions, the human desire for control and reassurance remains central. This balance between automation and oversight requires transparent design, clear accountability structures, and a focus on collaboration rather than delegation.

What reasons are given as to why final decisions must be finally performed from humans? On the one hand, it is the so-called moral responsibility that is only attributed to humans. According to the current view, only humans can include moral and ethical considerations in their decisions, such as weighing up the consequences of their actions. According to the current view, machines cannot experience and evaluate contradictions of ethical nature. In war contexts, for example, the Geneva Conventions require that responsibility for military decisions can be assigned to a human actor. Accordingly, a machine could not be held responsible. The error-proneness of algorithms and the assumed lack of flexibility of programmed mechanisms are also stated as a disadvantage compared to human capabilities. Humans would also like to retain control over decisions that determine life and death, as this is currently considered to be a deeply human and non-delegable responsibility, meaning that the corresponding control should also be assigned to humans. The (supposed) control associated with the final decision also provides psychological security, as humans are currently more likely to trust their own judgment and values than an automaton. This is why transparent automation concepts rely on collaboration rather than complete delegation, so that AI provides support but the ultimate responsibility remains with the human. From this point of view, the current discussion represents a further development of the supervisory control strategy according to Sheridan (Sheridan and Johannsen, 1976), which has been known since the 1970s. Another aspect to be mentioned is the freedom of decision that can be attributed to humans as individuals: Humans want to retain control over vital decisions or assign them to humans because the associated responsibility is inextricably linked to freedom of decision, i.e. the ability to make conscious decisions and to assign accountability on the basis of awareness and the freedom to decide. Accordingly, decisions made autonomously by machines cannot be assigned responsibility and therefore cannot be held accountable. Human supervision or final decision therefore ensures that ethical aspects, situational judgment and responsibility in the sense of accountability are retained and are not dehumanized, i.e. assigned to something other than humans.

### 1.2. *Background: The role of humans vs. machines in Decision-Making*

In (Söffker and Weber, 2006) two authors (Prof. Weber, a social philosopher and media scientist, and Prof. Söffker, an engineer specializing in automation and human-machine systems) published a debate about the role of humans and machines. Söffker emphasizes a balanced, hierarchical relationship in safety-critical systems where machines support with data insights but humans retain final authority, particularly in ethically sensitive decisions. Machines should provide insights and automate routine tasks but never replace human decision-making authority, aligning with human-centered automation principles. Söffker also highlights the importance of clear interaction protocols to prevent 'automation complacency,' advocating for collaborative systems where machines enhance, not undermine, human decision-making capabilities.

An opposing view, driven by the rise of AI and automation, suggests that AI can surpass human decision-making in critical areas by processing vast data sets quickly and consistently (New York Post, 2024). Proponents argue that minimizing human oversight reduces delays and emotional bias. However, critics (German Ethics Council, 2023) warn that overconfidence in AI may overlook ethical and contextual challenges AI systems still cannot fully address.

The current debate on human-AI teaming centers on decision-making authority. While AI performs in data analysis and automation, the question remains whether human oversight or AI autonomy leads to the most reliable outcomes, particularly in safety-critical contexts. Human-AI collaboration aims to merge human intuition and ethical reasoning with AI's speed and consistency, but the balance between control and automation requires ongoing research. Research on Human-AI Teams (HAT) explores cooperation between humans and autonomous systems to effectively manage complex tasks (O'Neill et al., 2020). Humans contribute to creativity and ethical reasoning, while AI handles data processing and precision decision-making, such as intelligent assistance systems in aviation. Ongoing studies focus on trust, transparency, and ethical considerations (Vössing et al., 2022; Shneiderman, 2020a; Wickramasinghe et al., 2020). Trust is highly context-dependent, with studies (Fahnenstich et al., 2024) showing higher trust in AI for high-risk decisions compared to low-risk situations.

Guideline standardization and design principles for trustworthy AI interfaces (Xu and Gao, 2024) further emphasize transparency and user adaptability. Explainable AI (XAI) approaches enhance clarity in decision processes (Chamola et al., 2023), ensuring effective use in critical environments (Endsley, 2023). Decision reliability also plays a central role. The DECREE (Decision Reliability Evaluation) model (Pyy, 2000) offers a framework for evaluating decision quality in human-AI collaborations. Cognitive biases like simplification and verification biases (Watts et al., 2020) can impair human judgment, emphasizing the need for balanced decision-making frameworks. The 'automation conundrum' describes the risk of excessive system autonomy reducing human intervention capabilities in emergencies (Endsley, 2016; Shneiderman, 2020b). Though high autonomy can reduce cognitive load, it may also limit situational awareness and control in critical situations.

The current discussion on Human-AI Teaming converges on the critical question of decision making. Who (and possibly under what conditions or in what situation) makes the right decision - humans or AI or both together? While advancements in AI technologies have gained optimism about their capabilities in real-time analysis, pattern recognition, and autonomous action, the actual debate in critical fields remains unresolved or is decided by a third, higher authority. Human-AI collaboration seeks to combine human intuition and ethical reasoning with AI's computational power and consistency. However, in safety-relevant or critical scenarios, the fundamental question of whether human oversight or AI autonomy leads to the most reliable outcome remains practically unanswered. This ambiguity underscores the need for ongoing interdisciplinary research and careful system design.

To restart the discussion, it helps to focus on the core goal of decision-support systems: ensuring clear and correct decisions. Technically, this means maximizing reliability while recognizing that all decisions, even automated ones, originate from human thinking. Every machine-made decision goes back to human-defined rules. The following discussion will explore how machine-based assistance compensates, reinforces, anticipates, or improves human decision-making.

## 2. Opportunities and Conflicts in Human-AI Teaming for Decision Support

### 2.1. *Comparison of humans and machines/algorithms*

If strengths and weaknesses in Human and AI Decision-Making are summarized, it can be concluded, that purely technical decisions outperform the human counterpart in consistency, data-driven accuracy, and the ability to process vast amounts of information rapidly. Decisions remain unaffected by fatigue or emotional stress. However, pure technical decisions by definition lack moral judgment, contextual awareness, and often struggle with novel or unexpected scenarios so show missing situated flexibility, especially if these were not covered in the related training data (Michael Pflanzer, 2023). Purely human decisions show exactly the opposite here and there: contextual awareness, moral reasoning, and accountability (Jan B. Schmutz, 2024; Michael Pflanzer, 2023). Human's flexibity allow to adapt to unforeseen events and make (theoretically) value-based decisions, but they are sensitive to fatigue, emotional bias, and inconsistencies also affected by time pressure.

Robust Human-AI Teaming should balance weaknesses and maintain strengths (O'Neill et al., 2020). Both humans and AI are prone to bias — humans through judgment, AI through biased training data or flawed algorithms, for which humans remain responsible (Michael Pflanzer, 2023). Limited AI transparency and human cognitive overload further impact decisions. The goal is to combine strengths (O'Neill et al., 2020) while preventing any weakness from becoming critical.

### 2.2. *Preliminary conclusion for a suitable Human-AI Teaming approach (HAT)*

As previously detailed the (actual) unbeatable strength of AI is to handle data-intensive tasks with high precision, while humans provide ethical oversight and situated judgment (based on actual design and programming philosophies). Real-time collaboration allows AI to handle repetitive, structured tasks and alert humans to anomalies, while humans validate and adjust decisions in critical moments. Based on the above theoretical discussion, the known practice (cf. chapter 3.2) should therefore be justifiable: human confirmation of machine preselected decisions so that humans always make the final decision, and machines therefore play the role of a (reliable preparation) tools.

However, some weaknesses remain: designers over-reliance on AI systems can reduce human vigilance, and human cognitive limitations may hinder real-time oversight of complex AI outputs. Despite these challenges, an optimized Human-AI system can/may generate a balanced synergy where AI's analytical power complements human intuition and accountability, leading to more robust, ethical (in the sense of humans last), and reliable decision-making outcome, theoretically. Another (open) relevant aspect will be how time pressure effects the interaction.

## 3. Findings from a Reliability Study

### 3.1. *Study Design and Methodology*

A study by (Shyshova et al., 2025) examined the reliability of safety-critical decisions under different decision modes: human-only, AI-only, and hybrid human-AI teaming (HAT). The study focused on object recognition in road traffic under time pressure. Two conditions were tested: human-only decisions and HAT. In HAT, two different sub-conditions were used: AI-First, where AI pre-marked objects for human validation, and Human-First, where humans judged first, followed by AI analysis, prompting reconsideration in case of discrepancies.

In Table 1 the correlations between human influence on different development phases (*1-*4) of decision making (months: conception (*1) / training (*2) to seconds: situative decision making (*4) are shown. Based on human influence on the final decision (*4) as a criterion of a positive morality of the associated decision making, only the pure human decision itself (H-Only) and the assisted decision with human final decision (AI-First) can be evaluated as positive (marked in green), the exclusively AI-based decision can accordingly only be evaluated negatively (marked in red). However, if not, only the final decision is considered and, therefore, the entire process chain is integrated. In table 1 it becomes clear that, on the contrary, the purely AI-based decision is also subject to a very clear influence of human design (* 1 to *4), which is not even subject to the human characteristic of fickleness or inconsistency as a program principle or algorithmically defined for the formulation of decisions; of course, the same applies to the formation of the H-AI teamings of interaction in different but similar ways.

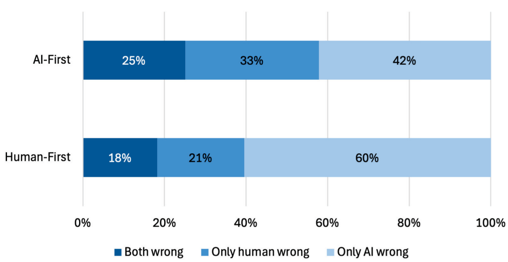


Fig. 1.: Distribution of wrong recognitions

### 3.2. *Results and Discussion on Consequences*

The evaluation of the results (cf. Tab. 3) shows that (beside the outperforming AI-Only results) the Human-only approach has the best results at median time limits of 6 and 8 seconds. In these time spans, the probability of human error (HEP) reached an average of 0.15, indicating high consistency and reliability under moderate time pressure. At very short time limits of 5 seconds, performance deteriorated slightly (HEP=0.20), while it decreased significantly at longer decision times (10 seconds), where the mean increased to 0.38, reflecting increasing inefficiency with longer decision times. This suggests that humans perform most reliably under moderate time pressure, while both extreme time constraints and too much time can lead to an increased error rate.

The AI system used is based on a YOLO-based neural network that was trained on a data set of 15,000 annotated images from the Udacity Self-Driving Car Dataset. The AI model achieved an average precision of 88.7 % across all classes at a threshold of 0.5 IoU (mAP@50), resulting in a probability of error of 0.113 or 11.3 % (1-mAP@50), so outperforms all other cases.

The AI-first workflow achieves stable results under time pressure, as the human is mainly used for validation, while the human-first workflow offers more flexibility but is less reliable. Hybrid human-AI teams (HAT) do not consistently produce better results than humans or AI alone, which could be due to additional coordination efforts and potential misunderstandings. It remains unclear whether the success of AI-First is due to humans being less critical of AI suggestions or more inclined to agree with existing recommendations. This result aligns with human-AI collaboration studies, typically on more complex interactions (Vaccaro et al., 2024).

In addition to the above ethical discussion, further contradictions arise in combination with the above results (last row of table 3), as the reliability of the overall result for the specific context gives a clear result.

## 4. Conclusions and Outlook

The comparison of road traffic object recognition shows that the AI-based approach achieves a significantly better recognition rate (lower error rate) than human-involved decisions and AI-assisted recognition. The purely human decision shows the best values, but the purely technical decision is superior in all cases, revealing a critical new problem.

This contribution evaluates assistance situations

Table 1.: Comparison of HMS or H-AI Teaming Approaches

| **HMS or H-AI Teaming Vs Human effect on the final decision** | **H-Only** | **H-First** | **AI-First** | **AI-Only** |
|---|---|---|---|---|
| Conception of decision criteria and decision-making (*1) | Present | Present | Present | Strong-Very strong |
| Data analysis: Training/Learning (*2) | Present | Present | Present | Strong |
| Finetuning of the Algorithm (*3) | Present | Present | Present | Present |
| Situative Decision Making (*4) | Very strong | Strong | Present | Not present |
| Reliability-based evaluation (EP) | 22.1 % | 19.8 % | 30.1 % | 11.3 % |

Judgments/relations: Not present, present, strong, very strong
Morally ethically relevant assessment: Positive, neutral/no judgment, negative
Reliability-based result: top, ok, flop.

Table 2.: Comparison of Human-First and AI-First corrctions

| | **Final result** | **Human-First** | | **AI-First** | |
|---|---|---|---|---|---|
| | | Count | % | Count | % |
| Initially matching results | Correct | 1666 | 49 % | 2213 | 65 % |
| | Incorrect | 314 | 9 % | 295 | 9 % |
| Correction by human | Correct | 468 | 14 % | 496 | 15 % |
| | Incorrect | 199 | 6 % | 386 | 11 % |
| Correction by AI | Correct | 169 | 5 % | - | - |
| | Incorrect | 574 | 17 % | - | - |
| **TOTAL** | | **3390** | **100 %** | **3390** | **100 %** |

Table 3.: Error probability results (mean values) for different time limits, with HO: 'Human-Only', AIF:'AI-first'; HF:'Human-first'; AIO:'AI-Only' (Shyshova et al., 2025)

| **Time limit** | **HO** | **AIF** | **HF** | **AIO** |
|---|---|---|---|---|
| 5 s | 0.20 | - | - | 0.11 |
| 6 s | 0.15 | 0.22 | 0.32 | 0.11 |
| 8 s | 0.15 | 0.19 | 0.28 | 0.11 |
| 10 s | 0.38 | 0.18 | 0.30 | 0.11 |

in AI-supported autonomy, emphasizing also life-critical or safety-relevant systems. The highest moral standards apply, especially in weapon systems, where fundamental decisions are predetermined. Here, international humanitarian law (IHL) and in particular the Geneva Conventions of 1949 and their additional protocols of 1977 form the central legal basis for the use of force in armed conflicts. Even in these cases, the final decision lies with humans, because (only) they can be held responsible; furthermore, machines/algorithms/programs (unlike humans) cannot - according to the actual understanding - make ethical considerations; this applies in particular to dilemma situations that do not have an algorithmic solution (as an argument). Therefore - and this is the current legal situation - a human decision-making authority is indispensable at the end of a fatal decision-making process. This article shows that, while maintaining the moral and ethical compass, a new conflict arises between the morally justifiable decision to make the final human decision and the best technical solution. At the same time, it is shown that there is indeed a strong human influence on the algorithm, although it remains to be researched how this is to be assessed and whether it is possible to map ethical principles algorithmically.

## References

Chamola, V., V. Hassija, A. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar (2023). A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, 1–1.

Endsley, M. (2016). From here to autonomy: Lessons learned from human–automation research. *Human Factors 59*, 001872081668135.

Endsley, M. R. (2023). Supporting Human-AI Teams:Transparency, explainability, and situation awareness. *Computers in Human Behavior 140*, 107574.

Fahnenstich, H., T. Rieger, and E. Roesler (2024). Trusting under risk – comparing human to ai decision support agents. *Computers in Human Behavior 153*, 108107.

Fitts (Ed.), P. M. (1951). Human engineering for an effective air-navigation and traffic-control system. *Washington, DC: National Research Council*.

German Ethics Council (2023). Human and Machine – Challenges of Artificial Intelligence.

Jan B. Schmutz, Neal Outland, S. K. E. G. A.-S. U. (2024). Ai-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*.

Michael Pflanzer, Zachary Traylor, J. B. L. V. D. C. S. N. (2023). Ethics in human–ai teaming: Principles and perspectives. *AI and Ethics*.

New York Post (2024). Five reasons the future of AI has already arrived. Accessed: 2025-01-13.

O'Neill, T. A., N. J. McNeese, A. Barron, and B. Schelble (2020). Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors The Journal of the Human Factors and Ergonomics Society 64*, 1–35.

Pyy, P. (2000). An approach for assessing human decision reliability. *Reliability Engineering and System Safety 68*(1), 17–28.

Sheridan, T. B. and G. Johannsen (1976). *Monitoring Behavior and Supervisory Control*. New York: Plenum Press.

Shneiderman, B. (2020a). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Trans. Interact. Intell. Syst. 10*(4).

Shneiderman, B. (2020b). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction 36*(6), 495–504.

Shyshova, O., A. Kamdoum 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. In *Proceedings of the IEEE International Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, submitted.

Söffker, D. and J. Weber (2006). On designing machines and technologies in the 21st century. an interdisciplinary dialogue. *International Review of Information Ethics 6*.

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.

Vössing, M., N. Kühl, M. Lind, and G. Satzger (2022). Designing transparency for effective human-ai collaboration. *Information Systems Frontiers 24*.

Watts, L. L., K. E. Medeiros, T. J. McIntosh, and T. J. Mulhearn (2020). Decision biases in the context of ethics: Initial scale development and validation. *Personality and Individual Differences 153*, 109609.

Wickramasinghe, C. S., D. L. Marino, J. Grandio, and M. Manic (2020). Trustworthy AI development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*, pp. 130–136. IEEE.

Xu, W. and Z. Gao (2024). Applying HCAI in Developing Effective Human-AI Teaming: A Perspective from Human-AI Joint Cognitive Systems. *Interactions 31*(1), 32–37.