From Exploitation to Augmentation: Navigating AI's Impact on Human Value- and Inference-based Decision-Making

Esther S. Kox[1]
Human-Machine Teaming, TNO, Soesterberg
Psychology of Conflict, Risk & Safety, University of Twente, Enschede
THE NETHERLANDS
esther.kox@tno.nl

M. Birna van Riemsdijk[2]
Human Media Interaction, University of Twente, Enschede
THE NETHERLANDS
m.b.vanriemsdijk@utwente.nl

José H. Kerstholt[3]
Human Behaviour and Collaboration, TNO, Soesterberg
Psychology of Conflict, Risk & Safety, University of Twente, Enschede
THE NETHERLANDS
jose.kerstholt@tno.nl

ABSTRACT

Multi-Domain Operations (MDO) drastically increase the complexity of military decision-making. Artificial Intelligence (AI) has the potential to augment decision-making processes and assist human decision-makers in dealing with this growing complexity. However, AI also poses a risk, as influential algorithms can be exploited by adversaries to influence perceptions, beliefs and behaviours and to manipulate public opinion to gain strategic advantages (e.g., in cognitive warfare). To understand AI's effects on human decision-making, we distinguish between two types of decision tasks: inference-based decisions (focused on situational understanding) and value-based decisions (focused on choosing a course of action). In the present paper, we argue that AI has differential effects on both types of decisions, augmenting and exploiting them. AI can enhance inference-based decisions by its computational power, but it can also manipulate human judgments by facilitating the spread of misinformation. For value-based decisions, AI can help elicit preferences and calculate optimal options to achieve military goals. However, AI-driven profile-based, micro-targeting techniques can be used to manipulate people into choices that are inconsistent with their true values. We aim to offer insights into how to best navigate these influences to mitigate the risks and to maintain meaningful human control.

[1] Esther Kox (0000-0002-2512-5665) - ORCID

[2] Birna van Riemsdijk (0000-0001-9089-5271) - ORCID

[3] Jose Kerstholt (0000-0002-5421-3090) - ORCID

1.0 INTRODUCTION

The concept of Multi-Domain Operations (MDO) has emerged as a strategic response to the evolving complexities of modern warfare (NATO, 2023). MDO involves the integration and coordination of military activities across the five domains sea, land, air, cyber and space and therefore necessitates sophisticated orchestration. This coordination aims to synthesize information from multiple sources, creating a comprehensive understanding of the battlefield that encompasses all military and non-military activities for both friendly and enemy forces. Furthermore, MDO expands the scope of military effects beyond the traditional physical dimension. It now incorporates both virtual and cognitive dimensions, reflecting the multifaceted nature of modern conflicts. This expansion recognizes that warfare in the 21st century is not confined to physical battlegrounds but extends into cyberspace and the realm of human cognition. MDO addresses several key challenges such as the limitations of conventional warfare in dealing with hybrid threats, the resurgence of near-peer adversaries and the rapid pace of technological advancements (NATO, 2023).

The introduction of MDO has placed significant demands on military personnel, pushing their cognitive capabilities to the limits. The sheer volume of data available in modern military operations will overwhelm human cognitive capabilities, not to mention coordination activities needed for synchronization across domains and effects. To address these challenges and to maintain a competitive edge, military decision-making should be supported by advanced technologies, with Artificial Intelligence (AI) being the most promising option. In its emerging role, AI is increasingly capable of making autonomous decisions, moving beyond its traditional function as a decision support system for human operators. This advancement is driven by improvements in machine learning algorithms, increased computational power, and the ability to process and analyze vast amounts of data in real-time.

Advancements in AI bring both opportunities and challenges. On one hand, it can significantly enhance operational efficiency, reduce human error, and enable faster responses to rapidly changing situations. On the other hand, it raises important ethical considerations and questions, especially in high-stakes environments. The central issue revolves around the concept of "meaningful human control" – how can we ensure that humans maintain appropriate oversight and ultimate authority over these increasingly autonomous systems? (Cavalcante Siebert et al., 2023).

A huge challenge to the concept of meaningful human control is cognitive warfare (CogWar). In this evolving landscape, AI is not used to enhance decision-making capabilities for military operators but is instead weaponized by adversaries to deceive and manipulate, thereby defeating the purpose of meaningful human control. CogWar refers to the strategic use of AI and other technologies to influence human cognition, perceptions, and decision-making processes (Masakowski, Y. R. & Blatny, 2023). In this context, AI systems can be deployed to generate and disseminate misinformation, create convincing deep fakes, or manipulate data streams in real-time. The goal is to sow confusion, erode trust, and ultimately compromise the decision-making capabilities of the opposing force. In such a complex and rapidly evolving environment, the notion of meaningful human control becomes increasingly vague and challenging to maintain.

To outline the effects of AI on human decision making and the consequences for meaningful human control, we distinguish two types of decision tasks that are differentially affected by AI: inference-based decisions (or judgments) and value-based decisions (or choice). Inference-based decisions typically require the combination of cues to form a judgment on a particular situation, such as detecting a military threat or predicting an enemy course of action (Doswell, 2004). Choices, on the other hand, concern the explicit trade-off of the consequences of a range of options, such as the choice between a safer or faster route, or between mission effectiveness and risk of collateral damage (Babushkina & Votsis, 2022). In addition, we make a distinction between AI that is used to support military decision makers (augmentation) and AI that

is used to attain strategic advantage through influencing a target audience (exploitation). This last category concurs with the concept of CogWar and may be directed towards the military or civilians. In the present paper, we argue that AI has differential effects on both types of decisions, both through enhancing and exploiting them, and consequentially for meaningful human control (Table 1).

Table 1: Examples of phenomena that augment or exploit different types of decision tasks

	Augmentation	Exploitation
Inference-based decisions (judgment)	**Pattern recognition; situation awareness**	**Spread and creation of mis- and disinformation with next-gen AI**
Value-based decision-making (choice)	**Decision and behaviour (change) support**	**AI-driven profile-based, micro-targeting techniques (hyper-nudging)**

2.0 INFERENCE-BASED DECISIONS (JUDGMENT)

Assessing a situation requires knowledge of which cues are (most) predictive for the situation at hand. Through practice and feedback this knowledge is built up over time, mostly resulting in automized, intuitive judgments. Particularly the naturalistic decision making (NDM) approach has emphasized the role of prior knowledge in how people make decisions in real-world settings (Klein, 2008). From their perspective on intuition and expertise, prior experiences enable people to quickly categorize situations to make effective decisions (Klein, 2008). While judgments can inform a decision, they ultimately require a specific decision rule, for example ‘if a target cannot be positively identified as hostile, do not engage’. In the present paper, judgments only concern the assessment, in this example hostility, and not the decision rule.

Today’s advanced computers can base judgments on even significantly larger amounts of information using pattern recognition (De Cremer & Kasparov, 2021). Thanks to advances in computing power, speed, performance and algorithms, we have computer systems that perform better than humans in certain tasks, without mimicking human reasoning. This has been illustrated by computer systems like Deep Blue and AlphaGo that have defeated human champions in the boardgames Chess and Go respectively. They were initially trained to mimic human play, but ultimately learned through self-play, and displayed moves unlike the usual human style of play (Campbell et al., 2002; J. X. Chen, 2016). While AI excels at identifying patterns in information, it is important to note that these examples involve well-defined and static decision tasks, which is a far cry from the dynamic and complex challenges faced in real military operations.

2.1 Augmentation

In the context of Multi-Domain Operations (MDO), AI offers significant advantages in enhancing information processing and synchronizing military and non-military capabilities (C. E. Lee et al., 2023). AI systems can efficiently select and integrate information across various domains and dimensions, allowing for rapid analysis of vast datasets from multiple sources. This capability enables the identification of patterns and correlations that might elude human analysts, as well as real-time synthesis of information to support both tactical and strategic decision-making.

For humans it might be difficult, however, to understand the reasoning behind AI-generated outcomes. As noted by Davidovic (2023), understanding these processes may even be considered a paradox as the AI was introduced to augment humans *beyond* their natural information processing limits (Davidovic, 2023). A well-known result of the inability to fully comprehend the AI's reasoning is the automation bias, the tendency of humans to readily and unreflectively accept advice from automation, particularly in situations where the AI's conclusions conflict with human judgment (Krügel et al., 2022; Wright et al., 2016). Accepting AI-generated solutions without sufficient critical evaluation potentially leads to over-reliance on machine intelligence.

The need for meaningful human control in judgment tasks highly depends on the trustworthiness of the AI. AI that is 100% correct or produces outcomes that are significantly better than those of humans can be trusted, and consequently require less human control. The complexity of many operational environments often introduces uncertainty, however, that affects the reliability of an AI agent's predictions (Kox et al., 2022).This uncertainty complicates the assessment of when it is appropriate to trust and rely on an AI system. As AI systems become more sophisticated and are deployed in increasingly complex scenarios, the challenge of calibrating trust and maintaining meaningful human control becomes more pronounced (de Visser et al., 2019; J. D. Lee & See, 2004). Humans need to trust AI enough to rely on its outcomes for the collaboration to be profitable, while they should also be aware of the limitations of the AI for the collaboration to be safe (Parasuraman & Riley, 1997).

For calibrated trust, humans need to construct an appropriate mental model of how the AI works (Jermutus et al., 2022) and form a cognitive representation of the machine's capabilities and limitations (Matthews et al., 2018). However, these representations of reality are personal and often, if not always, incomplete and inconsistent (Jones et al., 2011). To bolster meaningful human control over AI systems, several design principles are needed to make AI systems more transparent, understandable, and manageable for human operators. The key principles include Observability (ensuring that the AI system's actions and states are visible and can be monitored by human operators), Predictability (making the AI system's behaviour consistent and foreseeable), Explainability (making the AI system's decision-making process transparent and understandable to human users) and Directability (ensuring that human operators can guide, intervene, or override the AI system when necessary) (Johnson et al., 2014; Miller, 2017, 2022).

2.2 Exploitation

Although AI can augment human judgement, it also has the potential to distort it in two main ways. First, AI can distort human judgement by personalizing what people see and read online, potentially creating worldviews that no longer corresponds to reality. Given the overwhelming amount of available information online, algorithms have been designed to tailor the virtual choice environment to a user's online behaviour (e.g., personalized feeds or search results) (Morozovaite, 2021; Yeung, 2017). Consequently, users are increasingly presented with information that confirms and reinforces their current beliefs and preferences. Such "news personalization" and the creation of information bubbles or echo chambers can have major psychological consequences on different levels of magnitude (Reviglio & Agosti, 2020), such as reducing opportunities to self-determination, and limiting critical thinking, dialogue, creativity and solidarity by being less exposed to alternative points of view (Reviglio & Agosti, 2020), potentially increasing the polarization of opinions.

Second, in addition to such unintended effects, AI is also used to deliberately influence human judgments through enabling the creation of mis- and disinformation on the one hand and by facilitating the spread of it on the other. Disinformation, "intentionally false or deceptive communication tactics that actors use to advance their political or economic aims" (Tenove, 2020), is used to manipulate the public's perception and decision-making (Hung & Hung, 2022). Generative AI enables technologies like deepfakes (i.e., the creation

of highly realistic fake images, audios or videos that can alter the face or voice of an individual into someone else's (Mahmud & Sharmin, 2021)) which distort people's perceptions of reality. These AI generated media are difficult, if not impossible, to distinguish from real and can be used to create and proliferate fake news narratives, fraud and hoaxes (Mahmud & Sharmin, 2021). Opponents can use these techniques to, for example, sow distrust in authorities or create polarization and fragmentation (Bernal et al., 2020; Claverie et al., 2021; Kox et al., 2023). Malevolent actors can benefit from the social unrest and division that results from a deterioration of trust and the spread of distrust within society and exploit it for political or military gain.

Traditional understanding of human control in military operations assumes a clear distinction between reliable information and enemy deception. However, in the age of AI-driven CogWar, this distinction becomes blurred, making it difficult for human operators to discern reality from fabrication. To address this issue and enhance meaningful human control over AI-assisted decision processes, there is a growing recognition of the need to improve AI literacy. Enhancing AI literacy could contribute significantly to increasing meaningful human control in AI-assisted decision-making contexts as it empowers users to engage more critically with AI systems, understand their underlying mechanisms, and make more conscious choices about when and how to incorporate AI recommendations into their decision processes.

3.0 VALUE-BASED DECISIONS (CHOICE)

In contrast to judgments, choices require trade-offs between options that are differentially favourable on attributes. One course of action is fast but dangerous, the other one involves less danger but will cost significantly more time and effort. The distinction between AI and human intelligence is presumably most pronounced for this class of decision tasks as these tasks involve an affective evaluation (Busemeyer et al., 2019; Slovic et al., 2004).

Recent neuropsychological research has shown the positive role of emotions in decision making (Damasio & Carvalho, 2013; Fox et al., 2018). Human emotions also matter in military decision-making and an appropriate level of emotional involvement is required for all actors in the decision-making chain (Diggelen et al., 2023). Emotions not only result from making choices, but they also have anticipatory value when deciding (Loewenstein et al., 2001). Preferences are affected by emotions and change over time depending on task and context (Fox et al., 2018). This means that human involvement is constantly required in AI applications as to ensure that outcomes align with personal preferences and values.

3.1 Augmentation

AI provides opportunities for augmenting human value-based decision making in several ways. First, it can be supportive in eliciting human preferences and underlying values and interests (Jonker et al., 2016). Considering the constructive nature of human preferences and the role of emotions in value-based decision making, mapping out such interests needs to be a collaborative human-machine task. This elicitation process can help clarify underlying interests that were initially not apparent to the human user. Second, once such preferences have been elicited, mathematical optimization techniques can be used to calculate or approximate optimal choices related to these sets of preferences in combination with various decision-making strategies (Jonker & Aydoğan, 2021). In these applications, the AI system's role is to suggest choices that are superior to those the user might have made independently. Third, AI techniques can be used to create personal behavior support agents or 'electronic partners' that are aimed at supporting individuals in their activities (Neerincx & Grant, 2010; Oinas-Kukkonen, 2010; Van Riemsdijk et al., 2015). Rather than calculating optimal choices, the role of AI in these settings is to support people in their activities and

achieving their and their organization's goals in alignment with important values. These goals and values and how people respond to challenging situations differ from person to person and in different situations.
There is thus potential in using AI systems for supporting value-based decisions. However, it is important to note that while certain aspects of human values and preferences can be captured in computational models, people and social systems are inherently unpredictable (Birhane, 2021). This means that machine judgments about human needs and values are prone to incorrect interpretations. If a machine is to support people in a way that considers their personal norms and values, it is a fundamental requirement for such machines to be able to adjust their decision making based on input from people about what they find important (Birhane, 2021; Wang et al., 2022). Although AI can compare multiple choice options along a range of dimensions, value-based decisions are prone to context effects (Busemeyer et al., 2019) and should eventually be made by a human, that is, under meaningful human control.
While in judgment tasks meaningful human control typically concerns reliability and unbiased assessments, in choice tasks morality becomes the primary reason for maintaining meaningful human control. Unlike assessments, choices can lead to concrete actions with significant and lasting consequences, raising important questions of responsibility and accountability. This realization has led to the emergence and increasing importance of value-aware AI decision making. Such decision making can be enabled by capturing aspects of values and preferences in the AI system's reasoning and decision-making processes (Kayal et al., 2018; Kola et al., 2020; Van Riemsdijk et al., 2015; Tielman et al., 2018). This is, however, not enough due to the inherent human dimension in value-based decision making. At its core, value alignment should therefore be approached as an interaction problem, rather than as an optimization problem (P. Y. Chen et al., 2023).

3.2 Exploitation

Algorithms can also distort human choices in a way that hinder value-based decisions and aspired behaviour (change). For example by hypernudging, an algorithmic decision-guidance technique which harvests information about individuals' preferences and susceptibilities and exploits them to influence their future choices (Smith & de Villiers-Botha, 2021; Yeung, 2017). So far, it has mainly been used in the commercial setting to "reach the right user, with the right message, by the right means, at the right time, as many times as needed," (Morozovaite, 2021), making it almost impossible for the target to resist. Hypernudges are designed to unconsciously interfere with an individual's decision-making process (Reviglio & Agosti, 2020). In doing so, it can be used to steer individuals away from their true preferences and to manipulate people into choices that are inconsistent with their values, thereby undermining autonomous and authentic choice "in a way that is not experienced as forced" (Morozovaite, 2021). The phenomenon of hypernudging shows how "an AI can acquire and exploit the cognitive biases of a single individual, using them to change and shape the desired behaviour chosen by the choice architect in an unprecedented way" (Faraoni, 2023).
Through these personalization algorithms, social media platforms recommend content that fits a user's interest to maximize engagement, thereby disregarding content quality or social significance (Devito, 2016; Reviglio & Agosti, 2020). Through profiling (i.e., collating the traces that users leave behind online into data profiles), maleficent actors can use online social media to identify vulnerable targets to launch information warfare campaigns (Hung & Hung, 2022; Smith & de Villiers-Botha, 2021; Yeung, 2017).
Even though AI literacy might make users more aware of how algorithms can steer their decisions into directions that do not align with their values, that might not be enough. The mechanisms employed are sophisticated and operate on a level that frequently bypasses conscious awareness, making them challenging to resist, even for those who are aware of their existence. In this context, meaningful human control requires increased critical thinking skills and improved awareness of personal values and vulnerabilities to defend against manipulation. These skills may promote several aspects such as evaluating the information sources,

increasing mental resilience, identifying cognitive biases, and logical fallacies, constructing counter-arguments and reflecting on personal values and preferences. Critical thinking serves as a fundamental cognitive skill set that is essential for both individuals and societies to defend against and counter cognitive warfare tactics. It provides the tools necessary to navigate the complex information environment, resist manipulation, and maintain autonomy in thought and decision-making.

4.0 FUTURE RESEARCH QUESTIONS

4.1 Decision support

To date, research in the military domain has predominantly focused on autonomous systems, with particular emphasis on autonomous weapon systems in the context of meaningful human control. However, the application of AI in military contexts extends far beyond these systems, with significant potential to support decision-making processes in Multi-Domain Operations (MDO). While design requirements such as transparency and explainability remain crucial in this broader context, there is a pressing need for deeper insights into human cognition and decision-making processes when interacting with AI systems, with a focus on calibrated trust. This knowledge will be crucial for designing HATs that truly augment human capabilities, maintain meaningful human control, and effectively leverage the strengths of both human and artificial intelligence in complex MDO environments. As argued above meaningful human control and trust calibration highly depend on the task and reason, resulting in different implications for AI design and human competences. Given the asymmetry between AI and human epistemology, it needs to be investigated how "to appropriately integrate this product in the human decision-making process" and what "the epistemological constraints on such integration" are (Babushkina, 2022).

The ultimate objective in developing Human-AI Teams (HATs) is to achieve collaborative intelligence, a synergistic integration of human and machine capabilities that leverages the strengths and mitigates the weaknesses of both. In the context of MDO there has been a predominant focus on analytic reasoning, emphasizing information or intelligence-driven decision-making processes. This approach aligns well with the strengths of AI systems, which excel at processing vast amounts of data and identifying patterns that might elude human perception. However, human decision-making is more accurately characterized by experiential processes, drawing upon accumulated knowledge and past experiences. This intuitive approach to decision-making often involves rapid, subconscious processing that can be particularly valuable in complex, ambiguous, or time-sensitive situations. It encompasses elements such as tacit knowledge, gut feelings, and creative insights that are not easily achieved by current AI systems. The challenge lies in effectively integrating these two distinct approaches to decision-making: the analytic processes (both human and machine-driven) and the intuitive, experiential thinking characteristic of human cognition. This integration is crucial for enhancing overall decision-making capabilities, particularly in situations requiring creativity and adaptability to novel situations.

Another area of research in the development and deployment of AI systems, particularly in complex environments like military operations, concerns the allocation of requirements and responsibilities across the various actors involved in the AI system's lifecycle. This extends beyond the end-user to encompass a range of stakeholders, including AI developers, data scientists, system integrators, and policymakers. The need for a comprehensive framework that specifies requirements based on the type of AI system, the tasks it performs, and the reason for maintaining meaningful human control is becoming increasingly apparent. To develop such a comprehensive framework requires a more nuanced understanding of the AI lifecycle and its implications for eventual use in specific contexts. This approach would help ensure that meaningful

human control is maintained, ethical considerations are addressed, and the benefits of AI are realized while mitigating potential risks across various operational contexts.

4.2 Cognitive Warfare

CogWar poses an increasing threat to military operations, significantly complicating decision-making processes across all levels, from policymakers to military personnel and civilians. These campaigns often operate at a subconscious level, making them particularly insidious. To address this growing challenge, several key research areas emerge. A first line of inquiry involves understanding how adversaries design and implement influence campaigns. This includes identifying which cognitive processes are targeted and the methods employed. Gaining insight into these mechanisms and their effects on military personnel's cognitive processes is essential for developing effective defence strategies.

A second research topic is the development of technology for early detection of influence campaigns. In addition to the technology as such, insight is needed into distinct mechanisms as to extract indicators. For instance, studies have shown that bots exhibit distinct behavioural patterns compared to human users on the internet (Shu et al., 2020). Such insights can inform the development of technologies capable of identifying and removing disinformation or influence campaigns, thereby limiting their impact.

The final research direction centres on building resilience against cognitive influences. This involves increasing awareness of AI applications, improving overall AI literacy among users and to support critical thinking and value awareness. The goal is to empower individuals to make decisions that are authentic, autonomous, and aligned with their personal values, even in the face of sophisticated influence attempts.

These research areas are interconnected and vital for developing comprehensive strategies to counter the threats posed by cognitive warfare. By understanding the nature of influence campaigns, improving detection capabilities, and enhancing individual resilience, military organizations can better protect their decision-making processes and maintain operational integrity in increasingly complex information environments.

5.0 CONCLUSION

We made a distinction between two types of decision tasks, inference – and value-based tasks, and argued that AI has differential beneficial and harmful influences on both types of tasks. For each of the four categories we tried to identify the implications for meaningful human control and provided suggestions for research questions to ensure that hybrid teams of humans and machines attain the best decisions and to mitigate malicious applications of AI. Our main conclusion is that implications for both technology and human competences depend on the specific task, context and reasons for meaningful human control. This conclusion implies a comprehensive approach with a deep understanding of underlying mechanisms for both beneficial and malicious applications of AI.

To harness AI’s full potential while ensuring its safe and beneficial adoption in both the military and society, researchers from diverse fields such as ethics, law, psychology, sociology, and policymaking should collaborate to contribute to the legal, ethical and societal issues regarding future developments of AI. We argue that input from the social sciences is particularly required when AI systems evolve from support tools to buddies or team member, endowed with social skills, as it puts the role of emotions and unique human capabilities at the centre of the debate. Discussions on moral issues should not be left to AI designers; collaborative effort is needed to shape the future that serves humanity best.

References

Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. *Ethics and Information Technology*, *24*(2), 1–15. https://doi.org/10.1007/s10676-022-09629-y

Bernal, A., Carter, C., Singh, I., Cao, K., & Madreperla, O. (2020). Cognitive Warfare: An Attack on Truth and Thought. In *Journal of Language Relationship* (Vol. 12, Issue 1). https://doi.org/10.31826/jlr-2015-120101

Birhane, A. (2021). *Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence* (Issue October 2021) [University College Dublin]. http://arxiv.org/abs/2206.04179

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, *23*(3), 251–263. https://doi.org/10.1016/j.tics.2018.12.003

Campbell, M., Hoane, A. J., & Hsu, F. H. (2002). Deep Blue. *Artificial Intelligence*, *134*(1–2), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics*, *3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chen, J. X. (2016). The Evolution of Computing: AlphaGo. *Computing in Science and Engineering*, *18*(4), 4–7. https://doi.org/10.1109/MCSE.2016.74

Chen, P. Y., Tielman, M. L., Heylen, D. K. J., Jonker, C. M., & Van Riemsdijk, M. B. (2023). Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues. *Frontiers in Artificial Intelligence and Applications*, *368*, 93–107. https://doi.org/10.3233/FAIA230077

Claverie, B., Prébot, B., Buchler, N., & Cluzel, F. Du. (2021). What is Cognition? And How to Make it One of the Ways of the War. In *Cognitive Warfare: The Future of Cognitive Dominance*. NATO.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, *14*(2), 143–152. https://doi.org/10.1038/nrn3403

Davidovic, J. (2023). On the purpose of meaningful human control of AI. *Frontiers in Big Data*, *5*. https://doi.org/10.3389/fdata.2022.1017677

De Cremer, D., & Kasparov, G. (2021). AI Should Augment Human Intelligence, Not Replace It. *Harvard Business Review*, 1–9.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S. C., Shaw, T. H., Pak, R., & Neerincx, M. A. (2019). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, 1–20. https://doi.org/10.1007/s12369-019-00596-x

Devito, M. A. (2016). From Editors to Algorithms: A Values-Based Approach to Understanding Story Selection in the Facebook News Feed. *Digital Journalism*, 1–20. https://doi.org/10.1080/21670811.2016.1178592

Diggelen, J. Van, Metcalfe, J. S., Bosch, K. Van Den, Neerincx, M. A., & Kerstholt, J. H. (2023). Role of emotions in responsible military AI. *Ethics and Information Technology*, *25*(17), 2–5.

Doswell, C. A. (2004). Weather forecasting by humans - Heuristics and decision making. *Weather and Forecasting*, *19*(6), 01115–01126. https://doi.org/10.1175/waf-821.1

Faraoni, S. (2023). Persuasive Technology and computational manipulation: hypernudging out of mental self-determination. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1216340

Fox, A. S., Lapate, R. C., Shackman, A. J., & Davidson, R. J. (2018). *The Nature of Emotion: Fundamental Questions.* (Second edi). Oxford University Press.

https://doi.org/10.1037/h0074005
Hung, T. C., & Hung, T. W. (2022). How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars. *Journal of Global Security Studies*, *7*(4), 1–18. https://doi.org/10.1093/jogss/ogac016
Jermutus, E., Kneale, D., Thomas, J., & Michie, S. (2022). Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review. *Wellcome Open Research*, *7*, 65. https://doi.org/10.12688/wellcomeopenres.17550.1
Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction*, *3*(1), 43. https://doi.org/10.5898/jhri.3.1.johnson
Jonker, C. M., & Aydoğan, R. (2021). Deniz: A Robust Bidding Strategy for Negotiation Support Systems. *Studies in Computational Intelligence*, *905*, 29–44. https://doi.org/10.1007/978-981-15-5869-6_3
Jonker, C. M., Aydoğan, R., Baarslag, T., Broekens, J., Detweiler, C. A., Hindriks, K. V., Huldtgren, A., & Pasman, W. (2016). An Introduction to the Pocket Negotiator: A General Purpose Negotiation Support System. *Multi-Agent Systems and Agreement Technologies*, *1*, 13–27. https://doi.org/10.1007/978-3-319-59294-7
Kayal, A., Brinkman, W. P., Neerincx, M. A., & Birna Van Riemsdijk, M. (2018). Automatic resolution of normative conflicts in supportive technology based on user values. *ACM Transactions on Internet Technology*, *18*(4), 1–21. https://doi.org/10.1145/3158371
Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385
Kola, I., Jonker, C. M., Tielman, M. L., & Birna van Riemsdijk, M. (2020). Grouping situations based on their psychological characteristics gives insight into personal values. *CEUR Workshop Proceedings*, *2787*, 17–26.
Kox, E. S., Finlayson, N. B., Broderick-Hale, J. C., & Kerstholt, J. H. (2023). Calibrated Trust as a Means to Build Societal Resilience Against Cognitive Warfare. *Proceedings of NATO Symposium HFM-361 on Mitigating and Responding to Cognitive Warfare*, 1–15.
Kox, E. S., Siegling, L. B., & Kerstholt, J. H. (2022). Trust development in military and civilian Human-Agent Teams: the effect of social-cognitive recovery strategies. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00871-4
Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *Philosophy and Technology*, *35*(1). https://doi.org/10.1007/s13347-022-00511-9
Lee, C. E., Baek, J., Son, J., & Ha, Y. G. (2023). Deep AI military staff: cooperative battlefield situation awareness for commander's decision making. *Journal of Supercomputing*, *79*(6), 6040–6069. https://doi.org/10.1007/s11227-022-04882-w
Lee, J. D., & See, K. A. (2004). *Trust in Automation : Designing for Appropriate Reliance*. *46*(1), 50–80.
Loewenstein, G. F., Hsee, C. K., Weber, E. U., & Welch, N. (2001). Risk as Feelings. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.127.2.267
Mahmud, B. U., & Sharmin, A. (2021). *Deep Insights of Deepfake Technology : A Review*. http://arxiv.org/abs/2105.00192
Masakowski, Y. R., & Blatny, J. M. (2023). *Mitigating and Responding to Cognitive Warfare*.
Matthews, G., Panganiban, A. R., Bailey, R., & Lin, J. (2018). Trust in Autonomous Systems for Threat Analysis: A Simulation Methodology. *International Conference of Virtual, Augmented and Mixed Reality*, *10910 LNCS*, 116–125. https://doi.org/10.1007/978-3-319-91584-5_10

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. http://arxiv.org/abs/1706.07269

Miller, T. (2022). Are we measuring trust correctly in explainability, interpretability, and transparency research? *TRAIT 2022: Trust and Reliance in AI-Human Teams*, *1*(1), 1–11. http://arxiv.org/abs/2209.00651

Morozovaite, V. (2021). Two Sides of the Digital Advertising Coin: Putting Hypernudging into Perspective. *Market and Competition Law Review*, *5*(2), 105–145. https://doi.org/10.34632/mclawreview.2021.10307

NATO. (2023). *Alliance Concept for Multi-Domain Operations (MDO)*. *2022*.

Neerincx, M. A., & Grant, T. (2010). Evolution of electronic partners: Human-automation operations and epartners during planetary missions. *Journal of Cosmology*, *12*, 3825–3833.

Oinas-Kukkonen, H. (2010). Behavior change support systems: A research model and agenda. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6137 LNCS*(May), 4–14. https://doi.org/10.1007/978-3-642-13226-1_3

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Reviglio, U., & Agosti, C. (2020). Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media. *Social Media and Society*, *6*(2). https://doi.org/10.1177/2056305120915613

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), 1–40. https://doi.org/10.1002/widm.1385

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. *Risk Analysis*, *24*(2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Smith, J., & de Villiers-Botha, T. (2021). Hey, Google, leave those kids alone: Against hypernudging children in the age of big data. *AI and Society*. https://doi.org/10.1007/s00146-021-01314-w

Tenove, C. (2020). Protecting Democracy from Disinformation: Normative Threats and Policy Responses. *International Journal of Press/Politics*, *25*(3), 517–537. https://doi.org/10.1177/1940161220918740

Tielman, M. L., Jonker, C. M., & van Riemsdijk, M. B. (2018). What should I do? Deriving norms from actions,values and context. *CEUR Workshop Proceedings*, *2134*, 35–40.

Van Riemsdijk, M. B., Jonker, C. M., & Lesser, V. (2015). Creating Socially Adaptive Electronic Partners: Interaction, reasoning and ethical challenges. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2*(Aamas), 1201–1206.

Wang, A., Kapoor, S., & Narayanan, A. (2022). *Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy*. *1*(1), 1–45. https://papers.ssrn.com/abstract=4238015

Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2016). *The Effect of Agent Reasoning Transparency on Automation Bias: An Analysis of Response Performance* (Vol. 9740, pp. 465–477). https://doi.org/10.1007/978-3-319-39907-2_45

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information Communication and Society*, *20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

From exploitation to augmentation: Navigating AI's impact on human value- and inference-based decision making