OPEN ACCESS

Give us a hand, mate! A holistic review of research on human-machine teaming

Jitu Patel ,[1] M Boardman,[1] B Files,[2] F Gregory,[2] S Lamb,[3] S Sarkadi,[4] M Tešić,[5,6] N Yeung[7]

[1]DSTL, Salisbury, UK
[2]US Army Research Laboratory, Aberdeen Proving Ground, Maryland, USA
[3]MOD, London, UK
[4]King's College London, London, UK
[5]Birkbeck University of London, London, UK
[6]Leverhulme Centre for the Future of Intelligence, Cambridge, UK
[7]University of Oxford Social Sciences Division, Oxford, UK

Correspondence to
Dr Jitu Patel; jmpatel@dstl.gov.uk

Received 1 May 2024
Accepted 17 October 2024

ABSTRACT

Defence has a significant interest in the use of artificial intelligence (AI)-based technologies to address some of the challenges it faces. At the core of future military advantage will be the effective integration of humans and AI into human-machine teams (HMT) that leverages the capabilities of people and technologies to outperform adversaries. Realising the full potential of these technologies will depend on understanding the relative strengths of humans and machines, and how we design effective integration to optimise performance and resilience across all use cases and environments.

Since the first robot appeared on the assembly line, machines have effectively augmented human capability and performance; however, they fall short of being a team member—someone you can ask to give you a hand! Working in teams involves collaboration, adaptive and dynamic interactions between team members to achieve a common goal. Currently, human-machine partnership is typically one of humans and machines working alongside each other, with each conducting discrete functions within predicable process and environments. However, with recent advances in neuroscience and AI, we can now envisage the possibility of HMT, not just in physical applications, but also complex cognitive tasks.

This paper provides a holistic review of the research conducted in the field of HMT from experts working in this area. It summarises completed and ongoing studies and research in the UK and USA by a broad group of researchers. This work was presented in the HMT thematic session at the Sixth International Congress on Soldiers' Physical Performance (ICSPP23 London).

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence and autonomy are developing at an accelerating pace, but defence remains a fundamentally human endeavour.
⇒ Advances in human sciences, including neuroscience, cognitive science and teaming science, as well as in autonomy, including artificial intelligence, robotics and computer science, will be needed to support fully functioning human-machine teams that can address near-term and far-term defence challenges.

WHAT THIS STUDY ADDS

⇒ This targeted review identifies and summarises key efforts to leverage human-autonomy teams to address important defence capabilities.
⇒ These efforts span topics including brain-computer interfaces, trust in human-machine teams, explainable artificial intelligence, theory of mind, cognitive models and hybridised human-machine systems.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ By reviewing recent progress and identifying gaps and promising findings, this targeted review points the way for future research efforts that could enable human-autonomy teams to better address needed capabilities of the future.

To cite: Patel J, Boardman M, Files B, *et al*. *BMJ Mil Health* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/military-2024-002737

INTRODUCTION

Recent advances in artificial intelligence (AI) and its application across a multitude of fields have highlighted the potential impact that this technology could have on how tasks, especially cognitive tasks typically conducted by humans, might change in the future. Consequently, Defence has a significant interest in the use of AI-based technologies to address some of the challenges it faces. However, the complex, dynamic and uncertain nature of military activity, and operational contexts combined with ethical and legal considerations will require defence to adopt a human-machine teaming (HMT) approach to the use of AI in many applications.

At the core of future military advantage will be the effective integration of humans, AI and robotics into warfighting systems—human-machine teams—that leverages the capabilities of people and technologies to outperform adversaries. Realising the full potential of these technologies will depend on understanding the relative context-dependent strengths of humans and machines, and how we design effective interaction to optimise performance and resilience across all use cases and environments.

This paper provides a holistic review of the research conducted in the field of HMT from experts who are undertaking ongoing working in this area. It summarises completed and ongoing studies and research in the UK and USA that was presented in the HMT theme session at the Sixth International Congress on Soldiers' Physical Performance (ICSPP London, September 2023).

BRAIN-COMPUTER INTERFACE

In future, AI agents with the highest level of full autonomous capabilities will be used across civilian, medical and military operations. However, numerous challenges must be addressed to realise trusted human-AI interactions that enable enhanced team performance.[1] Autonomous AI agents used in

national security and defence will not be restricted to completing assistive tasks as subservient entities. Instead, intelligent agents will have distinct roles within hybrid teams of multiple humans and/or multiple agents executing complex sequences of tasks for activities like command and control. The emergent joint cognitive systems have the potential to supersede the strengths of humans or intelligent agents alone. Achieving trusted human-machine teaming with the ability to solve complex decision-making tasks will require additional pathways to accomplish shared situational awareness.[2]

The US Office of the Secretary of Defense and the UK Ministry of Defence jointly sponsored the Bilateral Academic Research Initiative (BARI) Pilot Programme to pursue high-risk basic research to uncover new human-agent interaction principles for enhancing shared human-AI situational awareness (SA). New methods incorporating brain-computer interface (BCI) decoding of human state were explored to identify objective indicators of human trust, decision confidence and communication states that may facilitate real-time AI model update.[3] The idea of direct brain-computer communication has been around for over 50 years[4] with the field of BCIs evolving to include non-invasive and non-medical applications and use in healthy subjects.[5] Collaborative BCIs were explored as a potential conduit for transparent teaming of intelligent autonomous agents with their human counterparts.[6]

Brain and body signals from human teammates provide potential real-time indicators for translating conscious and subconscious human behaviours, and mental states, into effects on human-agent team interaction. The BARI pilot programme facilitated the creation of hybrid collaborative BCI capable of decoding human decision confidence[7] and means for providing group feedback to an agent resulting in accelerated perceptual group decision-making.[8] In addition, access to decision confidence of human users enabled enhanced team performance in tasks involving human-AI teams.[9] Potentially, hybrid BCI signals can also assist in the assessment of fatigue/attention levels of human decision makers and allow breaks or workload reductions when cognitively overloaded, thereby offering greater potential for shared SA in future teams. The knowledge gained from hybrid collaborative BCI approaches applied to individuals and groups of humans in human-AI teams may inform a system of transparent governance principles that will guide future research, development and acquisition.

TRUST

Trust is a crucial element of human-machine teaming, just as it is in effective human-human teams where trust underpins effective information sharing, division of labour and delegation of duties.[10] Trust must be appropriately calibrated to the task at hand, the situation and corresponding abilities of team members. Yet research has documented striking divergence between cases where AI and automated systems are under-trusted and therefore under-utilised and others where obviously malfunctioning systems continue to be relied on by human decision makers. These failures of trust highlight the need to build AI systems that are sensitive to the principles of trust that underpin human social interactions.

To date, principles governing trust and influence in human decision-making, as identified in experimental psychology and neuroscientific studies, have not been systematically explored in the context of HMT. The BARI project aimed to fill this gap: identifying key principles of interpersonal trust and applying these to understand human-machine trust. The overall BARI goals and the research thrusts are encapsulated in the following conceptual architecture, which envisages a team of human and AI (virtual human) teammates working flexibly together in complex decision environments. This architecture involves HMT at two levels: each human decision maker operates with a 'virtual personal assistant' BCI to optimise their individual output, then output from these 'local' HMTs are combined with AI decision makers ('virtual humans') for the overall ('global' HMT) team decision.

Conceptual architecture is shown in Figure 1 and developments are organised against four primary research thrusts:

Thrust 1 (virtual human)

Humanising machines for human teaming. The research was focused on developing virtual humans/agents.

Thrust 2 (virtual personal assistant)

Translating human conscious and unconscious behaviours, mental states and body language into machine-interpretable form, and modulating effects of communication. The research focus was on developing multimodal, real-time, hybrid BCIs.

Thrust 3 (virtual team assistant)

Optimising decision integration, group performance measures and team/teamwork coordination. The aim was to develop a virtual team assistant.

Thrust 4 (integration and test)

The goal was to integrate and test the components developed in Trusts 1, 2 and 3.

Within this architecture, the BARI project identified several factors that influence the effectiveness of the local HMT: the user state of the human in the loop (cognitive load; level of task focus), the timing of input from their virtual personal assistant, and the nature of this input.[11] We also find wide individual differences in the degree to which human decision makers trust AI systems, even when carefully controlling for objective performance of both human and AI teammates.[10]

Human factors are also critical to the design of the global HMT to ensure effective sharing and use of information across the team.[12] The BARI architecture focuses on sharing of decision confidence as an effective way to weight individual team-member opinions and arbitrate in cases of conflict, using BCIs as described above to optimise this process. A key open question for future research is how to ensure mutual trust as HMTs become more complex in their architecture and capabilities of AI systems increase, where factors such as need for effective shared responsibility will become paramount.[13]

EXPLANATIONS

Explainable artificial intelligence (XAI) seeks to make the decision-making processes of AI systems transparent, aiming to foster appropriate trust among users. By generating explanations for the behaviours of AI systems, XAI aims to make explicit correlations and associations identified by these systems within data. However, merely making these correlations explicit does not convert them into causal relationships. It is a well-established fact in cognitive science that people can interpret explanations causally.[14] In this context, it becomes crucial to explore whether people attribute causal interpretations to those explanations, even when such interpretations are not justified by the explanations themselves.[15]

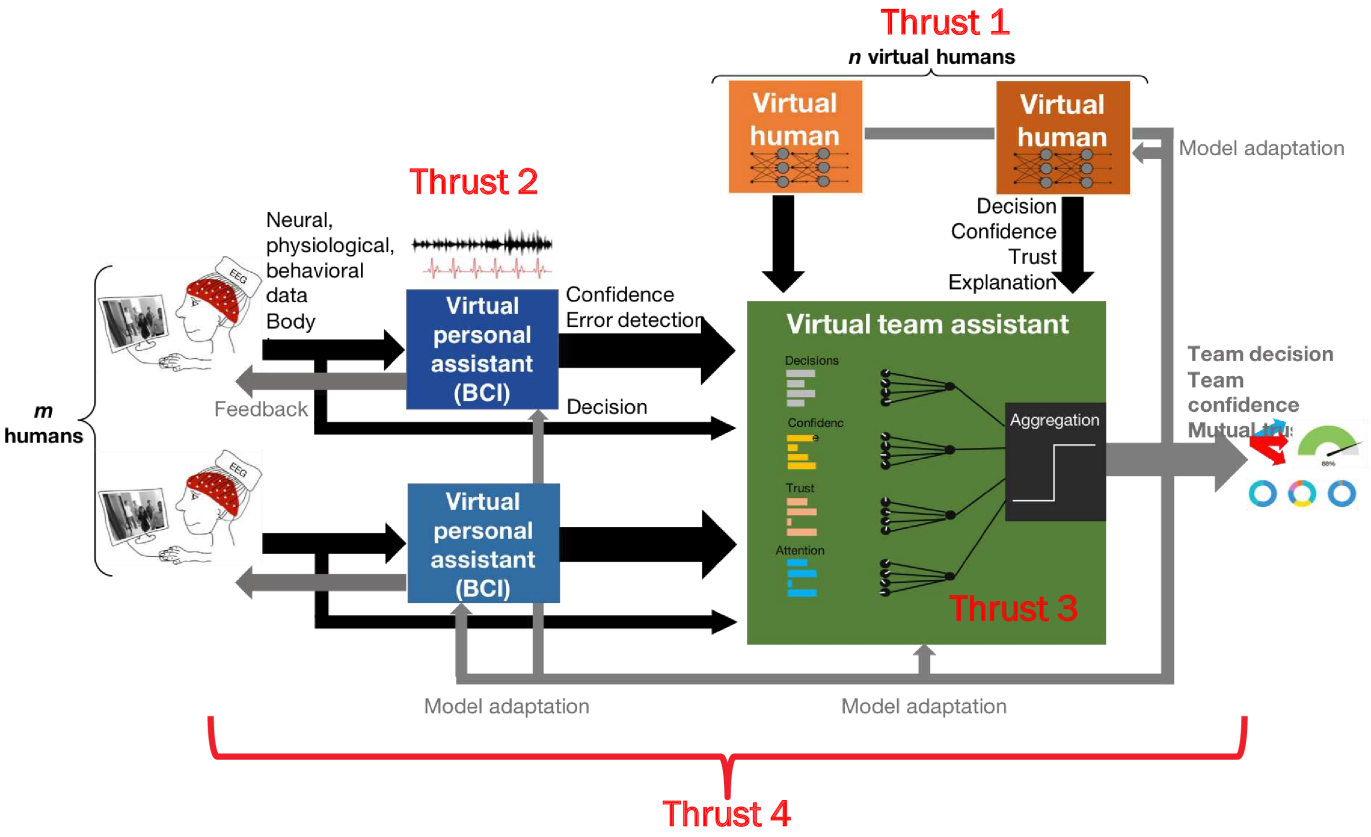

Figure 1 Joint cognitive system for collaborative human-artificial intelligence decision-making. BCI, brain-computer interface.

Counterfactual (CF) explanations have emerged as a prominent method within XAI due to the fact that people often use CF explanations and their potential to suggest actionable recourse.[15] For example, an AI system may predict that an individual is likely to default on a loan based on attributes such as salary and education level. A CF explanation could state: "Had the individual's salary been higher than £X, the AI system would have predicted a successful loan repayment." This might motivate the person to take action in the real world to increase their salary. However, such an explanation communicates the AI system's identification of a strong correlation between salary and loan default risk. It does not establish a causal relationship between the two, meaning that actions taken solely on the basis of the counterfactual explanation, such as looking for ways to increase salary with the hope of reducing default risk, would not be justified.

A wealth of research in cognitive science and psychology has highlighted humans' predisposition to interpret counterfactuals as indicative of causal relationships between events. For instance, the counterfactual 'Had it not rained, the road would not have been slippery' is commonly interpreted as rain causing the road to be slippery. In a series of experiments, we explored whether people interpret the relationships between events in CF explanations causally. Our findings indicate that CF explanations of AI systems' predictions can indeed influence people's causal beliefs, making them more likely to view events mentioned in CF explanations as causally connected. Drawing inspiration from literature on misinformation and health warning messaging, we also investigated whether it is possible to correct the unjustified change in causal beliefs. We discovered that highlighting the fact that AI systems identify correlations, rather than causal relationships, can mitigate the impact of CF explanations on people's causal beliefs.[15]

This exploration into the psychological impact of AI explanations seeks to promote more mindful and responsible use of XAI. We aim to prevent unintended misconceptions created by using tools, such as CFs, that are familiar to people in novel ways and contexts.

DECEPTION

Deception is the intentional process of an agent to cause another agent to have a false belief given an ulterior goal of the deceiver.[16] Deceptive AI research aims to study deception in the context of human-machine ecosystems for high-level reasoning, team formation and decision-making in defence and intelligence analysis. Three main approaches were used to explore deception as a sociocognitive process that takes place between humans and machines.

The first approach is the modelling of cognitive agent architectures to enable the representation of human-like internal reasoning and communication processes responsible for deception and deception detection. One of these processes, crucial for team coordination is Theory of Mind, that is, mentalisation, for example, "I know that you know that I know…."[16]

The second approach explores a higher level of abstraction by assuming the presence of the underlying cognitive agent architectures. This is used for modelling the sociocognitive costs and rewards of communicating knowledge in order to simulate deception and deception detection as evolutionary and adaptive processes.[17]

By simulating evolutionary sociocognitive processes, we have shown (1) that decentralised knowledge management is better at dealing with deception than centralised approaches, (2) that a decentralised interrogation and punishment in knowledge

sharing is resilient in the face of large-scale disinformation attacks, (3) that an arms race in Theory of Mind between deceivers and deception detectors happens in agent societies and detectors need to always operate at a higher level of Theory of Mind in order to maintain cooperation; (4) that lying and deceiving respond to distinct evolutionary pressures of communication; and (5) that agents choose deception over truthfulness when detectors are present, but only when agents care to do better than others and not when they only care for themselves.[18]

The third approach explores how humans perceive their interactions with deceptive AI as part of social ecosystems. One example of this approach is described in Sarkadi *et al*[19] where US-based participants were asked to evaluate deceptive AI behaviour versus deceptive human behaviour in future-of-work scenarios. Results showed that participants did not care whether the deception was performed by a human or an AI.

Finally, all three approaches emphasise that we need to carefully consider the complexities of emergent cognitive and behavioural processes responsible for deception in an ever-evolving human-machine interactions.

COGNITIVE MODELS

The adoption of HMT for intelligence analysis is proposed to increase the pace, volume and accuracy of data that is analysed. Within this section, we describe learning from a longitudinal qualitative study of expert intelligence analysts, as they adopt an imperfect machine vision tool that supports identification and tracking of objects while analysts conduct scene recognition within live and historical motion videos.[20] This tool is being adopted within a time constrained, complex operating environments. We investigated how the two models of cognition and decision-making can inform trust and adoption of these tools.

Klein's Recognition Primed Decision-Making model (RPD)[21] was selected as it has been shown to describe time constrained decision-making under uncertainty, thereby matching the operating environment observed. Second, Cummings' relative-strength model,[22] shown in Figure 2, was selected, as it informed our understanding of the distributed decision-making between the tool and analysts that can be applied to the design of human-machine interaction. The relative-strength model was adapted using Kamkar's description of object detection, recognition, object tracking and activity and scene recognition[23] as these functions described the activities the intelligence analysts were undertaking.

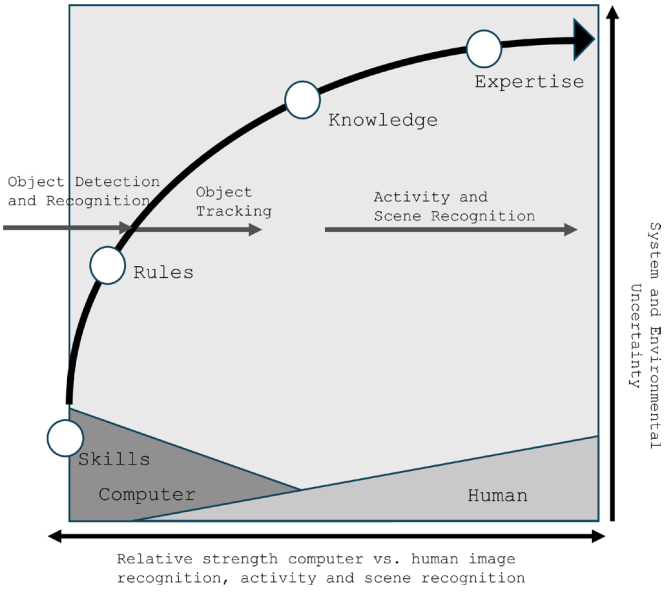

Figure 2 An adapted relative strengths model.

A thematic analysis[24] of analysts interviews on tool adoption confirmed that the design and use of these tools could be informed by Klein's RPD model and the adapted Cummings' relative strength model.

Second, it was found that Cummings' relative strength model could be further developed, while it identifies uncertainty of the operating environment as driving a need to move from skills to rules, knowledge and expertise, it does not recognise the perceived criticality of the situation, nor the imperfection of the tool, as it missclassifies objects. Analysts describe that in highly critical situations they would be less likely to trust the tool, conducting and checking a greater portion of object detection, recognition and tracking. Further research should investigate if there is a relationship between situational criticality and the uncertainty of the tool's classifications, further developing this model.

STRENGTHENING TEAMWORK

HMT covers a range of possible scales, from small teams to large systems of interconnected, heterogeneous teams of humans and machine systems. It also covers a range of interactions, from system-as-tool to system-as-teammate to hybridised human-machine systems. US Army Research Laboratory's (ARL) Strengthening Teamwork for Robust Operations in Novel Groups (STRONG) collaborative research alliance is building a cross-disciplinary collaborative ecosystem to push research and development toward the hard problems of large-scale, high complexity, high uncertainty HMTs that leverage the full range of possible human/machine interactions.[25 26] A motivating application space for this foundational research is future Command and Control, in which large teams of humans and machines plan and coordinate actions of thousands of human and machine agents in a complex, uncertain environment.[27]

STRONG is examining synergistic hybrid human-machine intelligence to enable potentially many humans and autonomous systems to jointly think, ideate and plan in contexts so complex and ill-defined to be outside the capabilities of humans or machines alone. Related concepts exist within the scientific literature, such as hybrid intelligence[28] and intelligent cognitive extenders,[29] which consider machine intelligence that extends the capabilities of humans. STRONG seeks to expand on these frameworks to include research on how to help humans better integrate with and extend the capabilities of machines. Progress is needed in human-focused areas, including neurotechnologies and technological fluency.[30] Progress is also needed in scalable interactive machine learning,[27] including the need for algorithms that can accept human interactions and feedback in complex, dynamic situations; resilient human/machine teams with adaptable roles and configurations; and algorithms and interactions that can operate hierarchically in terms of spatial, temporal and organisational scale.

A central premise of STRONG is that progress on these critical research areas depends crucially on close collaboration among experts from many disciplines. Human/system integration in complex, uncertain and ill-defined problem spaces goes well beyond the simplified problems and interactions studied by individual disciplines. To facilitate this, ARL and the UK Defence Science and Technology Laboratory are working together to build an expansive, cross-disciplinary collaborative ecosystem

to facilitate advances that will help define the future of human-machine teaming.

FINAL THOUGHTS

Since the first robot arm started working in the General Motors assembly line in 1961, machines have evolved and proliferated in all domains from agriculture, defence, hospitals, manufacturing and into our homes where they carry out household chores such as gardening and vacuuming. These machines have augmented human capability and performance; however, they fall short of being a team member—someone you can ask to give you a hand! Working in teams involves collaboration, adaptive and dynamic interactions between team members to achieve a common goal. Currently, human-machine partnership is typically one of humans and machines working alongside each other with each conducting discrete functions within predicable process and environments. However, with advances in neuroscience, AI and robotics over the last decade, we can now envisage the possibility of HMT, not just in physical applications, but also complex cognitive tasks.

Contributors JP, MB, BF, FG, SL, SS, MT and NY made equal contributions. JP is the guarantor.

Funding This study was funded by US DoD (N/A) and UK MOD (N/A).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

Jitu Patel http://orcid.org/0009-0009-4926-0712

REFERENCES

1 National Academies of Sciences, Engineering, and Medicine. *Human-AI teaming: state-of-the-art and research needs*. Washington, DC: The National Academies Press, 2022.
2 Endsley MR. Supporting Human-AI Teams:Transparency, explainability, and situation awareness. *Comput Human Behav* 2023;140:107574.
3 Boldt A, Yeung N. Shared neural markers of decision confidence and error detection. *J Neurosci* 2015;35:8:3478–84:.
4 Vidal JJ. Toward direct brain-computer communication. *Annu Rev Biophys Bioeng* 1973;2:157–80.
5 van Erp J, Lotte F, Tangermann M. Brain-Computer Interfaces: Beyond Medical Applications. *Computer (Long Beach Calif)* 2012;45:26–34.
6 Poli R, Valeriani D, Cinel C. Collaborative brain-computer interface for aiding decision-making. *PLoS One* 2014;9:e102693.
7 Sadras N, Sani OG, Ahmadipour P, *et al*. Post-stimulus encoding of decision confidence in EEG: toward a brain-computer interface for decision making. *J Neural Eng* 2023;20:056012.
8 Bhattacharyya S, Valeriani D, Cinel C, *et al*. Anytime collaborative brain-computer interfaces for enhancing perceptual group decision-making. *Sci Rep* 2021;11:17008.
9 Valeriani D, O'Flynn LC, Worthley A, *et al*. Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario. *J Neural Eng* 2022;19:056036.
10 Celaya A, Yeung N. Confidence and trust in human-machine teaming. *HDIAC J* 2019;6:21–5.
11 Davidson MJ, Macdonald JSP, Yeung N. Alpha oscillations and stimulus-evoked activity dissociate metacognitive reports of attention, visibility, and confidence in a rapid visual detection task. *J Vis* 2022;22:20.
12 Kämmer JE, Choshen-Hillel S, Müller-Trede J, *et al*. A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision (Wash D C)* 2023;10:107–37.
13 El Zein M, Bahrami B, Hertwig R. Shared responsibility in collective decisions. *Nat Hum Behav* 2019;3:554–9.
14 Tešić M, Hahn U. Explanation in ai systems. In: Muggleton S, Chater N, eds. *Human-like machine intelligence*. Oxford University Press, 2021: 114–36.
15 Tešić M, Hahn U. Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that? *Patt (N Y)* 2022;3:100635.
16 Sarkadi Ş. Deceptive AI and Society. *IEEE Technol Soc Mag* 2023;42:77–86.
17 Sarkadi Ş. Self-Governing Hybrid Societies and Deception. *ACM Trans Auton Adapt Syst* 2024;19:1–24.
18 Sarkadi S, Lewis PR. The triangles of dishonesty: modelling the evolution of lies, bullshit, and deception in agent societies. Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024); 2024
19 Sarkadi S, Mei P, Awad E. Should my agent lie for me? a study on attitudes of us-based participants towards deceptive ai in selected future-of-work scenarios. Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS); 2023:345–54.
20 Lamb SC, Ramchurn SD, Norman TJ, *et al*. Cognitive models to inform the design of ai tools for intelligence analysts. 2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense); Rome, Italy, 2023:461–5.
21 Klein GA. A recognition-primed decision (RPD) model of rapid decision making. In: *Decision making in action: models and methods*. 1993.
22 Cummings M. Informing Autonomous System Design Through the Lens of Skill-, Rule-, and Knowledge-Based Behaviors. *J Cogn Eng Decis Mak* 2018;12:58–61.
23 Kamkar S, Ghezloo F, Moghaddam HA, *et al*. Multiple-target tracking in human and machine vision. *PLoS Comput Biol* 2020;16:e1007698.
24 Braun V, Clarke V. Thematic analysis revised. *J Chem Inf Model* 2019;53:1689–99.
25 DeCostanza AH, Marathe AR, Bohannon A, *et al*. Enhancing human-agent teaming with individualized, adaptive technologies: a discussion of critical scientific questions, ARL-TR-8359. 2018. Available: https://brain.ieee.org/brain-storm/enhancing-human-agent-teaming
26 Metcalfe JS, Perelman BS, Boothe DL, *et al*. Systemic Oversimplification Limits the Potential for Human-AI Partnership. *IEEE Access* 2021;9:70242–60.
27 Madison A, Novoseller E, Goecks VG, *et al*. Scalable interactive machine learning for future command and control (arXiv:2402.06501) [arXiv]. 2024.
28 Akata Z, Balliet D, de Rijke M, *et al*. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer (Long Beach Calif)* 2020;53:18–28.
29 Hernández-Orallo J, Vold K. AI extenders: the ethical and societal implications of humans cognitively extended by ai. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019:507–13.
30 Pollard KA, Files BT, Oiknine AH, *et al*. *How to prepare for rapidly evolving technology: focus on adaptability (ARL-TR-9432)*. DEVCOM Army Research Laboratory, 2022.