nature reviews psychology

https://doi.org/10.1038/s44159-025-00499-x

Perspective

Check for updates

A cognitive approach to human–AI complementarity in dynamic decision-making

Cleotilde Gonzalez [1] & Hoda Heidari [2]

Abstract

As artificial intelligence (AI) becomes increasingly integrated into complex decision-making environments, there is a growing need to develop AI systems that complement human capabilities. AI and humans offer distinct strengths: AI excels at processing large datasets, identifying statistical patterns and optimizing predefined objectives, whereas humans are skilled at navigating uncertainty, novelty and interpersonal challenges. The synergy between humans and AI is particularly vital in dynamic decision-making domains — such as disaster response situations — in which rapid analysis of AI results must be balanced with human judgement and ethical considerations. In this Perspective, we provide a conceptual framework to integrate human decision-making with AI, focusing on cognitive AI: a computational approach that models human cognitive processes to create AI systems that learn and make decisions in ways similar to those of humans. We discuss the elements and necessary capabilities of cognitive AI and how to realize human–AI complementarity in decision-making while considering ethical risks. By advancing these areas, researchers can lay the groundwork for adaptive and cognitively grounded human–AI teamwork that is aligned with human values and goals.

Sections

[1]Social and Decision Sciences Department, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA. e-mail: coty@cmu.edu

Introduction

Humans make thousands of decisions per day, from trivial choices — like what movie to watch — to very complex and consequential choices — like what cancer treatment to take. Some of those decisions are static, meaning that the individual needs to make a single decision in a steady environment. Examples of such decisions are picking a name for a newly born child or an insurance plan from a fixed set of options. However, many decisions are dynamic in nature: the environment changes autonomously over time, the decision-making task itself might evolve — possibly in response to previous decisions — and the decision-maker can learn from and adapt to previous observations. For example, during an emergency response operation, a firefighter conducts real-time assessment of the situation and adjusts their actions based on rapidly changing conditions in the world[1]. As in this example, dynamic decision-making involves making a sequence of interdependent decisions in a constantly evolving environment[2]. In dynamic decision-making, humans often face multiple practical constraints — such as time pressure, workload and limited resource availability — that complicate decision-making[3].

Data-driven AI systems play an essential part in supporting dynamic decision-making in many domains[4,5]. These computational models leverage vast amounts of data to learn and make predictions by identifying key patterns through statistical modelling[6,7]. Sometimes, full automation (systems that operate without human intervention) of decision-making is feasible and effective with data-driven AI. For instance, data-driven AI systems can make autonomous decisions in online advertising or by finding and displaying the shortest path to a destination on a virtual map. However, in many dynamic tasks, full automation is not possible or appropriate because data-driven AI can fail when faced with novel, rapidly changing conditions, ambiguity or incomplete information — in such cases, human input and oversight are necessary[8]. For example, human oversight is often necessary to ensure quality control and compliance with legal requirements, especially in critical tasks like medical diagnosis[9,10]. Additionally, tasks that involve societal tradeoffs — such as pre-trial criminal risk assessments of defendants — require human professionals to adhere to ethical norms and involve complex judgements that are currently beyond the reach of AI predictions[11–15].

Thus, in dynamic decision-making tasks, data-driven AI is mostly used as a tool that takes advantage of the power of computing and data-driven automation to provide recommendations and advice to humans, who make the final decisions[16]. Examples include AI-powered socially assistive robots for elderly care, nursing and healthcare[17,18]. Current data-driven AI systems have major difficulties in adapting to dynamic environments, accounting for human input in real time, and automatically adjusting their recommendations on the basis of changing goals, shifting contexts, user feedback and evolving patterns in the environment[19].

Despite these limitations, there is a vision of a future in which AI is not a mere tool or assistant but a member of a team including human decision-makers[20–25]. In this vision, humans and AI form a synergistic relationship known as human–AI complementarity, in which both can operate at comparable levels of autonomy and complement each other's strengths and weaknesses so that they arrive at a more informed and balanced decision together than either could alone. Autonomy refers to the capacity of an agent (human or artificial) to make and act on decisions independently, without external control. Levels of autonomy describe the extent to which an agent acts with external intervention, ranging from full intervention (no autonomy) to no external intervention (full autonomy). For example, in a medical decision situation, a physician and a diagnostic AI system might operate at comparable levels of autonomy during the diagnosis and treatment of a patient. The AI system independently analyses the patient's medical history, laboratory results and imaging scans to identify a potential diagnosis and treatment recommendations, while the physician evaluates the patient's symptoms, asks clarifying questions and weighs up the emotional and social factors. Together, they arrive at a more informed decision than either could alone.

The goal of human–AI complementarity is to combine human strengths with the computational power of AI to produce better decisions than either the humans or AI could achieve independently[26–30]. However, current human–AI systems often fail to achieve true complementarity, and the reasons for this shortfall remain unclear[30]. Achieving human–AI complementarity will require addressing a complex set of factors, and integrating different perspectives and potential alternative approaches[5,16,20,31–33].

Of the possible approaches to achieve human–AI complementarity, one with great promise is cognitive AI. Cognitive AI aims to emulate and simulate the human mind as an information-processing system. To emulate means to replicate the cognitive process that led to human decisions, whereas to simulate means to model these processes computationally to understand better how decisions emerge. Thus, cognitive AI functions similarly to early AI systems, which simulate human beings to emulate the way humans' minds work[34]. Cognitive AI, in combination with data-driven AI, has the potential to enable effective complementarity during dynamic decision-making. Cognitive AI aims to model human-like reasoning, memory and decision-making to mimic how people perceive, interpret and respond to complex tasks. By contrast, data-driven AI relies on large-scale statistical patterns in data to identify correlations and make predictions, without necessarily modelling how humans think. Although the two approaches are distinct, together they can enable human–AI complementarity by combining human-aligned reasoning with powerful data-driven insights.

In this Perspective, we outline a cognitive approach to human–AI complementarity in dynamic decision-making tasks. We begin by presenting the notion of cognitive AI and the role that it would have in supporting human decisions and fostering seamless collaboration with humans in teams. Then we discuss the need to develop dynamic models of the environment, the cognitive AI capabilities needed to advance AI, and how to realize the vision of cognitive AI for human–AI complementarity. Next, we highlight the broader societal and ethical implications of advancing cognitive AI and conclude with concrete recommendations to guide future work towards achieving human–AI complementarity.

We use the term human–AI complementarity here to describe systems that enable interactions at the cognitive level rather than interactions with the physical world[35]. Similarly, although work on human–robot interaction is discussed throughout the Perspective, the discussions of challenges regarding a physical environment or a robot's sensing capabilities[36] are outside the scope of cognitive AI.

Cognitive AI for human–AI complementarity

Our vision is to develop cognitive AI systems that emulate humans' cognitive information processing. Such cognitive AI systems would not only assist humans but would also be able to simulate human beings and act as teammates alongside humans. Cognitive AI would be capable of understanding and responding to human actions and engaging in meaningful interactions with humans, thus revolutionizing how humans communicate and collaborate with technology.

Cognitive AI aims to emulate and simulate the human mind as an information-processing system, similar to what was envisioned in the origins of AI[34,37,38], unified theories of cognition[39] and the beginnings of human–computer interaction[40,41]. This design enables AI systems to reason, adapt and interact in ways that align with how humans think, enabling greater transparency and interpretability. Unlike data-driven AI, cognitive AI seeks to explain and replicate how and why decisions are made, mirroring human mental functions rather than just matching outputs. Furthermore, cognitive AI must mimic human limitations and constraints, because doing so will allow us to interpret human behaviour and to predict (and prevent) human error.

Although this idea might appear anthropomorphic[42,43] inasmuch as it involves attributing human-like characteristics to AI, it is important to clarify that cognitive AI does not aim to replicate human identity or appearance. Rather, it draws on formal models of cognition to enhance reasoning, learning and decision-making in ways that are aligned with human cognitive processes. Although anthropomorphism can emerge in human users' perceptions of AI, especially in contexts where systems behave autonomously or use natural language[44,45], the goal of cognitive AI is not to elicit social attribution or human-likeness, but to support complementary roles in joint tasks. As such, the use of terms like 'teammate' or 'collaboration' refers to functional integration and shared goals, not to metaphorical or emotional resemblance. This distinction is critical to avoid conflating cognitively grounded modelling with anthropomorphic design intent[46].

Whereas the origins of AI and human–computer interaction are focused on a broad set of problems and the full range of capacities of the human mind, here we focus on cognitive AI as applied to dynamic decision-making. For example, managing resources during the real-time evolution of disasters like wildfires, hurricanes and floods is an extremely complex task. These situations demand decision-makers to evaluate tradeoffs among multiple alternatives and make choices under high uncertainty and with many constraints, such as limited time and resources[47–49]. It is critical to balance speed and accuracy (rapid decisions can save lives but acting too quickly on incomplete information can lead to misguided responses), to consider resource trade-offs (immediate solutions like temporary shelters can address short-term needs but be unsustainable in the long term), and to balance human safety against economic costs (evacuations can be necessary to save lives but disrupt communities and economies). These decisions often require coordination and collaboration among multiple decision-makers and poor teamwork can result in conflicts and inconsistencies across efforts.

In such decisions, data-driven AI acts as an assistant to human decision-makers, often by providing recommendations or aid through synthesis or explanations (Fig. 1a). For example, algorithms can analyse and synthesize large amounts of data such as weather forecasts and satellite images in real time[50,51]. Autonomous vehicles, such as drones and ground robots, can assist in search-and-rescue operations and access hard-to-reach and dangerous areas, reducing the risk to human rescuers[52]. Disaster robotics can be used to assist disaster managers and search-and-rescue missions[53–56]. Data-driven AI has been used to assess flood mapping and project impact and to verify inundation models, as well as to aid human decision-makers to optimize resource allocation and support logistics, and to help with long-term recovery efforts[57–59]. Although using data-driven AI as an assistant to human decisions has proved very valuable, it does not achieve human–AI complementarity[21,60]. In this AI-as-assistant approach, the AI does not need to fully represent a dynamic environment, nor need it have a dynamic mental model of the human. Instead, humans are tasked with maintaining a mental model of the environment and a mental model of the AI (for instance, how likely it is to err on specific cases) in order to use the AI's recommendations effectively[61]. Because the human is tasked with interpreting the outputs of the AI, the AI designers need to consider transparency, trustworthiness, accuracy, confidence and the interpretability of AI explanations. These aspects are critical to the adoption of AI-as-assistant[62,63].

In our approach to human–AI complementarity, cognitive AI serves as a dynamic computational representation of the mental model of the human, which can be used by data-driven AI for decision support (Fig. 1b), or to simulate a human partner in a human–AI team (Fig. 1c). Although both decision support and assistance involve helping users, they differ in purpose and complexity. Assistance helps humans to execute their own decisions, often by simplifying tasks or offering convenient options. By contrast, decision support enhances the quality of decision-making by providing relevant insights, interpreting data, evaluating alternatives and predicting outcomes.

Both of these aspects of our approach rely on the possibility of knowledge tracing using cognitive AI. Knowledge tracing is a method of estimating and updating a human's knowledge state over time based on their observable behaviours and responses in sequential tasks. This method[64] has now been used in multiple dynamic decision-making tasks to predict the decision a human will make at a given point in time, based on a history of actions taken previously[65,66]. Tracing individual human actions is similar to how intelligent tutoring systems support student learning, by mimicking the strategies and knowledge of human tutors and optimizing them for interaction with the learner[67,68]. In dynamic decision-making, cognitive AI provides a model of the human's mental state, which can then enable data-driven AI to fine-tune its use of large datasets and deliver more accurate, timely and personalized decision support.

Embedding human-like cognitive processes of cognitive AI into human–AI interactions enables data-driven AI to reason in ways that are aligned with human thought, facilitating mutual understanding and true complementarity. Because cognitive AI explicitly models the steps and constraints of human decision-making, it enhances the transparency and interpretability of AI behaviour. This alignment fosters trustworthiness, as the system's outputs can be traced back to human-relevant reasoning patterns, rather than to opaque statistical associations. In turn, users are more likely to have confidence in the AI's recommendations, and the ability to simulate individual mental models enables personalized decision support that improves both the accuracy and the relevance of AI assistance.

Data-driven AI can also use the mental model of the human provided by cognitive AI to change the choice architecture of a dynamic decision-making task (Fig. 1b). Choice architecture interventions involve re-structuring or altering the descriptions or other aspects of the choice options to influence decision-making[69,70]. Data-driven AI can adjust the choice options that are available to the decision-maker or alter the way the choice options are presented to the decision-maker[71]. It can also adjust the time and mode of presentation of those options, by using the model of the human provided by cognitive AI.

Thus, using cognitive AI for decision support enables human–AI complementarity by aiding human reasoning and judgement in cognitively demanding tasks, with the aim not just to help the human to act but to think along with them. This idea of using cognitive AI for decision support has been illustrated in cognitively demanding tasks, including specific applications like cyber deception training for detecting phishing emails[65,66,72]. However, substantially more research is needed

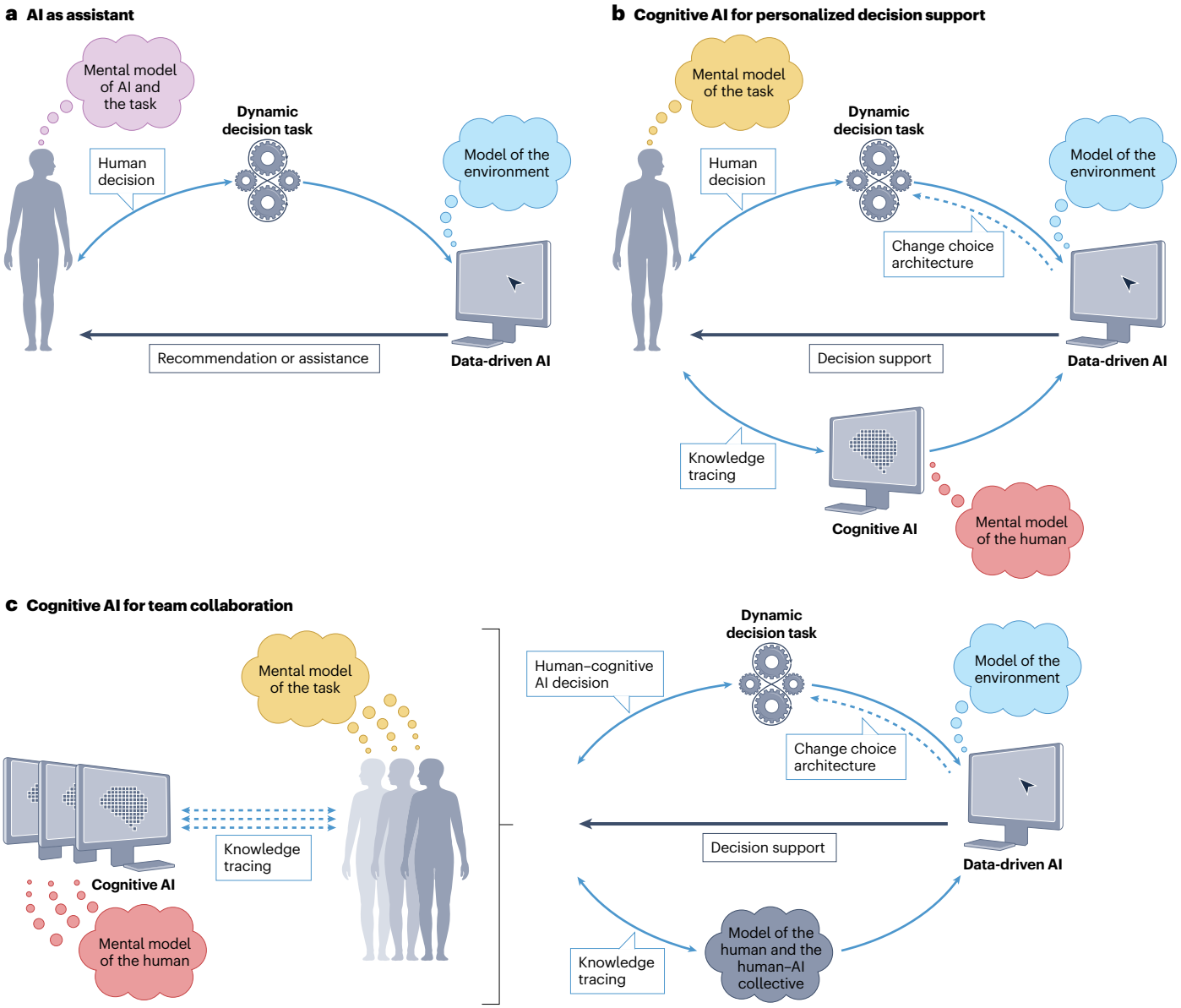

Fig. 1 | Types of human–AI complementarity. a, Data-driven AI is used as an assistant to provide recommendations and advice to the human decision-maker. **b**, Cognitive AI and knowledge-tracing methods are used to create a mental model of the human and provide data-driven AI with specific predictions of human actions. Using the input from cognitive AI, data-driven AI calibrates the recommendations and decision support or adjusts the choice architecture to support human decision-making. **c**, Cognitive AI acts autonomously as part of a team and can also apply knowledge tracing to generate mental models of specific humans in a team or of the team as a whole. Cognitive AI provides data-driven AI with a model of the human and a model of collective human and cognitive AI teamwork. Data-driven AI can calibrate its decision support towards the team as a whole or just to individual members of the team. It can also adjust the choice architecture in the environment for humans to improve future decisions.

to demonstrate its broader utility and to establish its role in enabling effective human–AI complementarity.

Cognitive AI can also emulate human cognitive decision processes and function as a human-like partner within a team (Fig. 1c). Teams are typically composed of heterogeneous agents, each with distinct roles, responsibilities and capabilities, who work interdependently towards a shared goal[43]. Interdependence is critical, creating the need for coordination, communication and joint activity among team members[73]. In dynamic decision-making tasks, outcomes are often achieved not by a single individual but through collaboration within a team of agents. Collaboration is the process by which agents — human or artificial — work together in a team to achieve a common goal. In this context, cognitive AI, like human teammates, can be designed to specialize in particular tasks or to operate effectively under specific conditions. Different cognitive AI teammates might take different roles, contributing their strengths when relevant, and stepping back when not,

thereby enhancing collective performance and enabling more adaptive, flexible and intelligent team behaviour[74–76]. To achieve human–AI collaboration in a team, cognitive AI would need not only the ability to act autonomously, to make decisions, to adapt to the environment and to improve its own performance over time[31,77]. It will also need to have the capability to engage in interdependent processes with other agents in the team to draw the allocation, update and retrieval of cognitive resources from other team members[78,79].

Using AI as a teammate is currently only an emerging idea[20,21,31], and its future development will require advanced definitions, formalizations, experiments and research progress in multiple disciplines as they relate to the design of AI and the study of team work and emergent collaboration and adaptation in hybrid teams. A central vision for this emerging area is to achieve effective human–AI complementarity in dynamic decision-making environments by advancing cognitive AI, not only for decision support, but also as a capable and adaptive teammate. Although cognitive AI can contribute in various ways, its ability to model human reasoning and adapt to a dynamic environment makes it especially well suited to collaborative roles that require coordination, shared understanding and mutual responsiveness.

AI models of dynamic environments

To deliver decision support, adapt the choice architecture and/or influence the behaviour of agents in a team, any AI (cognitive or data-driven) must maintain accurate representations of the changing environment. Dynamic decision-making environments are characterized by complexity, uncertainty and continuous change over time. In such settings, decisions must often be made with incomplete information, under time pressure, and in response to evolving goals or external conditions. The environment might be only partially observable, require multi-step planning and involve feedback loops in which earlier decisions influence future states. A model of the environment must therefore capture temporal dynamics and causal relationships, and the AI must be able to adapt to new or unforeseen information. Thus, AI systems must be updated in real time, accounting for human actions and changes in environmental demands. Here we review these capabilities in existing AI systems.

Reinforcement learning AI is designed to operate in dynamic environments and it is relatively well suited to maintaining a representation of such environments. For example, reinforcement learning agents learn to make decisions by interacting repeatedly with an environment[80]. Reinforcement learning agents take actions and observe the outcomes of those actions (in the form of rewards or penalties) and maintain a goal to maximize cumulative rewards over time. To this end, reinforcement learning agents learn about the reward corresponding to different courses of action in the environment (which necessitates a certain level of exploration or trial and error), while taking actions that guarantee high rewards (which necessitates exploitation, or using the information obtained so far to take the action that leads to high rewards). Using techniques like Q-learning[81,82], reinforcement learning agents gradually improve both their understanding of the environment and their decision-making strategy, aiming to find an optimal policy that leads to high long-term rewards in the environment. These agents maintain a model of the environment including the human in it and can adapt their behaviour on the basis of the feedback received from human users (through satisfaction scores or the rate at which a human follows their recommendations), which over time improves their ability to respond appropriately to human collaborators. For example, reinforcement learning agents are used to help robots to learn to perform tasks while interacting with their physical environment, to learn complex behaviours or to adapt to changing conditions. These agents enable robots to acquire new skills autonomously, such as walking, grasping objects or navigating through environments[83]. Thus, reinforcement learning agents are particularly suitable for dynamic, complex environments, including those that involve human interactions[84], with some important caveats.

At least three factors need to be addressed to achieve competent reinforcement learning models of dynamic decision environments. One factor is that reinforcement learning agents often require substantial amounts of training data (human interactions and interactions with the environment) to learn effectively. In environments in which each action has a real-world cost (for instance, resource allocation in disaster management), learning from real-world data can be impractical (if the data are scarce or expensive), which limits reinforcement learning speed and reliability. Second, training reinforcement learning agents in high-dimensional environments is exceedingly resource-intensive, requiring substantial computational power and time. Third, effective dynamic decision-making demands long-term planning: the ability to act under uncertainty and to perform well even when rewards are delayed or infrequent. In such settings, reinforcement learning agents often struggle to learn optimal decision-making policies, because their learning is highly dependent on the design, timing and structure of the reward function[85,86]. Crafting effective reward functions typically requires substantial domain knowledge: an expert understanding of the task, of the environment and of what constitutes desirable behaviour[87]. This domain knowledge is distinct from the knowledge in training data, which consists of statistical patterns observed during interaction or simulation. Even when domain knowledge is available, converting it into rewards that a reinforcement learning agent can learn from is often a complex and non-trivial task.

In addition to reinforcement learning, systems that use Bayesian inference offer a promising framework for learning in dynamic decision-making environments. By continuously updating beliefs about the state of the world based on incoming evidence, Bayesian methods enable agents to represent uncertainty, adapt to change and refine their understanding of the environment over time. For example, Bayesian learning from demonstration enables an agent to generalize from limited human demonstrations and dynamically incorporate new information to improve task performance[88,89]. This probabilistic, data-efficient approach is well suited to dynamic settings, because it supports ongoing learning, targeted information gathering by querying the human for the most informative data points, and the construction of flexible models of the environment that evolve from experience.

However, Bayesian models also face practical challenges, particularly in specifying prior distributions. A well known issue is that poorly chosen priors can lead to biased or unstable inferences, especially in complex or uncertain environments. Although some existing models use human input to calibrate priors, developing a principled, generalizable method for doing so remains difficult. This challenge is particularly relevant in the context of human–AI complementarity in dynamic decision-making, where priors must not only reflect the structure of the changing environment, but also align with human reasoning and adapt to evolving human inputs. The need for priors that are both context-sensitive and cognitively compatible makes this an open problem at the intersection of statistical learning and human-centred AI. Some models aim to integrate Bayesian approaches with reinforcement learning to enhance adaptability and uncertainty

in dynamic environments. For instance, techniques like dropout-based Bayesian approximation in deep learning demonstrate how uncertainty estimates can be incorporated into generative models, enabling more robust and calibrated predictions[90]. Similarly, Bayesian reinforcement learning methods emphasize the role of structured priors and inductive biases in learning and generalization[90,91].

In summary, some data-driven AI systems are appropriate for maintaining a dynamic representation of the changing environment. But these systems must advance their capabilities to achieve a representation of dynamic environments, including reducing the number of human interactions with the environment required to learn about the environment effectively, improving their ability to perform long-term dynamic planning in the absence of immediate observations of rewards, and defining a theoretical approach for generalizing priors for dynamic decision-making tasks.

Necessary AI capabilities

Beyond a model of the dynamic environment, there are other capabilities that cognitive AI and data-driven AI must have to achieve human–AI complementarity. Regardless of whether cognitive AI is used for decision support or as a teammate, cognitive AI must be competent and reliable; capable of human-accessible communication; and enable flexible interactions with humans and with data-driven AI (Table 1). These capabilities are discussed below, where we draw examples and implications from research across different types of AI.

Competence and reliability

Data-driven AI can demonstrate competence (the ability to perform a task effectively and accurately) and reliability (the consistency of performance over time and across tasks) in domains in which decision accuracy is easily quantified and optimal solutions are well defined. For example, data-driven AI typically performs well in static decision tasks that are well structured, have clear objectives and offer large datasets for pattern recognition[7]. These include applications such as image classification, optimization problems and data analytics, where AI systems can process large amounts of data to identify patterns that are often not evident to humans[92,93].

However, in dynamic decision-making environments, where accuracy is difficult to measure and optimal solutions might not exist, it becomes more challenging to assess these qualities. Data-driven AI demonstrates low competence in the absence of large amounts of data and in dynamic decision-making tasks that involve uncertainty, ambiguity, rapidly changing environments, delayed decision effects and time constraints[94]. In such cases, it is especially important for AI systems to recognize the limits of their knowledge and determine when to defer to human judgement.

Cognitive AI can support the competence and reliability of AI systems by applying metrics that are often used to evaluate human decision-making in dynamic tasks. These metrics include benchmarks or comparisons to expert judgements, calculated for each decision within a sequence[95]. In addition to outcome-based measures, studies emphasize the importance of process metrics, such as the sequence and timing of decisions, and the ability to adapt to changing circumstances[96,97]. Similar metrics could be applied to evaluate the competence and reliability of cognitive AI itself and to help to determine when human intervention might be necessary. Determining the competence and reliability of AI in dynamic tasks is essential for promoting its adoption in decision support and for ensuring the effective use of cognitive AI in team collaboration under dynamic conditions[14,98].

Although some AI systems have been created to be competent in specific dynamic – and somewhat ambiguous – tasks, they are often applicable only to those specific tasks. For example, AI is highly competent in dynamic tasks such as board games[99,100], the game Go[101,102] and computer games[103]. However, an AI system that performs well in one task (such as Go) does not perform well in other tasks (such as search and rescue operations)[104]. AI systems are currently created to be competent in a particular task configuration but they are not reliable, because they cannot easily generalize to new tasks or to changing configurations of the same task.

Some cognitive AI systems have aimed to address the characteristics of dynamic decision-making tasks, demonstrating human-like competence in their capability to maintain performance under changing configurations of the same task[105]. However, research is required to advance the competence of cognitive AI in combination with data-driven AI, and to achieve reliability across different tasks[106,107].

In summary, current AI systems lack the competence and reliability needed for effective human–AI complementarity in dynamic decision-making. Whereas data-driven AI excels in static, well defined tasks, it

Table 1 | Requirements for cognitive AI in dynamic environments

Property	Description	Flooding disaster scenario
Competence and reliability	Cognitive AI must maintain high and predictable performance in the presence of uncertainty and ambiguity, especially when decision rules or metrics are not clearly defined, and when data are sparse or unavailable. It must also support the development of tractable notions of decision quality in dynamic tasks	During a severe flood with disrupted communications, cognitive AI helps emergency managers to prioritize evacuation zones despite gaps in sensor data and rapidly shifting water levels. It adapts to limited information by using human inputs and learned patterns to make consistent recommendations and evaluates decision quality based on reducing exposure risk rather than fixed optimization targets
Human-accessible communication capabilities	Cognitive AI must communicate effectively with humans through understandable, transparent formats. Effective communication includes improving uncertainty quantification, enhancing interpretability, reducing misleading outputs and tailoring communication to different users	In coordinating flood relief, cognitive AI conveys projected inundation zones through interactive maps, provides confidence estimates for shelter accessibility, and communicates resource deployment options in plain language. It gives field responders concise visual cues while offering technical justification and uncertainty levels to command centre staff
Flexible interactions	Cognitive AI must represent and update models of human mental states, predict likely human actions, collaborate with data-driven AI systems, and support flexible team configurations. It should incorporate meta-cognitive processes and enable shared decision-making across human–AI teams	When responding to flooding in an urban area, cognitive AI monitors responder fatigue and task load, dynamically adjusting its support (for example, by proposing reassignment of personnel or shifting from autonomous decision-making to human oversight). It integrates with drone-based data-driven AI, anticipating human goals (like rescue priority) and adjusting its recommendations accordingly to fit evolving team roles

struggles with uncertainty, with changing conditions and with tasks in which optimal solutions are unclear. Cognitive AI offers a promising path by incorporating human-centred metrics of competence and reliability, but further development is needed to investigate how to generalize across tasks and to determine when to defer to human judgement.

Human-accessible communication

To achieve human–AI complementarity in dynamic decision-making environments, cognitive AI also needs to provide accessible ways to communicate with humans. Effective communication is essential for coordinating actions, sharing intentions and resolving misunderstandings, especially in environments in which decisions must be made quickly and under conditions of uncertainty. Without clear communication, humans might misinterpret or fail to trust AI recommendations, ultimately reducing team performance[108]. Prior work has proposed various approaches to close the communication gap between humans and AI, including uncertainty quantification[109], explanations[110,111], visualizations, and other media for communication (such as natural language).

Bayesian approaches[112], ensemble methods[113] and other techniques[114] quantify model uncertainty and are used to represent epistemic uncertainty — the uncertainty stemming from limited knowledge or information about the system or process in question. Although these methods offer a rigorous way to estimate confidence in model predictions, they are often computationally expensive and rely on ad hoc assumptions (such as the choice of prior distributions), which can make their outputs difficult for human decision-makers to interpret and rely on[115]. AI must be capable of conveying uncertainty in ways that are understandable to users, providing intuitive explanations or calibrated confidence estimates[116]. Research shows that communicating uncertainty clearly (through confidence intervals or probabilistic forecasts) can improve human trust, calibration and decision quality, particularly in high-stakes or ambiguous environments[16]. Thus, effective uncertainty representation might not only support transparency but also enhance AI adoption in human–AI teams.

Another approach to human-accessible communication is algorithmic transparency, including explainability and interpretability[62,98]. Explainability refers to the ability of AI to make its internal processes and decision logic understandable to human users, and especially how specific inputs influence outputs. Doing so is particularly important in complex models like deep neural networks, where the reasoning behind predictions is often opaque[117]. Explainability enables users to assess how and why a model arrived at a decision, which in turn supports trust, accountability and appropriate reliance on AI systems. Interpretability refers to the degree to which a human can understand the cause of a decision made by an AI model[62]. Although related, explainability goes a step further by providing a post hoc explanation that helps users to understand the reasoning behind the model's output, even when the model itself is too complex to interpret directly.

Given the black-box nature of many modern AI models (such as deep learning models with millions of parameters), a substantial body of research has proposed approaches for explainability[98,110,117–119]. For instance, local explanations aim to explain how an AI model arrived at a specific decision or prediction, identifying which features contributed most to an individual decision, making complex models more transparent[120,121]. Such techniques can improve human understanding of AI predictions and foster trust, which in turn supports better decision-making in collaborative human–AI settings[120,121]. For example, in medical diagnosis, local explanations can show why an AI predicted a specific disease for a specific patient by highlighting the most influential clinical features. However, the fidelity and useability of explainability methods have not been adequately established and they are being actively researched[122].

Other research suggests that it is important to create models that are interpretable in the first place and avoid the need for explainability[118]. Creating interpretable AI may involve incorporating human feedback to help to adapt and improve AI behaviour in ways that better align with human expectations and support trust over time[84]. Moreover, cognitive AI is inherently interpretable because it is built on transparent, theory-driven models of human reasoning and decision-making processes. These structured representations make it easier to trace how inputs lead to outputs, and they can be integrated with data-driven AI to enhance both explainability and interpretability.

Visualizations are an important form of structured explanations that make complex model behaviour more accessible to human users[123]. Visualizations like heat maps, feature importance graphs and decision trees help users to understand how input features influence model outputs[124]. For example, sensitivity analysis shows how model predictions change in response to variations in the input features, highlighting which variables matter most[124]. These visual tools support communication by making model reasoning more transparent, which in turn helps users to interpret, evaluate and trust AI recommendations.

Natural language is another human-accessible form of communication, which can be tailored to the needs of different human users. For example, a doctor might need a detailed, technical explanation of an AI-provided diagnosis including citations to reputable sources, whereas a patient might only need a simplified version (without technical jargon) to understand the diagnosis. Large language models (LLMs), a form of generative AI, can engage in natural language communication, providing a smooth integration with humans[125]. However, LLM agents can produce content that is factually incorrect, irrelevant, incoherent or misleading[126,127]. Because LLMs generate content based on statistical patterns rather than true understanding, they can produce outputs that seem plausible but are not appropriate[128]. Human over-reliance on such outputs can lead to costly mistakes. Moreover, it is often challenging for human teammates to understand why a generative AI agent makes certain decisions or recommendations owing to the opacity of the agent's decision-making process and internal models of the environment[84]. This lack of interpretability can pose challenges for human communication, in cases where understanding the rationale behind the teammates' decisions is critical for trust and appropriate reliance. These issues are specific to current LLMs and other generative AI approaches; and it might be possible to develop other forms of AI that communicate in natural language using structured, rule-based or cognitively grounded models that provide more reliable and interpretable outputs (for example, ref. 129).

In summary, cognitive AI needs to communicate in ways that are accessible to humans to enable human–AI complementarity. Doing so will require reduced computational costs and ad hoc assumptions for uncertainty quantification, improved approaches for explainability and transparency, and approaches that include human input to adapt and meet human expectations. In particular, LLMs must become more reliable in the accuracy of their responses and more adaptive to individual users, tailoring their communication style, content and level of detail to suit different users and contexts.

Flexible interactions

Humans, cognitive AI and data-driven AI must interact to reach decisions, and the nature of these interactions depends on how cognitive

AI is configured within the team. Cognitive AI can take on different roles, such as functioning as a decision-support tool or as a collaborative teammate, depending on the task demands and level of autonomy assigned. These are distinct use cases, but they do not necessarily require entirely separate systems; rather, a well designed cognitive AI system should be flexible enough to support both roles, adapting its level of engagement, communication and autonomy to fit the needs of the human team and the decision context.

A common configuration of human–AI interaction involves a single human working with a single AI agent. Even in this configuration, there are multiple ways in which humans and AI (whether cognitive or data-driven) can interact. Cognitive AI can be configured to act either for decision support, by using knowledge tracing and providing data-driven AI with a model of the human's mental state, or as a collaborative teammate alongside the human. In either role, there are multiple ways in which humans and AI (whether cognitive or data-driven) can interact[16]. First, AI can act as an advisor to the human, and the human then makes the final decision[16]. Second, humans can provide oversight of the AI decision[130,131]. Third, the human and the AI can make their own decisions independently and rely on a predefined aggregation function (for instance, a simple average, an uncertainty-weighted average, or an independent human referee) to combine those decisions into a final call[16,26]. Fourth, the human and the AI make their own decisions independently, but the AI additionally produces some auxiliary material characterizing its decision for the human (such as through explanations or uncertainty quantifications)[132]. This list is not an exhaustive set of possibilities even if we limit our attention to a single AI and a single human. In more complex human–AI teaming scenarios (Fig. 1c) that involve multiple humans and multiple cognitive AI agents collaborating on dynamic decision tasks, these interactions demand even greater flexibility. They require the team to continuously adapt roles, communication patterns and decision strategies to changing conditions and task requirements.

In summary, effective human–AI complementarity requires flexible interaction among humans, cognitive AI and data-driven AI, with cognitive AI configured to serve either as a decision-support tool or as a collaborative teammate. The structure of these interactions can vary widely, even within simple setups such as a single human working with a single AI agent. Interaction modes include advisory roles from AI without deciding, human oversight of AI decisions, independent decisions with aggregated outputs, and explanation-based support to inform human judgement. In more complex team settings that involve multiple humans and cognitive AI agents, these interactions must be more flexible and adaptive, requiring the team to adjust roles, communication styles and coordination strategies in real time as task demands evolve.

Realizing complementarity with cognitive AI

Initial efforts to develop cognitive AI for human–AI complementarity are ongoing and there is much work to build on. Existing efforts can be grouped broadly into three areas. First, researchers are integrating cognitive architectures with machine learning to create systems that are both human-aligned and computationally adaptive. Socio-cognitive frameworks model human roles, mental states and team structures to enable AI to function as collaborative teammates[133,134]; resource-rational approaches account for cognitive limitations by linking high-level goals to algorithmic constraints[135]. Second, a growing body of work focuses on fusing cognitive models with generative AI to combine the interpretability of human-like reasoning with the scalability of modern machine learning. These efforts include embedding generative models into cognitive architectures to support complex socio-cultural reasoning[136], enhancing cognitive models of decision-making with data-efficient generative techniques[137], and developing systems capable of integrating structured mental representations and adaptive generative mechanisms[138]. Third, cognitive AI is being applied to real-world tasks such as cyber defence, user training and behavioural intervention[139]. These systems simulate human responses to threats, anticipate behaviour in social-engineering scenarios, and design effective decision environments, which demonstrates their potential to enhance decision-making and collaboration in dynamic, high-stakes contexts[140]. Collectively, these efforts mark foundational steps towards building cognitive AI systems that support effective and adaptive human–AI teaming.

Despite these promising initial efforts, current approaches to developing AI for human–AI complementarity remain limited in several critical ways. Many systems focus narrowly on either emulating human behaviour or on optimizing task performance, without fully capturing the dynamic, individualized and context-sensitive nature of human cognition. There is often a lack of integration between cognitive representations and adaptive capabilities, resulting in AI systems that are either too rigid to generalize or too opaque to support effective collaboration. The vision for realizing human–AI complementarity with cognitive AI will address the ability to represent, adapt and reason with a human's mental model over time, which can be used either to emulate human behaviour or to perform knowledge tracing of a user's decisions over time to predict their future actions[77]. To serve this purpose, cognitive AI must be grounded in cognitive science principles, incorporate a certain amount of flexibility for variability in human behaviour, and have some level of autonomy in representing and updating models of the human and environment.

Cognitive science fundamentals

Cognitive AI cannot be developed simply by scaling up data-driven AI with more computing power or larger datasets[104]. Instead, cognitive AI requires fundamentally different architectures, which are grounded in cognitive science and aim to model the processes underlying human memory, learning and decision-making[77,141]. Building cognitive AI systems is a first step towards human–AI complementarity for dynamic decision-making[34,77].

An initial approach to generate cognitive AI is using cognitive architectures that intend to simulate human thought processes in a unified approach[41]. For instance, ACT-R (adaptive control of thought-rational) and SOAR (state, operator and result) are two cognitive architectures that represent human perception and action, memory, learning, problem-solving, decision-making and other capabilities[142,143]. The goal of cognitive architectures is to provide a comprehensive computational model of the human mind[39] and can inform the development of computational systems that align with human information processing by modelling key cognitive functions. When integrated with data-driven AI such as generative models or deep learning, these cognitive systems can be enhanced with greater scalability, pattern recognition and adaptability to complex environments[137,144]. This hybrid approach combines the structured reasoning and interpretability of cognitive architectures with the flexibility and data efficiency of data-driven AI.

Existing cognitive science approaches to modelling aspects of human cognition — particularly human decision-making in dynamic, uncertain environments — can be broadly classified into heuristic-based and learning-based systems. In dynamic decision-making

environments, humans often rely on cognitive heuristics: simplified decision rules that allow for efficient decisions without requiring full exploration of the environment. Heuristics reduce cognitive load by simplifying the processing of large amounts of information and by offering practical strategies for managing uncertainty in complex tasks[145]. Examples include heuristics like 'win-stay, lose-shift', the 'hot-stove effect' or 'probability matching', which guide behaviour on the basis of recent outcomes or the frequency of observed events[146,147]. However, heuristics are inherently imprecise and are often described descriptively rather than formalized computationally[148]. To evaluate their effectiveness in dynamic decision-making tasks, formal computational models of these heuristics must be developed and tested for complex, evolving tasks[148]. Although many researchers have proposed formal models of heuristics, such as lexicographic rules[149], or elimination-by-aspect[150], these models are relatively rare in psychology, especially in applications to complex, dynamic environments where human decision-making unfolds over time[151,152]. For the development of cognitive AI, heuristics offer valuable insights into human-like strategies, making them an important foundation for building interpretable and adaptive AI systems.

Cognitive learning agents are intended to learn sequential decisions from experience by updating their behaviour based on feedback over time. These agents are often grounded in reinforcement learning or Bayesian learning frameworks, which provide mechanisms for adapting to dynamic environments[85,91]. For example, a key distinction in reinforcement learning between model-free and model-based approaches illustrates two different strategies for dynamic decision-making[153]. Model-free agents learn action values directly from experience without constructing an internal model of the environment, whereas model-based agents build and update such a model to simulate future outcomes, so that in this way their process resembles deliberate decision-making. This distinction captures important aspects of human cognition and serves as a foundation for developing cognitively plausible learning agents. Other cognitive learning agents aim to explain the cognitive process by which humans make decisions in dynamic tasks[96]. For example, instance-based learning theory is an approach that aims to mimic human decision-making in dynamic tasks by using past experiences to inform current decisions[77,96]. It posits that people rely on a combination of specific instances or examples from memory, rather than on abstract rules or generalized knowledge, to make decisions in dynamic and uncertain environments. This approach aligns with how humans often recall and use past experiences in real life to solve new problems[154]. Cognitive AI systems can rely on these computational learning models of human-like cognition to represent the human's mental model of the dynamic environment and the human's own intentions with regard to decision-making. Taken together, these learning-based approaches offer a promising foundation for building adaptive, interpretable and human-aligned cognitive AI systems that can function in dynamic decision settings.

Role flexibility

Cognitive AI can take definite roles in a team and interdependencies with other team members can be clearly defined, in order to pursue collaboration towards a common goal. Cognitive AI can be used by powerful data-driven AI to personalize and time the decision support to the human[65,66,155]. Furthermore, cognitive AI can also be a teammate to humans, contributing to dynamic tasks as part of a team in which creativity, collaboration and diverse actions lead to better outcomes. A team configuration defines the roles and interdependencies of the human and the AI in a collaborative team, and it should leverage the complementary strengths of humans and AI to optimize the team decision process.

AI that is optimal for independent decision-making might not be the best teammate[27,156]. To use cognitive AI as a teammate, AI systems need to adjust their decision support to align with both individual and collective human mental models. To ensure effective collaboration in the team, there must also be a clear mechanism for identifying and resolving conflicts or discrepancies between human judgements and cognitive AI decisions. For example, human–AI complementarity can emerge when cognitive AI approaches a task using a different strategy than the human teammate[157]. In such cases, rather than requiring perfect agreement, the system can flag the divergence, explain the reasoning in human-understandable terms, and support the human in reevaluating their assumptions. Doing so enables both agents to benefit from their distinct perspectives, which leads to more robust decisions that reflect both experience-based intuition and systematic reasoning.

Given the high uncertainty and unpredictability of dynamic decision-making tasks, it is important that interactions between humans and AI are flexible, so that humans can adapt roles or override AI when needed. A flexible configuration will also enable the human to adjust the AI role as their trust improves or assign tasks to the teammate that is best equipped to handle the task.

Autonomy and mental models

High levels of autonomy and self-directed behaviour are desirable in cognitive AI. Autonomy is needed to enable the AI to act as a teammate and collaborate with human partners in complex, rapidly changing scenarios. Cognitive AI would complement human capabilities by adapting to new information as it arrives from the environment thanks to its dynamic model of the environment and the dynamic human–AI mental model. When working in a team, cognitive AI should be capable of acting autonomously and working in capacities similar to those of their human partners to collaborate towards a common goal.

Cognitive AI would maintain a mental model of the human's desires and preferences that could be used to anticipate the human's needs and challenges, interpret their intentions, and explain human behaviours. Shared mental models are the knowledge structures that are common to all members of human teams. These shared mental models enable humans to collaborate and coordinate effectively in dynamic environments[158]. Humans create, store and manipulate the internal models of the dynamic systems with which they interact — such as flight operations in aviation, command and control systems in military settings, or emergency response procedures in crisis management — as team members must continuously update their understanding of goals, roles and environmental conditions to act effectively as a unit[159]. Humans also develop a mental model of their AI partners, how they work and the expectations regarding their behaviour[21]. These human–AI shared mental models are not static, but rather evolve dynamically as team members interact and adapt to change with technology[160].

However, the shared mental model formed between humans and AI, which encompasses the dynamic system, team structure, team roles, individual capabilities and other relevant features, can have a crucial role in enabling collaboration with cognitive AI in team settings (Fig. 1c). To function effectively within a human–AI team, cognitive AI will need to understand the shared team goals and its own role within the team. One formulation of this concept originates in studies of collective intelligence and is captured by the transactive

systems framework, which can be formalized for computational implementation[78,79,133]. In this framework, team members (human or artificial) maintain awareness not only of their own knowledge and capabilities but also of what others in the team know and can do. This 'who knows what' structure enables efficient information sharing, task allocation and coordination. Applying this framework to AI systems means designing them to represent and update knowledge about their teammates' expertise, responsibilities and roles. For cognitive AI, doing so could involve creating internal models of human teammates' mental states and capabilities, enabling the system to anticipate when to offer support, when to defer and how to contribute more effectively to collective goals. Embedding such transactive memory mechanisms into cognitive AI would enhance its ability to act as an integrated, adaptive member of human–AI teams.

The concept of a shared mental model is related to theory of mind: the human capability of understanding the beliefs and desires of others. Computational models of theory of mind have been of great interest in the computational cognitive sciences, where they have been used to predict human actions and assist in collaborative tasks[135,161–163]. Models such as Bayesian theory of mind[161,162,164] aim to infer a human's mental state from observed behaviour, enabling artificial agents to adapt their responses in socially and contextually appropriate ways. Bayesian models have also been extended to support sequential decision-making by incorporating hierarchical structures that reflect how knowledge might be organized across levels of abstraction. In hierarchical Bayesian models, the learning process occurs at multiple levels of abstraction, such as estimating low-level task parameters for specific actions and higher-level strategies. This layered representation enables the model to adapt flexibly to new and rapidly changing environments by generalizing from past experience while remaining sensitive to context-specific variations[165].

Human theory of mind involves multiple cognitive and social capacities and further specificity and formalization of these capacities are needed to develop them effectively in AI systems. For example, social intelligence capabilities can be designed to enable AI to engage human users in more natural and contextually appropriate conversations[166]. Similarly, techniques such as affective computing and sentiment analysis can help to advance cognitive AI by enabling it to detect human emotional states from text, voice or facial expressions and adjust its responses accordingly to support more empathetic and effective interaction[167].

Cognitive AI should help to maintain a dynamic mental model of the environment that reflects how humans perceive, interpret and reason about complex environments over time. Unlike data-driven AI models that often represent the environment statistically or algorithmically, a cognitive AI system must capture the mental model as experienced by the human, including key decision-making variables, perceived cause-and-effect relationships, feedback loops that represent causality, and emergent dynamics within the environment[168,169]. Moreover, cognitive AI should be capable of adapting to novel tasks and personalizing its reasoning and support according to its understanding of each human teammate's goals, prior experiences and preferences. This personalization enables more effective and context-sensitive collaboration within human–AI teams, by providing data-driven AI with predictions about human decisions so that it can adjust the recommendations, environment and interventions according to the changing preferences of humans over time[65,66,155,170]. This capability could be realized through techniques such as reinforcement learning, in which the system learns optimal actions by receiving feedback from human users, and meta-learning, in which the system learns to learn from a variety of tasks[171]. However, in contrast to reinforcement learning and other data-driven approaches, cognitive AI can generate theory-driven predictions even in the absence of empirical training data[172,173]. These predictions are possible because cognitive models are grounded in formal theories of human cognition that describe how decisions are made on the basis of mechanisms such as memory retrieval, similarity and experience-based reasoning. For example, instance-based learning models can make accurate predictions about human decision-making without being trained on human data, and use cognitive principles to simulate sequential decisions[173].

In summary, the development of cognitive AI represents a transformative step towards creating systems that can either emulate human decision-making or function as teammates alongside humans. By integrating insights from cognitive science and leveraging advanced data-driven AI, these systems will be able to emulate human-like thought processes, adapt dynamically to changing environments, and maintain the shared mental models that are critical for effective collaboration. Beyond merely responding to commands, cognitive AI will predict human needs, engage in meaningful interactions, and assist in decision-making, all while fostering trust and transparency. This ambitious vision seeks to merge theoretical foundations with practical innovations, ensuring that cognitive AI enhances human capabilities and complements human judgement in complex, dynamic scenarios. These ideal features of cognitive AI are applicable to any dynamic decision-making tasks[174], such as disaster management (Box 1).

Ethical considerations and risks

The development of cognitive AI for human–AI complementarity requires careful consideration of human values. Data-driven AI systems lack the ability to reason about ethical and societal implications of their decisions in novel environments, but this ability is key to the trustworthiness of AI teammates[175,176]. Moreover, cognitive AI systems can have major long-term ramifications for society; therefore, responsible conduct of research in this field requires careful consideration of the societal effect of deploying such AI at scale in consequential domains.

Ethical competence

Some of the key ethical principles that are usually invoked when AI is used to make or assist high-stakes decisions are beneficence, fairness and justice, transparency and privacy[177]. Accounting for these principles is the minimum requirement for ethical competence – the ability to identify and understand relevant values and ethical principles, recognize ethical conflicts and make decisions that align with those principles and values. Data-driven AI agents lack capabilities for moral agency and responsibility and cannot take intentional action[35], but our vision of cognitive AI requires ethical competence.

Beneficence requires establishing that cognitive AI contributes positively to the life plans and wellbeing of individuals and communities who are affected by it while upholding their fundamental rights[178,179]. For example, beneficent cognitive AI should have an accurate mental model of their teammates– including what contributes to the wellbeing of their human teammates and how the AI can benefit the teammates.

Fairness and justice prohibit discrimination against individuals or groups based on characteristics such as race, gender, age, socioeconomic status or other protected attributes[180,181]. To be considered fair, cognitive AI should have an accurate model of the decision-making environment in which it operates, including the social and political context that renders certain individuals and communities disadvantaged.

Box 1 | Human–AI complementarity in a disaster management scenario

This case study outlines the ideal capabilities of cognitive AI in a dynamic decision-making context, using the example of a major urban flood. These capabilities demonstrate how cognitive AI could support human–AI complementarity through autonomy, communication, adaptation and social understanding.

Autonomous operations and self-directed behaviour

- Monitor environmental and sensor data (such as rainfall, river levels and road closures) to detect flood risks in real time
- Assess the vulnerability of affected areas and predict the effect on infrastructure and population
- Develop evacuation plans and direct autonomous vehicles to transport at-risk residents
- Coordinate drone and robot deployment to distribute emergency supplies such as water, medical kits and food to stranded populations
- Initiate structural assessments and coordinate robotic repairs to critical infrastructure, such as bridges or levees, when human access is unsafe

Human-like communication with diverse users

- Communicate with the public using multimodal methods, including spoken language, text messages, visual displays and social media updates
- Provide clear and adaptive messaging that varies in technical detail depending on the audience (for instance, for emergency responders, government officials or local residents)
- Broadcast evacuation updates across multiple platforms (such as mobile alerts, radio and social media) to maximize reach
- Engage in two-way communication, enabling residents to report their locations, health status or access needs, which cognitive AI uses to adjust evacuation routes and supply distribution in real time

Social intelligence, personalization and shared mental models

- Personalize evacuation instructions by accounting for household characteristics (such as age, mobility and access to transportation)
- Tailor alerts and resources for vulnerable groups (such as older people, people with disabilities or non-native language speakers)
- Prioritize the deployment of emergency services to densely populated or high-need areas based on social, geographic and economic data
- Plan infrastructure recovery efforts by estimating the social and economic impact of damage, and scheduling repairs accordingly

Real-time adaptation to a changing environment

- Dynamically update its mental model of the disaster as new information (such as weather forecasts or floodwater movements) becomes available
- Reassign response resources, such as drones or rescue teams, based on shifting conditions, such as a levee breach or newly flooded area
- Generate detailed, up-to-date damage assessments using aerial drone footage and sensor data, supporting rapid decision-making by emergency planners

Transparency requires clarity and openness surrounding decisions and requires that this information is provided in an accessible manner to AI stakeholders, including developers, users and regulators[119,182]. Human-accessible communication and flexible interactions with human teammates are a hallmark of ethically competent cognitive AI.

Finally, developing ethically competent cognitive AI requires vast amounts of personal and sensitive data from human teammates and the environment. Privacy requires that these data are protected in compliance with privacy rights and to maintain confidentiality, security and appropriate use[183]. Trustworthy cognitive AI agents must be able to identify and protect sensitive data in a contextually appropriate manner.

Researchers have attempted to design data-driven AI that is compliant with the above values, although different stakeholders might prioritize these values differently depending on the context[184]. For example, AI developers tend to prioritize values like transparency and technical robustness, whereas members of the public are more likely to emphasize fairness and protection from harm[184]. Furthermore, there are several barriers to operationalizing these values in practice, including a lack of organizational incentives and accountability structures, misalignment between ethical goals and business metrics, limited resources and tooling for ethical AI development, and insufficient cross-functional collaboration. Even when practitioners are motivated to build responsible AI, these structural and cultural challenges within organizations often prevent meaningful change. Addressing these barriers is essential before ethically competent cognitive AI systems can be developed[185].

Long-term ramifications

If not designed and deployed carefully, cognitive AI presents substantial long-term risks and ramifications for human society. Many of these concerns are relevant for data-driven AI agents, even if they do not emulate human cognition[177,185]. But cognitive AI that closely emulates human capabilities exacerbates these risks because such systems might be perceived as more trustworthy, leading to overreliance, misinterpretation of intent and diminished human oversight when cognitive AI acts in a fully autonomous manner. Moreover, the ability to mimic human-like reasoning can obscure accountability and raise more acute ethical concerns around manipulation, bias replication and decision opacity.

Training and deploying all kinds of AI consumes vast amounts of energy and the increased demand for powerful processors and data centres could result in substantial environmental degradation, including deforestation, water pollution and habitat destruction[186,187]. Also, AI that is capable of performing many complex tasks as efficiently as humans without suffering from the same physiological limitations and psychological biases could reduce the demand for human labour and change the nature of work[188,189]. It is also possible that human decision-makers could become overly dependent on cognitive AI[190,191], leading to a decline in human critical thinking and problem-solving skills[192]. For instance, as people become less accustomed to storing and

retrieving information themselves, this process could weaken memory retention and recall abilities over time[193]. Furthermore, as cognitive AI requires the maintenance of an accurate model of humans, these systems will need to monitor human decision-makers in real time, leading to increased surveillance, micromanagement and potentially manipulation. Over time, human decision-makers might feel a loss of control or agency in their work, leading to disengagement and dissatisfaction[194].

Researchers and developers of cognitive AI must be cognizant of the above issues and proactively seek to mitigate them to ensure that their work serves humanity in the long term and does not create such unintended, yet serious, risks. Accounting for the ethical and long-term societal implications of developing highly capable AI is particularly critical if it is to be deployed at scale in high-stakes domains. In addition to the widely applicable issues discussed here, real-world evaluation and deployment requires domain-specific ethical and societal considerations to be accounted for. For instance, in dynamic disaster management scenarios, cognitive AI systems that assist with evacuation planning or resource allocation must be designed to avoid reinforcing pre-existing social inequities, for example, by ensuring that marginalized communities are not deprioritized owing to biased data or assumptions embedded in the model. Researchers must pay special attention to justifying any claims of generalizability of their findings to other domains.

In conclusion, although this Perspective focuses primarily on the computational and cognitive foundations required to enable human–AI complementarity, we acknowledge that the development of autonomous cognitive AI systems raises important ethical questions. The increasing sophistication of AI systems that aim to emulate aspects of human cognition, such as reasoning, memory and learning, blurs the boundaries between tool and collaborator and demands careful ethical scrutiny[195]. Questions about accountability, agency and the moral status of AI systems arise when these systems operate with high autonomy and social intelligence, particularly in high-stakes domains. Although our goal is not to replicate human consciousness or identity, but to functionally model cognition for collaborative utility, we recognize that such capabilities might elicit anthropomorphic perceptions and expectations that have real social and psychological implications[45,46]. A comprehensive ethical analysis of these issues is beyond the scope of this paper, but cognitive AI development should be accompanied by ongoing, interdisciplinary discussions that include cognitive scientists, ethicists and policymakers. We hope that our work contributes to grounding those discussions in both technical feasibility and theoretical clarity, expanding the space of ethical inquiry around human–AI interactions.

Conclusions

We have described a path towards human–AI complementarity in dynamic decision-making environments, where cognitively grounded AI is critical to enhancing human capability through integration with data-driven AI technologies. Realizing this vision will require sustained collaborative efforts across disciplines. We identify four key areas for immediate action: developing infrastructure, advancing cognitive AI capabilities, empirically evaluating of human–AI complementarity, and addressing the ethical and societal implications of cognitive AI.

First, there is an urgent need for open-source simulation platforms that support cognitive AI research and evaluation of human–AI teams in varied, interactive decision-making contexts. Existing platforms often lack support for multi-agent hybrid human–AI interaction or are tailored to narrow use cases[196–198]. More flexible, general-purpose environments are needed to test, refine and compare cognitive AI models in dynamic, team-based scenarios[199,200].

Second, research must continue to enhance both the cognitive plausibility and adaptive functionality of cognitive AI systems. This includes progress in cognitive architectures, computational theories of human decision-making and hybrid models that combine symbolic reasoning with data-driven learning. Especially promising are adaptive mechanisms that infer human mental models and tailor AI behaviour accordingly in real time[65,66,71,155].

Third, experimental studies are needed to assess how cognitive AI systems function in team settings with human collaborators. Current research often relies on 'Wizard of Oz' methods, in which a human secretly controls the AI behind the scenes to simulate advanced capabilities that AI does not yet fully possess, or uses very simplified scenarios with clearly defined options and static decision options that do not reflect the ambiguity, time pressure and interdependence typical of real-world settings[4,31,201,202]. Although such methods have enabled researchers to study human responses to future AI behaviour and are valuable for isolating variables or testing early hypotheses, they might not capture the full range of challenges and dynamics that emerge in authentic human–AI collaboration. As a consequence, there remains a critical gap in understanding how cognitive AI systems perform in complex, high-stakes and evolving team environments. Future studies should test real cognitive AI agents across tasks, measuring outcomes such as decision quality, user confidence and collaborative fluency, particularly in high-stakes, dynamic environments.

Fourth, as cognitive AI systems grow more capable, understanding their social and ethical effects has become vital. Although existing research has begun to explore trust, acceptance and moral tradeoffs within human–AI interactions[180,183,203], the field lacks solid research on how cognitive AI might influence ethical reasoning or long-term human wellbeing. For instance, a cognitive AI system used in healthcare resource allocation could influence an emergency manager's ethical decisions by consistently framing allocation choices in ways that prioritize efficiency over fairness, potentially reshaping the human's moral reasoning over time. These dimensions — such as influence on human values, fairness and long-term dependence — must be integral to system design and evaluation.

By addressing the above challenges, the field can move towards truly complementary human–AI teams that can improve decision-making while upholding human values. The ultimate goal should not only be more effective decisions, but also systems that protect, enhance and empower human agency and wellbeing in complex environments.

Published online: 17 October 2025

References

1. Klein, G. A. *Sources of Power: How People Make Decisions* (MIT Press, 2017).
2. Edwards, W. Dynamic decision theory and probabilistic information processings. *Hum. Factors* **4**, 59–74 (1962).
3. Gonzalez, C., Fakhari, P. & Busemeyer, J. Dynamic decision making: learning processes and new research directions. *Hum. Factors* **59**, 713–721 (2017).
4. Tejeda, H., Kumar, A., Smyth, P. & Steyvers, M. AI-assisted decision-making: a cognitive modeling approach to infer latent reliance strategies. *Comput. Brain Behav.* **5**, 491–508 (2022).
5. Inkpen, K. et al. Advancing human-AI complementarity: the impact of user expertise and algorithmic tuning on joint decision making. *ACM Trans. Comput. Hum. Interact.* **30**, 1–29 (2023).
6. Halevy, A., Norvig, P. & Pereira, F. C. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
7. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition* (Pearson Education, 2016).
8. Leenes, R. In *Regulating New Technologies in Uncertain Times* (ed. Reins, L.) 3–17 (TMC Asser Press, 2019).

9. Lin, P. in *Autonomous Driving* (eds Maurer, M. et al.) 69–85 (Springer, 2016).
10. Bien, N. et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med.* **15**, e1002699 (2018).
11. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
12. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
13. Green, B. & Chen, Y. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proc. ACM Human-Computer Interact.* **5**, 148 (2021).
14. Amodei, D. et al. Concrete problems in AI safety. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1606.06565 (2016).
15. Boden, M. et al. Principles of robotics: regulating robots in the real world. *Conn. Sci.* **29**, 124–129 (2017).
16. Li, J., Yang, Y., Liao, Q. V., Zhang, J. & Lee, Y.-C. As confidence aligns: understanding the effect of AI confidence on human self-confidence in human-AI decision making. In *Proc. 2025 CHI Conf. on Human Factors in Computing Systems*, 1–16 (ACM, 2025).
17. Feil-Seifer, D. & Mataric, M. Socially assistive robotics. *IEEE Robot. Autom. Mag.* **18**, 24–31 (2011).
18. Papadopoulos, I. et al. A systematic review of the literature regarding socially assistive robots in pre-tertiary education. *Comput. Educ.* **155**, 103924 (2020).
19. Dellermann, D. et al. The future of human–AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2105.03354 (2021).
20. O'Neill, T., McNeese, N., Barron, A. & Schelble, B. Human–autonomy teaming: a review and analysis of the empirical literature. *Hum. Factors* **64**, 904–938 (2022).
21. Steyvers, M. & Kumar, A. Three challenges for AI-assisted decision-making. *Persp. Psychol. Sci.* https://doi.org/10.1177/17456916231181102 (2023).
22. Carroll, M. et al. (2019). On the utility of learning about humans for human-ai coordination. In *33rd Conf. on Neural Information Processing Systems*, https://proceedings.neurips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html (NeurIPS, 2019).
23. Strouse, D., McKee, K. R., Botvinick, M., Hughes, E. & Everett, R. Collaborating with humans without human data. In *Proc. 35th Int. Conf. Neural Information Processing Systems (NIPS '21)* 14502–14515 (ACM, 2021).
24. Ying, L. et al. Assessing adaptive world models in machines with novel games. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2507.12821 (2025).
25. Collins, K. M. et al. Building machines that learn and think with people. *Nat. Hum. Behav.* **8**, 1851–1863 (2024).
26. Rastogi, C., Leqi, L., Holstein, K. & Heidari, H. A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.10806 (2022).
27. Bansal, G., Nushi, B., Kamar, E., Horvitz, E. & Weld, D. S. Is the most accurate AI the best teammate? Optimizing AI for teamwork. *Proc. AAAI Conf. Artif. Intell.* **35**, 11405–11414 (2021).
28. Donahue, K., Chouldechova, A. & Kenthapadi, K. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *2022 ACM Conf. on Fairness Accountability and Transparency*, 1639–1656 (ACM, 2022).
29. Steyvers, M., Tejeda, H., Kerrigan, G. & Smyth, P. Bayesian modeling of human–AI complementarity. *Proc. Natl Acad. Sci. USA* **119**, e2111547119 (2022).
30. Vaccaro, M., Almaatouq, A. & Malone, T. W. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 2293–2303 (2024).
31. McNeese, N. J., Demir, M., Cooke, N. J. & Myers, C. Teaming with a synthetic teammate: insights into human-autonomy teaming. *Hum. Factors* **60**, 262–273 (2018).
32. McNeese, M. & Mcneese, N. J. Humans interacting with intelligent machines: at the crossroads of symbiotic teamwork. *Living Robots* https://doi.org/10.1016/B978-0-12-815367-3.00009-8 (2020).
33. Cabrera, ÁA., Perer, A. & Hong, J. I. Improving human–AI collaboration with descriptions of AI behavior. *Proc. ACM Human- Computer Interact.* **7**, 136 (2023).
34. Simon, H. in *An Overview of Machine Learning* (eds Carbonell, J. G. et al.) 25–37 (Springer, 1983).
35. Evans, K. D., Robbins, S. A. & Bryson, J. J. Do we collaborate with what we design? *Top. Cogn. Sci*. https://doi.org/10.1111/tops.12682 (2023).
36. Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A. & Alami, R. Artificial cognition for social human–robot interaction: an implementation. *Artif. Intell.* **247**, 45–69 (2017).
37. Simon, H. A. Cognitive science: the newest science of the artificial. *Cogn. Sci.* **4**, 33–46 (1980).
38. Simon, H. Artificial intelligence: an empirical science. *Artif. Intell.* **77**, 95–127 (1995).
39. Newell, A. *Unified Theories of Cognition* (Harvard Univ. Press, 1994).
40. Card, S. K., Moran, T. P. & Newell, A. *The Psychology of Human–Computer Interaction* (CRC Press, 2018).
41. Card, S. K. & Newell, A. in *Human Performance Models for Computer-Aided Engineering* (eds Elkind, J. I. et al.) 173–179 (Elsevier, 1990).
42. Cohen, M. C., Mancenido, M. V., Chiou, E. K., & Cooke, N. J. Teamness and trust in AI-enabled decision support systems. In *2nd Int. Conf. Hybrid Human-Artificial Intelligence (HHAI-WS)* (ed. Hirzle, P. K. M. et al.) Vol. 2456, 175–187 (CEUR, 2023).
43. Cooke, N. J. et al. From teams to teamness: future directions in the science of team cognition. *Hum. Factors* **66**, 1669–1680 (2024).
44. Ciechanowski, L., Przegalinska, A., Magnuski, M. & Gloor, P. In the shades of the uncanny valley: an experimental study of human–chatbot interaction. *Future Gener. Comput. Syst.* **92**, 539–548 (2019).
45. Epley, N., Waytz, A. & Cacioppo, J. On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* **114**, 864–886 (2007).
46. Marakas, G. M., Johnson, R. D. & Palmer, J. W. A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *Int. J. Hum. Comput. Stud.* **52**, 719–750 (2000).
47. Gonzalez, C. & Brunstein, A. Training for emergencies. *J. Trauma* **67**, S100–S105 (2009).
48. Comfort, L. K., Boin, A. & Demchak, C. C. *Designing Resilience: Preparing for Extreme Events* (Univ. Pittsburgh Press, 2010).
49. Schade, C., Kunreuther, H. & Koellinger, P. Protecting against low-probability disasters: the role of worry. *J. Behav. Decis. Mak.* **25**, 534–543 (2012).
50. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
51. Rolnick, D. et al. Tackling climate change with machine learning. *ACM Comput. Surv.* **55**, 42 (2023).
52. Adams, S. M., & Friedland, C. J. A survey of unmanned aerial vehicle (UAV) usage for imagery collection in disaster research and management. In *Proc. 9th International Workshop on Remote Sensing for Disaster Response* vol. 8, 1–8 (2011).
53. Murphy, R. R. *Disaster Robotics* (MIT Press, 2017).
54. Casper, J. & Murphy, R. R. Human–robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Trans. Syst. Man. Cybern. B* **33**, 367–385 (2003).
55. Murphy, R. et al. Two case studies and gaps analysis of flood assessment for emergency management with small unmanned aerial systems. In *Int. Symp. Safety, Security, and Rescue Robotics (SSRR)* 54–61 (IEEE, 2016).
56. Murphy, R. R. & Tadokoro, S. User interfaces for human–robot interaction in field robotics. In *Disaster Robotics: Springer Tracts in Advanced Robotics* (ed. Takadoro, S.) 507–528 (Springer International, 2019).
57. Sun, W., Bocchini, P. & Davison, B. D. Applications of artificial intelligence for disaster management. *Natural Hazards* **103**, 2631–2689 (2020).
58. Abid, S. K. et al. Toward an integrated disaster management approach: how artificial intelligence can boost disaster management. *Sustainability* **13**, 12560 (2021).
59. Tan, L., Guo, J., Mohanarajah, S. & Zhou, K. Can we detect trends in natural disaster management with artificial intelligence? A review of modeling practices. *Natural Hazards* **107**, 2389–2417 (2021).
60. Fok, R. & Weld, D. S. In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. *AI Mag*. https://doi.org/10.1002/aaai.12182 (2024).
61. Endsley, M. R. From here to autonomy: lessons learned from human-automation research. *Hum. Factors* **59**, 5–27 (2017).
62. Gilpin, L. H. et al. Explaining explanations: an overview of interpretability of machine learning. In *5th Int. Conf. Data Science and Advanced Analytics (DSAA)* 80–89 (IEEE, 2018).
63. Bach, T. A., Khan, A., Hallock, H., Beltrão, G. & Sousa, S. A systematic literature review of user trust in AI-enabled systems: an HCI perspective. *Int. J. Human-Computer Interact.* **40**, 1251–1266 (2024).
64. Corbett, A. T. & Anderson, J. R. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-adapted Interact.* **4**, 253–278 (1995).
65. Cranford, E. A. et al. Personalized model-driven interventions for decisions from experience. *Top. Cogn. Sci*. https://doi.org/10.1111/tops.12758 (2024).
66. Cranford, E. A. et al. Toward personalized deceptive signaling for cyber defense using cognitive models. *Top. Cogn. Sci.* **12**, 992–1011 (2020).
67. Anderson, J. R., Boyle, F. & Yost, G. The geometry tutor. In *Proc. Int. 9th Joint Conf. Artificial Intelligence (IJCAI '85)* Vol. 1 (ed. Joshi, A.) 1–7 (Morgan Kaufman, 1985).
68. Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. Cognitive tutors: lessons learned. *J. Learn. Sci.* **4**, 167–207 (1995).
69. Johnson, E. J. et al. Beyond nudges: tools of a choice architecture. *Market. Lett.* **23**, 487–504 (2012).
70. Thaler, R. & Sunstein., C. *Nudge: Improving Decisions About Health, Wealth and Happiness* (Yale Univ. Press, 2008).
71. Callaway, F., Hardy, M. & Griffiths, T. L. Optimal nudging for cognitively bounded agents: a framework for modeling, predicting, and controlling the effects of choice architectures. *Psychol. Rev*. https://doi.org/10.31234/osf.io/7ahdc (2022).
72. Malloy, T., Ferreira, M. J., Fang, F. & Gonzalez, C. Training users against human and GPT-4 generated social engineering attacks. In *International Conf. on Human-Computer Interaction* (ed. Moallem, A.) 54–74 (Springer Nature Switzerland, 2025).
73. Johnson, M. & Bradshaw, J. M. in *Engineering Artificially Intelligent Systems (Lecture Notes in Computer Science Vol. 13000)* (eds Lawless, W.F. et al.) https://doi.org/10.1007/978-3-030-89385-9_8 (Springer, 2021).
74. Woolley, A. W., Aggarwal, I. & Malone, T. W. Collective intelligence and group performance. *Curr. Dir. Psychol. Sci.* **24**, 420–424 (2015).
75. Malone, T. W. & Bernstein, M. *Handbook of Collective Intelligence* (MIT Press, 2022).
76. Galesic, M. et al. Beyond collective intelligence: collective adaptation. *J. R. Soc. Interf.* **20**, 20220736 (2023).
77. Gonzalez, C. Building human-like artificial agents: a general cognitive algorithm for emulating human decision-making in dynamic environments. *Persp. Psychol. Sci*. **19**, 860–873 (2024).
78. Woolley, A. W. & Gupta, P. Understanding collective intelligence: investigating the role of collective memory, attention, and reasoning processes. *Perspect. Psychol. Sci*. **19**, 344–354 (2024).

79. Gupta, P. & Woolley, A. W. Articulating the role of artificial intelligence in collective intelligence: a transactive systems framework. *Proc. Hum. Factors Ergon. Soc. Ann. Meet.* **65**, 670–674 (2021).
80. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement learning: a survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996).
81. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
82. Van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double Q-learning. *Proc. Conf. AAAI Artif. Intell.* **30**, 2094–2100 (2016).
83. Kober, J., Bagnell, J. A. & Peters, J. Reinforcement learning in robotics: a survey. *Int. J. Rob. Res.* **32**, 1238–1274 (2013).
84. Christiano, P. F. et al. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **30**, 4302–4310 (2017).
85. Sutton, R. S. & Barto, A. G. *Reinforcement Learning* 2nd edn (MIT Press, 2018).
86. Patil, V. *Credit Assignment and Abstraction for Sequential Decision Making*. PhD thesis, Univ. Linz (2024).
87. Icarte, R. T., Klassen, T. Q., Valenzano, R. & McIlraith, S. A. Using reward machines for high-level task specification and decomposition in reinforcement learning. *ICML* **80**, 2112–2121 (2018).
88. Argall, B. D., Chernova, S., Veloso, M. & Browning, B. A survey of robot learning from demonstration. *Rob. Auton. Syst.* **57**, 469–483 (2009).
89. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).
90. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. 33rd International Conf. on Machine Learning* (eds Florina Balcan, M. & Weinberger, K. Q.) 1050–1059 (PMLR, 2016).
91. Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* **14**, 357–364 (2010).
92. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.11929 (2020).
93. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. YOLOv4: optimal speed and accuracy of object detection. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2004.10934 (2020).
94. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
95. Brehmer, B. Dynamic decision making: human control of complex systems. *Acta Psychol.* **81**, 211–241 (1992).
96. Gonzalez, C., Lerch, J. F. & Lebiere, C. Instance-based learning in dynamic decision making. *Cogn. Sci.* **27**, 591–635 (2003).
97. Dutt, V. & Gonzalez, C. Accounting for outcome and process measures in dynamic decision-making tasks through model calibration. *J. Dynam. Decision Making* https://doi.org/10.11588/jddm.2015.1.17663 (2015).
98. Gunning, D. & Aha, D. W. DARPA's explainable artificial intelligence program. *AI Mag.* **40**, 44–58 (2019).
99. Campbell, M., Hoane, A. J. Jr & Hsu, F.-H. Deep blue. *Artif. Intell.* **134**, 57–83 (2002).
100. Schaeffer, J. et al. Checkers is solved. *Science* **317**, 1518–1522 (2007).
101. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
102. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
103. Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The Arcade Learning Environment: an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253–279 (2013).
104. Binz, M. et al. A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).
105. Konstantinidis, E., Harman, J. L. & Gonzalez, C. Patterns of choice adaptation in dynamic risky environments. *Mem. Cogn.* **50**, 864–881 (2022).
106. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1412.6572 (2014).
107. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
108. Bradshaw, J. M., Hoffman, R. R., Johnson, M. & Woods, D. D. The seven deadly myths of 'autonomous systems'. *IEEE Intell. Syst.* **28**, 54–61 (2013).
109. Abdar, M. et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
110. Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access.* **6**, 52138–52160 (2018).
111. Bansal, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proc. CHI Conf. Human Factors in Computing Systems* 81 (ACM, 2021).
112. Gelman, A. & Hill, J. *Analytical Methods for Social Research: Data Analysis Using Regression and Multilevel/hierarchical Models* (Cambridge Univ. Press, 2006).
113. Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms* (Whittles, 2012).
114. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
115. Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2003).
116. Steyvers, M. et al. What large language models know and what people think they know. *Nat. Machine Intell.* **7**, 221–231 (2025).
117. Barredo Arrieta, A. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion.* **58**, 82–115 (2020).
118. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Machine Intell.* **1**, 206–215 (2019).
119. Dwivedi, R. et al. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.* **55**, 194 (2023).
120. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
121. Lundberg, S. M., & Lee, S.-I. A unified approach to interpreting model predictions. *Neural Information Processing Systems* **30**, 4765–4774 (2017).
122. Bhatt, U. et al. Explainable machine learning in deployment. In *Proc. Conf. Fairness, Accountability, and Transparency* 648–657 (ACM, 2020).
123. Chatzimparmpas, A., Martins, R. M., Jusufi, I. & Kerren, A. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf. Vis.* **19**, 207–233 (2020).
124. Cortez, P. & Embrechts, M. J. Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf. Sci.* **225**, 1–17 (2013).
125. Lee, S., Hwang, S., Kim, D. & Lee, K. Conversational agents as catalysts for critical thinking: challenging social influence in group decision-making. In *Proc. Extended Abstracts of the CHI Conf. on Human Factors in Computing Systems* (eds., Yamashita, N. et al.) 1–12 (Association for Computing Machinery, 2025).
126. Zhang, Y. et al. Siren's song in the AI ocean: a survey on hallucination in large language models. *Comput. Linguist.* https://doi.org/10.1162/COLI.a.16 (2025).
127. Ye, H., Liu, T., Zhang, A., Hua, W. & Jia, W. Cognitive mirage: a review of hallucinations in large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2309.06794 (2023).
128. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.08399 (2023).
129. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Neural Information Processing Systems*, 9459–9474 (NeurIPS, 2020).
130. Lykouris, T. & Weng, W. Learning to defer in congested systems: the AI-human interplay. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.12237 (2025).
131. Raghu, M. et al. The algorithmic automation problem: prediction, triage, and human effort. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1903.12220 (2019).
132. Bansal, G. et al. Updates in human–AI teams: understanding and addressing the performance/compatibility tradeoff. *Proc. Conf. AAAI Artif. Intell.* **33**, 2429–2437 (2019).
133. Gupta, P., Nguyen, T. N., Gonzalez, C. & Woolley, A. W. Fostering collective intelligence in human–AI collaboration: laying the groundwork for COHUMAIN. *Top. Cogn. Sci.* https://doi.org/10.1111/tops.12679 (2023).
134. Gonzalez, C., Admoni, H., Brown, S. & Woolley, A. W. COHUMAIN: building the socio-cognitive architecture of collective HUman-MAchine INtelligence. *Top. Cogn. Sci.* **17**, 180–188 (2023).
135. Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Top. Cogn. Sci.* **7**, 217–229 (2015).
136. Dancy, C. L. & Workman, D. On integrating generative models into cognitive architectures for improved computational sociocultural representations. *Proc. AAAI Symp. Ser.* https://doi.org/10.1609/aaaiss.v2i1.27685 (2024).
137. Malloy, T. & Gonzalez, C. Applying generative artificial intelligence to cognitive models of decision making. *Front. Psychol.* **15**, 1387948 (2024).
138. Liu, Y., Liu, Y. & Shen, C. Combining minds and machines: investigating the fusion of cognitive architectures and generative models for general embodied intelligence. *Proc. AAAI Symp. Ser.* https://doi.org/10.1609/aaaiss.v2i1.27693 (2024).
139. Du, Y., Prébot, B., Xi, X. & Gonzalez, C. Towards autonomous cyber defense: predictions from a cognitive model. *Proc. Hum. Factors Ergon. Soc. Ann. Meet.* **66**, 1121–1125 (2022).
140. Du, Y., Prebot, B., Malloy, T., Fang, F. & Gonzalez, C. Experimental evaluation of cognitive agents for collaboration in human-autonomy cyber defense teams. *Comput. Hum. Behav. Artif. Hum.* **4**, 100148 (2025).
141. van Rooij, I. et al. Reclaiming AI as a theoretical tool for cognitive science. *Comput. Brain Behav.* https://doi.org/10.1007/s42113-024-00217-5 (2024).
142. Anderson, J. R. & Lebiere, C. J. *The Atomic Components of Thought* (Psychology Press, 2014).
143. Laird, J. E. *The Soar Cognitive Architecture* (MIT Press, 2019).
144. Laird, J., Lebiere, C. & Rosenbloom, P. A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Mag.* **38**, 13–26 (2017).
145. Gigerenzer, G., Todd, P. M. & ABC Research Group T. *Simple Heuristics That Make Us Smart* (Oxford Univ. Press, 2000).
146. Denrell, J. & Le Mens, G. in *Sampling in Judgment and Decision Making* (eds Fiedler, K. et al.) 90–112 (Cambridge Univ. Press, 2023).
147. Denrell, J. & March, J. G. Adaptation as information restriction: the hot stove effect. *Organ. Sci.* **12**, 523–538 (2001).
148. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
149. Gigerenzer, G. & Goldstein, D. Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* **103**, 650–669 (1996).
150. Einhorn, H. J. & Hogarth, R. M. Behavioral decision theory: processes of judgement and choice. *Annu. Rev. Psychol.* **32**, 53–88 (1981).
151. Rieskamp, J. & Otto, P. E. SSL: a theory of how people learn to select strategies. *J. Exp. Psychol. Gen.* **135**, 207–236 (2006).
152. Scheibehenne, B., Rieskamp, J. & Wagenmakers, E.-J. Testing adaptive toolbox models: a Bayesian hierarchical approach. *Psychol. Rev.* **120**, 39–64 (2013).
153. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).

154. Gonzalez, C. Learning and dynamic decision making. *Top. Cogn. Sci.* **14**, 14–30 (2022).
155. Gonzalez, C., Aggarwal, P., Lebiere, C. & Cranford, E. Design of dynamic and personalized deception: a research framework and new insights. In *Proc. Ann. Hawaii Int. Conf. System Sciences*, 1825–1834 (ScholarSpace, 2020).
156. Hamade, K., McIlroy-Young, R., Sen, S., Kleinberg, J. & Anderson, A. Designing skill-compatible AI: methodologies and frameworks in chess. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.05066 (2024).
157. Case, N. How to become a centaur. *J. Des. Sci*. https://doi.org/10.21428/61b2215c (2018).
158. Cannon-Bowers, J. A., Salas, E. & Converse, S. A. Cognitive psychology and team training: shared mental models in complex systems. *Hum. Factors Bull.* **33**, 1–4 (1990).
159. Craik, K. *The Nature of Explanation* (Cambridge Univ. Press, 1943).
160. Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E. & Cannon-Bowers, J. A. The influence of shared mental models on team process and performance. *J. Appl. Psychol.* **85**, 273–283 (2000).
161. Baker, C. L. *Bayesian Theory of Mind: Modeling Human Reasoning About Beliefs, Desires, Goals, and Social Relations*. PhD thesis, MIT (2012).
162. Baker, E. R., D'Esterre, A. P. & Weaver, J. P. Executive function and theory of mind in explaining young children's moral reasoning: a test of the hierarchical competing systems model. *Cogn. Dev.* **58**, 101035 (2021).
163. Nguyen, T. N. & Gonzalez, C. Theory of mind from observation in cognitive models and humans. *Top. Cogn. Sci.* **14**, 665–686 (2022).
164. Baker, S. T., Leslie, A. M., Gallistel, C. R. & Hood, B. M. Bayesian change-point analysis reveals developmental change in a classic theory of mind task. *Cogn. Psychol.* **91**, 124–149 (2016).
165. Courville, A. C., Daw, N. D. & Touretzky, D. S. Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10**, 294–300 (2006).
166. Bickmore, T. & Cassell, J. in *Advances in Natural Multimodal Dialogue Systems. Text, Speech and Language Technology* (eds van Kuppevelt, J. C. J. et al.) https://doi.org/10.1007/1-4020-3933-6_2 (Springer, 2005).
167. Picard, R. W. *Affective Computing* 252 (MIT Press, 1997).
168. Ford, A. *Modeling The Environment* 2nd edn 488 (Island Press, 2010).
169. Sterman, J. System dynamics: systems thinking and modeling for a complex world. *MIT Libraries* https://dspace.mit.edu/handle/1721.1/102741 (2002).
170. Aggarwal, P. et al. Designing effective masking strategies for cyberdefense through human experimentation and cognitive models. *Comput. Secur.* **117**, 102671 (2022).
171. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning 70* (eds Precup, D. & Teh, Y. W.) 1126–1135 (2017).
172. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
173. Bugbee, E. H. & Gonzalez, C. Making predictions without data: how an instance-based learning model predicts sequential decisions in the balloon analog risk task. *Proc. Ann. Meet. Cognitive Science Society* (eds Culbertson, J. et al.) 3167–3174 (Cognitive Science Society, 2022).
174. Gonzalez, C., Vanyukov, P. & Martin, M. K. The use of microworlds to study dynamic decision making. *Comput. Hum. Behav.* **21**, 273–286 (2005).
175. Glikson, E. & Woolley, A. W. Human trust in artificial intelligence: review of empirical research. *Acad. Manag. Ann.* **14**, 627–660 (2020).
176. Rossi, F. Building trust in artificial intelligence. *J. Int. Aff.* **72**, 127–134 (2018).
177. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Machine Intell.* **1**, 389–399 (2019).
178. London, A. J. & Heidari, H. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through AI systems. *Minds Machines* **34**, 41 (2024).
179. Raji, I. D., Kumar, I. E., Horowitz, A. & Selbst, A. The fallacy of AI functionality. In *Conf. Fairness, Accountability, and Transparency* Vol. 12, 959–972 (ACM, 2022).
180. Barocas, S. & Hardt, M. *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023).
181. Berk, R., Heidari, H. & Jabbari, S. Fairness in criminal justice risk assessments: the state of the art. *Sociol. Meth. Res.* https://doi.org/10.1177/0049124118782533 (2021).
182. Lipton, Z. C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *ACM Queue* **16**, 31–57 (2018).
183. Lee, H.-P. et al. Deepfakes, phrenology, surveillance, and more! A taxonomy of AI privacy risks. In *Proc. CHI Conf. Human Factors in Computing Systems* (eds Mueller, F. F. et al.) Vol. 79, 775 (ACM, 2024).
184. Jakesch, M., Buçinca, Z., Amershi, S. & Olteanu, A. How different groups prioritize ethical values for responsible AI. In *Conf. Fairness, Accountability, and Transparency (FAccT '22)* 310–323 (ACM, 2022).
185. Rakova, B., Yang, J., Cramer, H. & Chowdhury, R. Where responsible AI meets reality: practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Human-Computer Interact.* **5**, 7 (2021).
186. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the carbon emissions of machine learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.09700 (2019).
187. Luccioni, A., Viguier, S. & Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *J. Machine Learn. Res.* **24**, 253:1-253:15 (2022).
188. Frank, M. R. et al. Toward understanding the impact of artificial intelligence on labor. *Proc. Natl Acad. Sci. USA.* **116**, 6531–6539 (2019).
189. De Stefano, V. 'Negotiating the algorithm': automation, artificial intelligence, and labor protection. *International Labour Organization* https://www.ilo.org/publications/negotiating-algorithm-automation-artificial-intelligence-and-labour (2019).
190. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In *Proc. ACM Human–Computer Interact. (CSCW1)* 5, 21 (ACM, 2021).
191. Vasconcelos, H. et al. Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Human–Computer Interact.* **7**, 1–38 (2023).
192. Spector, J. M. & Ma, S. Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. *Smart Learn. Environ.* **6**, 8 (2019).
193. Bai, L., Liu, X. & Su, J. ChatGPT: the cognitive effects on learning and memory. *Brain-X* **1**, e30 (2023).
194. Levy, K. E. C. The contexts of control: information, power, and truck-driving work. *Inf. Soc.* **31**, 160–174 (2015).
195. Gyevnar, B. & Kasirzadeh, A. AI safety for everyone. *Nat. Machine Intell.* **7**, 531–542 (2025).
196. Balietti, S. NodeGame: real-time, synchronous, online experiments in the browser. *Behav. Res. Meth.* **49**, 1696–1715 (2017).
197. Chevalier-Boisvert, M. et al. Minigrid & miniworld: modular & customizable reinforcement learning environments for goal-oriented tasks. In *37th Conf. Neural Information Processing Systems (NeurIPS)* (eds. Oh, A. et al.) 73383–73394 (Institute of Electrical and Electronics Engineers, 2023).
198. Zhang, L., Ji, Z. & Chen, B. CREW: facilitating human-AI teaming research. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2408.00170 (2024).
199. McDonald, C. & Gonzalez, C. Controllable complementarity: subjective preferences in human–AI collaboration. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2503.05455 (2025).
200. Carvalho, W., Goddla, V., Sinha, I., Shin, H. & Jha, K. NiceWebRL: a Python library for human subject experiments with reinforcement learning environments. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2508.15693 (2025).
201. Mayer, L. W. et al. Human–AI collaboration: trade-offs between performance and preferences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2503.00248 (2025).
202. Schelble, B. G., Flathmann, C., McNeese, N. J., Freeman, G. & Mallick, R. Let's think together! Assessing shared mental models, performance, and trust in human–agent teams. *Proc. ACM Human–Computer Interact.* **6**, 13 (2022).
203. Lipton, Z. C. The mythos of model interpretability. *ACM Queue* **16**, 31–57 (2018).

Acknowledgements
The authors acknowledge the support of the AI Research Institutes Program funded by the National Science Foundation under the AI Institute for Societal Decision Making (AI-SDM), award number 2229881. The authors acknowledge occasional use of generative AI to polish and shape their writing in earlier drafts of this paper.

Author contributions
Both authors researched data for the article. Both authors contributed substantially to discussion of the content. C.G. wrote the content on cognitive AI; H.H. wrote the content on ethical considerations and risks. Both authors contributed equally to other sections. C.G. edited the manuscript after the reviews and H.H. reviewed the edits before submission.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Reviews Psychology* thanks Stefan Herzog and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.