nature human behaviour

Perspective

https://doi.org/10.1038/s41562-024-01991-9

Building machines that learn and think with people

Received: 11 April 2024

Accepted: 23 August 2024

Published online: 22 October 2024

Check for updates

Katherine M. Collins[1,9] ✉, Ilia Sucholutsky[2,9], Umang Bhatt[3,4,9], Kartik Chandra[5,9], Lionel Wong[5,9], Mina Lee[6,7], Cedegao E. Zhang[5], Tan Zhi-Xuan[5], Mark Ho[3], Vikash Mansinghka[5,10], Adrian Weller[1,4,10], Joshua B. Tenenbaum[5,10] & Thomas L. Griffiths[2,8,10]

What do we want from machine intelligence? We envision machines that are not just tools for thought but partners in thought: reasonable, insightful, knowledgeable, reliable and trustworthy systems that think with us. Current artificial intelligence systems satisfy some of these criteria, some of the time. In this Perspective, we show how the science of collaborative cognition can be put to work to engineer systems that really can be called 'thought partners', systems built to meet our expectations and complement our limitations. We lay out several modes of collaborative thought in which humans and artificial intelligence thought partners can engage, and we propose desiderata for human-compatible thought partnerships. Drawing on motifs from computational cognitive science, we motivate an alternative scaling path for the design of thought partners and ecosystems around their use through a Bayesian lens, whereby the partners we construct actively build and reason over models of the human and world.

Computers have long been seen as tools for thought. Steve Jobs called computers "bicycles for the mind": tools that dramatically increase the efficiency, productivity and joy of thinking. Now, 30 years later, this metaphor is beginning to change. Computer systems are increasingly referred to not as vehicles but as "copilots"[1,2]: we have moved from designing tools for thought to actual partners in thought.

The current wave of artificial intelligence (AI) technologies, particularly language models, have catalysed this transition (key terms are defined in Glossary). Users no longer have to know how to write code to engage intimately with computers; we can now interface through the medium of natural language. Humans already think alone and together, and these thoughts are often communicated through the medium of language[3]. We long have done so—from developing new modes of thinking through questioning and debate to teaching and learning through language. The apparent power of these new systems (which getting closer to the kind of AI imagined in the field's early days[4–9])—as well as challenges faced by the current iterations of such systems—invites us to think about what it will take to build systems that truly act as effective thought partners. We argue that good thought partners are systems (1) that can understand us, (2) that we can understand and (3) that have sufficient understanding of the world that we can engage on common ground.

One path to building such thought partners is to scale foundation models (such as large language models (LLMs)[10]) with large amounts of human demonstrations and feedback, along with 'traces' of human thought scraped from web-scale data[11–13]. Although such an approach has produced systems that accurately mimic human behaviour (for example, producing fluent text), these machines do not robustly simulate human cognition (for example, explicitly reasoning about the world or other minds) in ways expected by a true thought partner[3,14–20].

What would it take to design systems that meet our criteria? One promising path is to design systems that build explicit models of the

[1]Department of Engineering, University of Cambridge, Cambridge, UK. [2]Department of Computer Science, Princeton University, Princeton, NJ, USA. [3]Center for Data Science, NYU, New York, NY, USA. [4]Alan Turing Institute, London, UK. [5]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. [6]Microsoft Research, New York, NY, USA. [7]Department of Computer Science, University of Chicago, Chicago, IL, USA. [8]Department of Psychology, Princeton University, Princeton, NJ, USA. [9]These authors contributed equally: Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong. [10]These authors jointly supervised this work: Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, Thomas L. Griffiths. ✉e-mail: kmc61@cam.ac.uk

Glossary

Collaborative cognition
The process by which two or more agents work together in some aspect(s) of thinking (for example, planning together, learning together or creating together).

Thought partner
Another entity (human or AI) that works with an agent to push forward some aspect(s) of thinking.

Artificial intelligence (AI)
Computational systems that are able to process inputs and engage in some aspect of learning, planning, reasoning and/or decision-making. Used interchangeably with machines.

Large language model (LLM)
A particular kind of AI system that learns a distribution over text, often trained on large amounts of web-scale text data. LLMs are a class of large-scale foundation models.

Agent
An entity that can process inputs, make decisions and take actions in some environment.

Dyad
A system with two agents (for example, human–human, human–AI or AI–AI).

Resource rationality
The idea that human behaviour and cognition can be viewed as rational under bounded constraints (for example, under limited working memory).

Probabilistic generative model
A model of how the data one observes about the world are generated by some probabilistic process, from which one can sample new observations and make queries about existing observations.

Probabilistic programming language (PPL)
A language for expressing probabilistic generative models as computer programs that interleave deterministic code (for example, arithmetic, logic or artificial neural networks) with random choices. PPLs allow users to specify probabilistic models and inference algorithms in a modular and compositional manner.

Bayesian inference
A method for updating one's beliefs over various aspects of the world, grounded in probability theory; in Bayesian inference, an agent updates their beliefs by assigning higher credence to hypotheses that better explain the evidence, weighted against the backdrop of their prior beliefs.

Affordance
Design features of a system that inform use.

Table 1 | Modes of collaborative thought

Mode	Ongoing challenges	Sampling of existing systems
Collaborative planning		
• Joint decision-making • Decentralized cooperation • Goal and task assistance	Reliable goal inference Value and intent alignment Scalable multi-agent planning	Collaborative robots[68,222] Video game sidekicks[223,224] Language-based assistants[35,225]
Collaborative learning		
• Pair and team problem-solving • Identification of knowledge gaps • New problem construction	Strong and robust problem-solving abilities Personalized curriculum pacing Problem construction of targeted difficulty	Programming learning aids[178,226–228] Mathematics tutors[15,229,230]
Collaborative deliberation		
• Debate and argumentation • Critical review and discussion • Consensus formation	Opinion diversity Verifiable reasoning Formation of common ground	Machine-assisted debating[231–233] Consensus writing and opinion mapping[234,235]
Collaborative sense-making		
• Explanation • Visualization • Data analytics	Exponential increases in data produced Accessible communication Fidelity of insights to the world	Probabilistic data modelling[158,159,161,236,237] Machine-assisted theory discovery[238–240]
Collaborative creation and ideation		
• Co-design • Idea critiquing • Brainstorming	Generation diversity Style consistency Modular customizability	Machine-assisted writing[72,74,241] Prompted image creation[242–244] Collaborative sketching[245–247]

Settings in which human–human and human–AI thought partners can engage.

task, world and human (where these models are structured[21] rather than distributionally learned from data)—drawing on formal frameworks grounded in cognitive psychology for understanding how humans think, alone and together. In this Perspective, we chart a new vision for the design of AI thought partners. Decades of work in the behavioural sciences provide valuable insights for designing human-centric, uncertainty-aware thought partners. Drawing on such research, we argue that effective thought partners are those that build models of the human and the world.

This toolkit includes foundation models[22–24] but is not limited to them. Indeed, foundation models such as LLMs are fuelling new motifs for thinking about human minds in computational terms (for example, "rational meaning construction"[16]) interleaved alongside techniques from probabilistic programming[25–29], goal-directed search[30–32] and other explicit, structured representations—for example, of agents thinking about other agents[33–35]. We already have tools that help us to build machines that learn and think like people[36]. We propose applying that toolkit to collaborative cognition—to build machines that learn and think with people.

What are thought partners?

When we think, we draw coherent inferences, make predictions and act on these predictions—from assessing what birthday present to gift a treasured friend, to formulating a new scientific hypothesis and experiment plan to evaluate a theory. We flexibly draw on prior knowledge and update our beliefs through experience (as we discuss below). We not only solve problems but imagine new ones[37]. And we think together. For generations, humans have discussed and debated ideas and developed ecosystems to disseminate such thoughts to new audiences. Much scientific innovation has come through collaboration, where advances are frequently fuelled by engaging with diverse partners who offer new ideas yet share our values[38].

Modes of collaborative thought

As an illustration of the many ways that people and machines might think with each other, we highlight a few modes of collaborative thought (Table 1). This set of modes, partly inspired by characterizations of thinking and reasoning in psychology[39,40], are not meant to be comprehensive of all aspects of thought. Rather, we see these modes as ripe for the further development of AI thought partners.

Example domains

We next outline a few diverse domains in which the development of AI thought partners able to truly collaborate with humans (Fig. 1) may be particularly valuable. We highlight common computational challenges that arise when considering what effective partnership might look like in each domain, foreshadowing our proposed desiderata. We later return to these case studies with concrete human-centric thought partner instantiations.

Thought partners for programming. Programming is a cognitively demanding activity that requires gaining fluency in translating human intentions into formal, machine-interpretable languages. It is no surprise that decades of effort have gone into designing tools to help people to program[41–45]. New 'programming assistant' tools such as GitHub Copilot have rapidly gained enormous popularity and attention, but these tools are often unreliable[46–48]—for example, failing to understand users' intentions[49] and generating bugs that may be particularly risky alongside beginner programmers[50]. Programming

https://doi.org/10.1038/s41562-024-01991-9

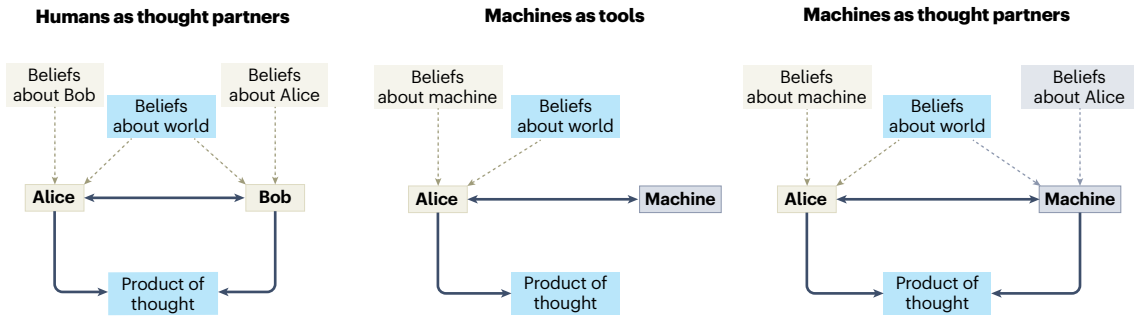

Fig. 1 | Examples of ecosystems for thinking. Humans have long thought together. Machines have expanded the efficiency of human thinking. Now, machines—powered by AI—open up new realms of computational thought partnership with humans.

involves much more than just accurate in-line code suggestions—which, at the time of writing, GitHub Copilot specializes in. Humans plan abstract, structural decisions and collaboratively learn, and we need partners who can answer our questions—such as why code behaves as it does or fails to work. A good collaborative programming partner seeks to understand not only the programming language but also their fellow programmer, inferring and reasoning about our overarching intentions and adapting to both what we know and what we do not know.

Thought partners for embodied assistance. Ensuring that embodied agents can form accurate and physically realizable plans is foundational for effective assistance we can trust—from guessing what a friend wants when we help them to cook[51], to working with someone with different physical abilities[52], to carrying out a high-stakes search-and-rescue mission[53]. Although much current research on embodied AI and assistive robots focuses on learning specific skills or following simple instructions[54–56], evaluations suggest that even state-of-the-art language models fine-tuned on extensive human feedback continue to struggle with tasks that require reliable, effective planning towards novel goals[57,58]. Instead, ideal assistive partners understand our actions, words and instructions as expressions of goals, beliefs and intentions[59–61] that are grounded in physical possibilities[62], while also understanding that these can be shared across multiple minds[63–65]. In addition, effective partners account for each other's limitations in perception, planning and world modelling, correcting for possible mistakes[66,67] and acting to make their intentions more legible[68,69].

Thought partners for storytelling. Another domain in which we may want thought partners is storytelling—for writers, filmmakers and even scientists. Storytelling is a complex, iterative cognitive process[70,71] with substantial opportunities for thought partners to collaboratively ideate and create with humans by helping to brainstorm new ideas, generate storylines and improve their writing style and tone[72–77]. For this process to be productive, a thought partner needs to understand more than just our authorial intentions and dispositions—they also need to understand the audience we are speaking to (that is, to understand the social world), including audience expectations and likely interpretations of the stories we are crafting for them.

Thought partners for medicine. Doctors need to sense-make, plan, deliberate and continually learn in the face of new medical evidence. A primary care doctor is not unlike Sherlock Holmes—collating and integrating disparate bits of evidence with their prior beliefs to make decisions under uncertainty. Yet, doctors rarely have enough time to engage deeply with each patient[78], driving high rates of burnout with knock-on effects on patient care quality[79]. Can we develop safe, reliable thought partners that can free doctors up to spend more time and communicate better with their patients? Already, foundation models are becoming proficient in medical assessments[80,81], seemingly capable of easing the heavy burden on doctors by assisting and partnering[82,83], and even providing preferable responses to patients[84]. Yet, it is not clear that these systems understand us (and our cognitive limitations), understand the world (underlying biology) and enable us to understand them (which, in this context, may be important for transparency and reliability[85–88]).

Desiderata

What then do we want from thought partners? There are many criteria for tools for thought that are of course relevant, such as efficiency, accuracy, robustness, fairness, cost and scalability. But the domains above illuminate that what is distinctive about a thought partner is its relationship to the user[89]. Looking to ideas from the behavioural sciences motivates three desiderata to guide the design of human-centred thought partners:

(1) You understand me: we would like our thought partners to understand our goals, plans, (possibly false) beliefs and resource limitations, taking into account what they have observed of us in the past and present to best collaborate with us in the future[90,91]. For example, a thought partner should adaptively change strategies when working with an expert, layperson or child, meeting us where we are.
(2) I understand you: we would like our thought partners to act in a way that is legible to us[68,92] and communicate with us in the way we intuitively understand[93–95].
(3) We understand the world: we would like our thought partners to be tethered to reality[96]. This means being accurate and knowledgeable as well as working with a shared representation of the world, domain or task[97–99]. Furthermore, our use of 'we' emphasizes that thought partnerships are fundamentally about synergy, moving beyond the sum of their parts.

Engineering human-centred thought partners

Our core proposal is that our three desiderata can be engineered explicitly, building on theoretical motifs from computational cognitive science and cognitively informed AI (summarized in Table 2), rather than left as emergent and potentially brittle properties arising implicitly in systems trained for other ends[20]. Here we articulate a framework for engineering thought partners designed to robustly and explicitly function as cooperative, collaborative actors. Humans are far from homogeneous, perfectly rational oracles, nor are we so unpredictable that it is hopeless to model human behaviour. We argue that models that explain human cognition and choice as approximately optimal solutions given goals and constraints provide an ideal starting point for designing thought partners, and that a Bayesian formalism provides a probabilistically sound common conceptual language that facilitates cross-talk between different disciplines[22,100,101].

Table 2 | Bayesian thought partner toolkit

Motif	Description	Sample references
Probabilistic mental models and inference	Humans update beliefs and draw inferences consistent with probabilistic generative models representing the world.	21,103,248
Structured knowledge representations	Humans have abstract, highly structured conceptual representations that include causality, agents and physical representations.	249–251
Hierarchical models	Humans construct and update hierarchical representations that separate concrete knowledge and belief from abstract ones.	106,184,252
Theory learning as programme synthesis	Human minds can be viewed as growing and editing theories of the world, expressed as programs, to improve their codebase (world models).	123,155,253
Resource-rationality	Humans make rational choices about how to allocate finite computational resources, including time and memory.	152,153,254
Goal-directed planning and search	Humans are intentional actors, who plan to achieve goals by reasoning about the (uncertain) effects of their (possible) actions in the environment.	255–257
Bayesian theory of mind	Humans represent other agents as intentional, intelligent actors and probabilistically infer their mental states from observations of actions.	126,134,258
Rational speech acts	Humans reason about language as an intentional, communicative action to infer speakers' underlying goals.	59,143,259
Learning to learn	Humans meta-learn (improve our overarching ability to learn) jointly with learning new concrete concepts and skills.	36,260–262

A range of computational cognitive motifs for reverse-engineering the mind in engineering terms, drawn from computational cognitive science, can be used to build human-centric thought partners that meet our desiderata.

Implementing our desiderata

What does it take to engineer real systems that meet our desiderata? First, we propose that a thought partner that understands us should explicitly model its human collaborator as such—as a cooperative agent with structured internal beliefs, knowledge and goals, as well as fundamental resource limitations. Second, engineering a thought partner that we can understand benefits from looking at how humans model other humans: just as a good human collaborator seeks to learn and adapt to the relative strengths, imperfections and computational bounds of their partner, we can build machine thought partners that also reason about the computational demands they are placing on another agent such that we can appropriately predict their behaviour[18,102]. Finally, to build thought partners that understand the world—and learn and think synergistically alongside us—we argue that it is valuable to build on structured computational toolkits for grounding shared goals and communication into the environment and domain in which collaboration takes place.

Computational cognitive science motifs

We now non-exhaustively spotlight several key insights about modelling humans, modelling humans modelling humans and modelling humans modelling the world from computational cognitive science—motifs for reverse-engineering the mind (Table 2)—that we believe can inform engineering of human-centred thought partners. Although we acknowledge that some communities within cognitive science may disagree with some of these theories, we emphasize that the computational underpinnings of the motifs hold tremendous engineering potential for building thought partners in practice.

Probabilistic models of cognition. Decades of work in computational cognitive science have demonstrated the power of modelling aspects of human cognition as Bayesian inference through structured probabilistic generative world models[21,103–106]. Such approaches have found empirical success in capturing a diversity of facets of human cognition including early word learning[107], visual perception[108,109], physical reasoning[99,110,111], concept learning[112–114], language processing and acquisition[104,115–117], causal inference in children[118,119] and adults[120,121], memory reconstruction[122], and theory formation[123,124], among many others. Probabilistic models of cognition, particularly those built using a Bayesian approach, have offered principled formalisms in capturing rapid belief updating[125] and how we may integrate our common-sense world knowledge with new evidence to inform the actions and decisions we take in the world[126]. Probabilistic inference over structured representations, particularly drawing on Bayesian modelling and tools such as meta-level Markov decision processes[127], has provided a computational account of how humans plan so flexibly, with the capability of forming rich hierarchical goals and subgoals, across varied timescales[100,126,128–130].

Theory of mind and communication. In our quest to build systems for collaborative cognition, we are guided by the success of Bayesian accounts of how we reason about others' mental states and how we communicate about them. In particular, Bayesian treatments of theory of mind have offered strong accounts of how we may rapidly reason about each other's beliefs, desires, goals and intentions[33,131–134]. We may build mental models[135,136] of our thought partners, which can in turn be used to support communication and collaboration, informing the way we teach[137–139], infer whether to rely on a partner for help[140] and support rapid, flexible adaptation to new conversation partners[141,142]. We call particular attention to the rational speech act framework[59,143], which models communicative partners as recursively reasoning about each other's minds to inform what to say (from the perspective of the speaker) and how to interpret a received utterance (as the listener). Bayesian models provide a useful framework for formalizing such rich cross-partner inferences, allowing both social cognition and communication to be modelled with the same computational toolbox[144,145].

Resource-rationality and tractable theory-building. Human brains also have limited resources such as time, memory and attention that shape what we think about, how long we spend thinking and even how we communicate our thoughts to others[146]. We thus sometimes make systematically biased inferences[147,148]. Such 'erroneous' judgements can be captured by modelling humans as making rational use of our finite resources—for example, via approximate inference[125,149] or bounded planning[67]. Crucially, human cognition is tractable[150]. Indeed, we can navigate large, potentially unbounded, hypothesis spaces to build theories of the world—a process that seems to demand some kind of heuristics and approximations, which may be resource-rational[17,130,146,151–154]. One approach to modelling minds advocates thinking about humans as "world model builders" (or "hackers")—conducting experiments and updating our beliefs about compressed representations of the world, where these representations may be expressed as programs[123,155]. Such representations—bolstered by tools such as program synthesis—help to explore suboptimal behaviour[156].

Scaling thought partners via probabilistic programming

If Bayesian thought partners are to reason over models of their human thought partner and the world, these models need to continually evolve as new facts come to light and as the human thought partner themselves grows in their expertise, beliefs and needs. Probabilistic programming[26] provides one powerful methodology for building, scaling and performing inference in these kinds of rich models. For example, probabilistic programs can be learned from data[157,158] and synthesized via LLMs that encode rich priors[16,159,160]. Probabilistic programs also

enable fast approximate inference in world models that cohere with human common-sense knowledge and domain expertise[161,162], where the learned models are themselves amenable to modular inspection and editing by humans. Modern probabilistic programming languages[25,27,163] offer not just generic inference but programmable inference—that is, they automate the mathematics for hybrids of optimization[164,165], dynamic programming[166] and Monte Carlo inference[167]. Although such frameworks are certainly not the only methods to handle uncertainty and build effective and robust thought partners, we believe they are one promising and cognitively grounded approach to instantiating thought partners today, as we discuss in our case studies.

Infrastructure around thought partners

The design of systems that learn and think with people necessitates careful construction not only of the thought partner (that is, the machine itself) but also of the infrastructure within which human and computational thought partners collaborate[102]. Questions such as 'When and where should a human be able to engage a computational thought partner to ensure effective and appropriate use?' or 'For a given problem, is the human or computational thought partner better suited to start first, in light of their respective strengths and weakness, costs of the task at hand and particular mode of thought?' inform the design of the workflow that surrounds thought partnership. This sociotechnical ecosystem may be dictated by external regulations, organizational practices or other principles[73,168–171] and is crucially informed by studies of human behaviour. For example, Article 14 of the European Union AI Act requires users of high-risk AI systems "to correctly interpret the high-risk AI system's output" and "to remain aware of the possible tendency of automatically relying or over-relying on the output." Satisfying such requirements not only begets careful design of thought partners (for example, that we can understand) but also demands careful design of the system of affordances[172,173] and infrastructure around thought partnerships (for instance, communicating back to humans information about their reliance strategies). Disentangling thought partners from the infrastructure around them provides a modular scaffold for addressing unintentional thought partnership behaviour, such as over-reliance[174] and "illusions of understanding"[175]. Bayesian modelling has already found success in inferring humans' reliance strategies[176] and regions of the task space where a human versus machine can complement one another[177].

Case studies in engineering thought partners

We now return to the example domains previously introduced and discuss specific case studies (depicted in Fig. 2). Our goal is to demonstrate the potential benefits of endowing thought partners with structured probabilistic models of the human and/or world and to provide a flavour of the kinds of infrastructure questions that may surround them to ensure that the thought partners we build work with people.

Thought partners for programming

We highlighted some visions for effective programming partnerships, such as a partner that can address 'why' questions. One recent idea, from Chandra et al.[178], is to apply the Bayesian toolkit to explain surprising behaviour of computer programs in a human-like way. Chandra et al.[178] apply Bayesian models of mental state inference and rational communication[179] to design a system called WatChat that answers questions such as "Why did program *p* output result *r*?" in a principled, human-like way. WatChat infers what erroneous mental model might cause the programmer to have expected something different (partner understands user) and generates an explanation that 'debugs' that mental model (user understands partner). WatChat represents possible mental models themselves as programs whose bugs correspond to possible misconceptions; mental models can thus be inferred by Bayesian program synthesis (Table 2). Such a framework can also be inverted to help to design new questions for teachers or self-driven learners to identify misconceptions.

Thought partners for embodied assistance

Recall the challenge of collaboratively planning uncertain tasks, from a search-and-rescue mission to everyday cooking, wherein we typically want to infer shared goals and communicative intent from our partners. This cooperative logic can be modelled in a Bayesian architecture called cooperative language-guided inverse plan search (CLIPS)[35]. By modelling humans as cooperative planners who use language to communicate joint plans to achieve their goals[65], CLIPS is able to infer those plans and goals from both the actions and instructions of human collaborators. This allows CLIPS to pragmatically follow human instructions, using context to disambiguate the multiple meanings that a request might have, while proactively assisting with the goals that underlie the instruction. For example, CLIPS can understand the likely intentions behind an instruction such as 'Can you prepare the vegetables while I knead the dough?', inferring the shared goal of making pizza. These capabilities are made possible by using probabilistic programming infrastructure[25] to unite algorithms for Bayesian inverse planning[33,132] and human–AI alignment[51,61,180] with LLMs. In particular, by using LLMs to evaluate the probability of a natural language instruction given a possible intention, CLIPS can infer intentions from natural language in a coherent Bayesian manner—demonstrating the power of combining tools from the Bayesian thought partner toolkit.

Thought partners for storytelling

Storytelling is about crafting experience. Can we also apply the toolkit to help storytellers design experiences from first principles? Recent work has shown that a system grounded in Bayesian theory of mind can predict and even design interventions on the audience's experience of a story[181,182]. Chandra et al.[183] conceive of storytelling as "inverse inverse planning": that is, starting with human social cognition, modelled as Bayesian inverse planning[33], and then optimizing narrative events to shape the model's inferences over time. They show how a variety of storytelling techniques—from plot twists to stage mime—can be expressed in the language of inverse inverse planning to create animations that have a desired cognitive effect on viewers. Herein, we also highlight the breadth of thought partners for media beyond language, though the framework does nicely suggest a variety of natural extensions, such as integration into tools for creative writing[72–77].

Thought partners for medicine

Finally, we envision medical thought partners that both understand us—reasoning about the doctor, patient and care team as agents with goals, beliefs and worries—and complement our capabilities, integrating swaths of evidence that exceed our cognitive capacities to inform diagnosis and treatment. Although no system yet meets our desiderata for these criteria, we believe that a range of motifs and tools from the Bayesian thought partner toolkit here can support the development of such systems for collaborative sense-making and deliberation. We imagine Bayesian thought partners that can update their medical world knowledge in light of new insights in biology—for example, editing a code snippet of the underlying probabilistic world model[16] or growing the representation in a non-parametric hierarchical Bayesian model[184]. Such a model can then, similar to WatChat, synthesize new questions to ensure that the human doctor's own medical world model is sound. Early work demonstrates that we can use elements of our toolkit, specifically probabilistic programming, to learn rich generative models for oncology and support efficient user queries[185]. Yet, effective medical thought partners beckon a broader view of the ecosystem in which they are deployed[89,186]. If a doctor is over-relying on the output of the thought partner or is overburdened amid a surge in patient queries, infrastructure around the human and thought partner can modulate when a patient query is either routed to a human or the AI thought partner, or is deemed to need collaborative planning[187]. Systems for routing based on probabilistic modelling are already proving successful in simulation[188].

https://doi.org/10.1038/s41562-024-01991-9

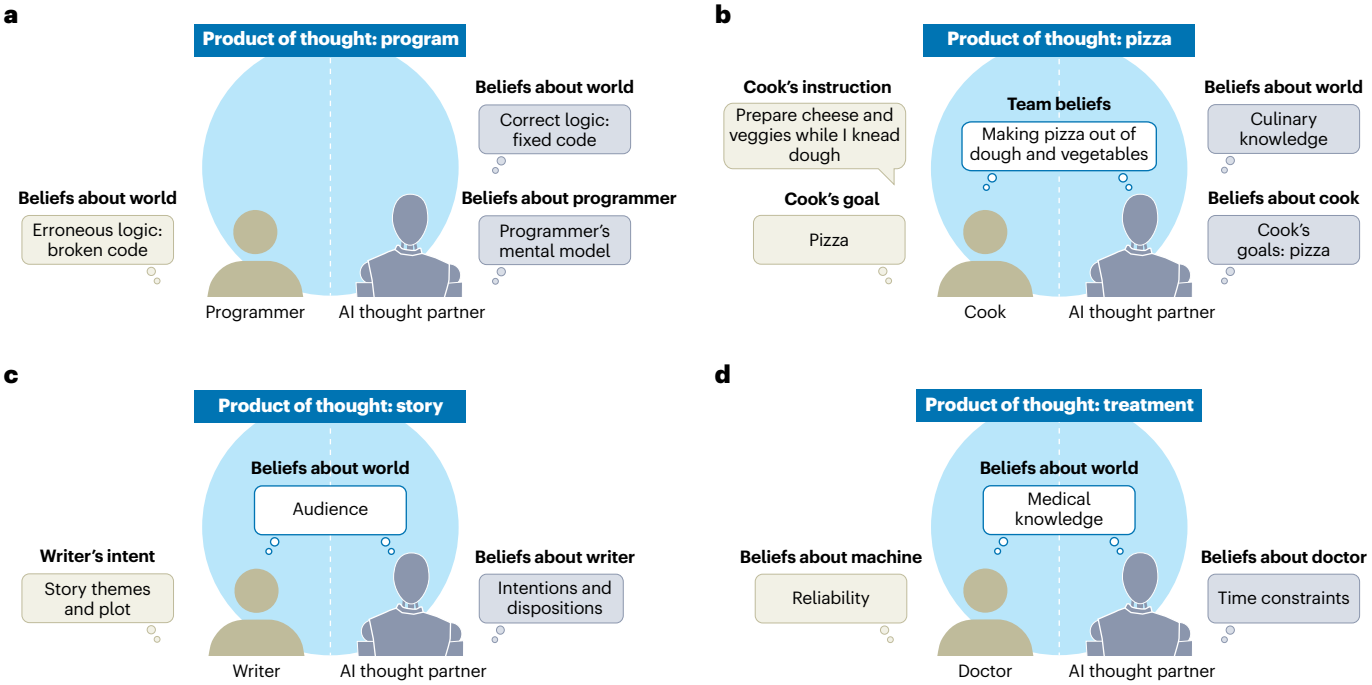

Fig. 2 | Case study depictions. a, WatChat infers the user's buggy mental model of the programming environment and interactively helps to 'patch' bugs in their understanding. **b**, CLIPS reasons explicitly about agents' goals, integrating (culinary) world knowledge and the human's utterances to infer appropriate actions. Both agents reason about the joint team plan (tomato and dough are needed to make pizza). **c**, Thought partners based on inverse inverse storytelling explicitly reason over models of the audience. **d**, Future thought partners for medicine can jointly reason with human doctors across modalities, a shared understanding of biology and patient needs, and a model of others' limitations.

Looking ahead

There is much exciting work to be done to characterize when and how to build thought partners across modes of collaborative thought, which can advance the dissemination and creation of new knowledge alongside humans. We next lay out several key challenges for researchers and designers intent on pursuing a human-centred programme of building machines that learn and think with people.

Non-dyad settings

Although there is substantial work to be done characterizing the space of possibilities for a single human and a single AI thought partner ('dyadic'), we envision a future where many humans and many machines ('non-dyadic'), across roles and specialties in increasingly complex social systems[189], engage in the realm of thought[190–192]. Already, researchers are exploring non-dyadic versions of many of the modes of thought and case studies laid out above, including collaborative learning with groups of humans accompanied by an AI thought partner[193] and medical robot collision avoidance systems that need to account for multiple humans[194]. As in the dyad setting, extensions to non-dyadic settings can be bolstered by a deepening understanding of human behaviour in groups—expanding the Bayesian thought partner toolkit—as is already underway in the study of convention formation[141,195]. Looking ahead, citizen science is a promising example of the opportunities of creating large networks of humans and thought partners: Zooniverse, a large-scale galaxy classification crowdsourcing project, serves as a case study for exploring smart task allocation, blending human and machine classifications, and infrastructure changes that affect human participation and performance with outcomes including both iterative scientific progress and serendipitous scientific discovery[196].

Evaluation

The assessment of thought partners demands a multi-faceted, cross-disciplinary suite of approaches. At minimum, the evaluation of AI thought partners must include some element of interactivity[197]. Recent works have highlighted deficits in static evaluation of foundation models[15,198], demonstrating the need for considering the interaction process in addition to the final output, the first-person perspective in addition to the third-party perspective, and notions of preference beyond quality. In addition to interactive user studies, we posit that studying different kinds of thought partners across modes of collaborative thought would benefit from a controlled, yet rich, playspace; games provide one such domain. Games offer a good formalism for the study of repeated interactions between multiple agents and grounds to explore rich patterns of thought in social collaborative settings[199–202].

Risks and important considerations

Computational thought partners are by no means a guaranteed or universal good and come with certain risks. We call out three such spheres of risk: (1) reliance, critical thinking and access; (2) anthropomorphization; and (3) misalignment.

First, AI thought partners could induce over-reliance and impair the development of critical thinking skills[175,203–205], potentially acting as "steroids" for the mind[206]. We are concerned about these risks; our emphasis on the infrastructure around thought partner use is explicitly intended to help practitioners to take steps to address these challenges, motivating further design of infrastructure modifications such as cognitive forcing functions[207,208]. Conversely, some people may under-rely on a thought partner, particularly if there is inadequate AI literacy training for how to best make use of new thought partners[209–211]. Research has already found that the kinds of queries people make of AI systems can be informed by the amount of prior experience they have interacting with chatbots[15], meaning that students, researchers and other practitioners in lower-income communities may be unable to maximize the value of thought partnering. It is important to ensure that the benefits of thought partners are not confined to an exclusive set of people.

Second, on the topic of anthropomorphization, we highlight an important distinction between human-centric and human-like thought partners[212]. Our desideratum 'I understand you' advocates for thought

partners whose behaviour we understand; although this could draw on how we understand other humans, we should however be careful about interpreting such machine thought partners as we do humans. As Weizenbaum[6] illuminated with the ELIZA system, there are risks to developing computer systems that present themselves as human-like in ways that they are not: for example, by leading users to attribute undue intention to systems' responses or (in the long run) leading society to devalue human intelligence[213]. Human-like thought partners should maintain categorical delineation between humans and machines to prevent over-reliance[203,214] and promote human dignity without encroaching on any partner's self-worth[7]. The term used to refer to a thought partner can affect the assumptions made about their capabilities (for example, 'teammate' implies the machine and human are on equal footing) or can detract from a partner's human-like nature (for example, 'tool' would be less anthropomorphic).

Lastly, we note that insufficiently accurate, robust or cognitively grounded models can yield misalignment with humans, leading intended AI thought partners to act towards the wrong goals[215], provide wrong or misleading information[216], or violate safety constraints[217]. A Bayesian approach to thought partnership can address some of these issues, enabling uncertainty-aware decision-making that avoids overconfidence[180,218,219]. Yet, while inferring human thoughts and behaviour can be used to design better collaborators, models of humans are inherently dual-use and can also be used to mislead, surveil or manipulate[220]. It is crucial to consider whether thought partners are aligned with society at large or merely superficially aligned with users while serving more powerful interests[221].

Conclusion

If we are to build helpful and reliable human–AI thought partnerships, we advocate for design that explicitly recognizes and engages with the richness and diversity of human thought in an often unpredictable world. We have argued, supported by several case studies, that those engineering thought partners and the infrastructure around their use can benefit from drawing on motifs from computational cognitive science and cognitive AI. The future of collaborative cognition is bright, but not without risk; continual collaboration and knowledge sharing among behavioural scientists, AI practitioners, domain experts and related disciplines is crucial as we strive to build machines that truly learn and think with people.

References

1. *GitHub Copilot: Your AI Pair Programmer* https://github.com/features/copilot (GitHub, 2022).
2. *Copilot for Microsoft 365—Microsoft Adoption* https://adoption.microsoft.com/en-us/copilot/ (Microsoft, 2023).
3. Fedorenko, E., Piantadosi, S. T. & Gibson, E. A. Language is primarily a tool for communication rather than thought. *Nature* **630**, 575–586 (2024).
4. Turing, A. Computing machinery and intelligence. *Mind* **59**, 433–460 (1950).
5. Clynes, M. E. & Kline, N. S. Cyborgs and space. *Astronautics* **14**, 26–27 (1960).
6. Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
7. Shneiderman, B. *Human-Centered AI* (Oxford Univ. Press, 2022).
8. Bundy, A. *The Computer Modelling of Mathematical Reasoning* (Academic Press, 1983).
9. Anderson, J. R., Boyle, C. F., Corbett, A. T. & Lewis, M. W. Cognitive modeling and intelligent tutoring. *Artificial Intelligence* **42**, 7–49 (1990).
10. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.07258 (2021).
11. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
12. Christiano, P. F. et al. Deep reinforcement learning from human preferences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1706.03741 (2017).
13. Lee, K. et al. Aligning text-to-image models using human feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.12192 (2023).
14. 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).
15. Collins, K. M. et al. Evaluating language models for mathematics through interactions. *Proc. Natl Acad. Sci. USA* **121**, e2318124121 (2024).
16. Wong, L. et al. From word models to world models: translating from natural language to the probabilistic language of thought. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.12672 (2023).
17. Zhang, C., Collins, K., Weller, A., & Tenenbaum, J. AI for mathematics: a cognitive science perspective. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.13021 (2023).
18. Gweon, H., Fan, J. & Kim, B. Socially intelligent machines that learn from humans and help humans learn. *Philos. Trans. R. Soc. A* **381**, 20220048 (2023).
19. Mahowald, K. et al. Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517–540 (2024).
20. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. & Griffiths, T. L. Embers of autoregression: understanding large language models through the problem they are trained to solve. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2309.13638 (2023).
21. 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).
22. Griffiths, T. L., Zhu, J.-Q., Grant, E. & McCoy, R. T. Bayes in the age of intelligent machines. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2311.10206 (2023).
23. Sumers, T., Yao, S., Narasimhan, K., & Griffiths, T. Cognitive architectures for language agents. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2309.02427 (2023).
24. Binz, M. & Schulz, E. Turning large language models into cognitive models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.03917 (2023).
25. Cusumano-Towner, M. F., Saad, F. A., Lew, A. K. & Mansinghka, V. K. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proc. 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* 221–236 (2019).
26. Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K. & Tenenbaum, J. B. Church: a language for generative models. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence* 220–229 (2008).
27. Bingham, E. et al. Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**, 973–978 (2019).
28. Ge, H., Xu, K. & Ghahramani, Z. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics* 1682–1690 (PMLR, 2018).
29. Goodman, N. D., Tenenbaum, J. B. & Gerstenberg, T. *Concepts in a Probabilistic Language of Thought* Tech. Rep. (Center for Brains, Minds and Machines, 2014).
30. van Opheusden, B. et al. Expertise increases planning depth in human gameplay. *Nature* **618**, 1000–1005 (2023).
31. Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482 (2024).

32. Yao, S. et al. Tree of thoughts: deliberate problem solving with large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.10601 (2023).
33. Baker, C. L., Saxe, R. & Tenenbaum, J. B. Action understanding as inverse planning. *Cognition* **113**, 329–349 (2009).
34. Jara-Ettinger, J., Schulz, L. E. & Tenenbaum, J. B. The naive utility calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* **123**, 101334 (2020).
35. Zhi-Xuan, T., Ying, L., Mansinghka, V. & Tenenbaum, J. B. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proc. 23rd International Conference on Autonomous Agents and Multiagent Systems* 2094–2103 (2024).
36. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
37. Chu, J. & Schulz, L. E. Play, curiosity, and cognition. *Annu. Rev. Dev. Psychol.* **2**, 317–343 (2020).
38. Yanai, I. & Lercher, M. J. It takes two to think. *Nat. Biotechnol.* **42**, 18–19 (2024).
39. Holyoak, K. J. & Morrison, R. G. *The Cambridge Handbook of Thinking and Reasoning* (Cambridge Univ. Press, 2005).
40. Holyoak, K. J. & Morrison, R. G. *The Oxford Handbook of Thinking and Reasoning* (Oxford Univ. Press, 2012).
41. Ko, A. J. & Myers, B. A. Designing the whyline: a debugging interface for asking questions about program behavior. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 151–158 (2004).
42. Ko, A. J. et al. The state of the art in end-user software engineering. *ACM Comput. Surv.* **43**, 21 (2011).
43. Muggleton, S. & De Raedt, L. Inductive logic programming: theory and methods. *J. Log. Program.* **19**, 629–679 (1994).
44. Anderson, J. R. & Reiser, B. J. The lisp tutor. *Byte* **10**, 159–175 (1985).
45. Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. Cognitive tutors: lessons learned. *J. Learn. Sci.* **4**, 167–207 (1995).
46. Imai, S. Is GitHub Copilot a substitute for human pair-programming? An empirical study. In *Proc. ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings* 319–321 (2022).
47. Nguyen, N. & Nadi, S. An empirical evaluation of GitHub Copilot's code suggestions. In *Proc. 19th International Conference on Mining Software Repositories* 1–5 (2022).
48. Wermelinger, M. Using GitHub Copilot to solve simple programming problems. In *Proc. 54th ACM Technical Symposium on Computer Science Education Vol. 1* 172–178 (2023).
49. Barke, S., James, M. B. & Polikarpova, N. Grounded copilot: how programmers interact with code-generating models. *Proc. ACM Program. Lang.* **7**, 85–111 (2023).
50. Dakhel, A. M. et al. GitHub Copilot AI pair programmer: asset or liability? *J. Syst. Softw.* **203**, 111734 (2023).
51. Fisac, J. F. et al. Pragmatic–pedagogic value alignment. In *Robotics Research: The 18th International Symposium ISRR* 49–57 (Springer, 2020).
52. Ranz, F., Hummel, V. & Sihn, W. Capability-based task allocation in human–robot collaboration. *Procedia Manuf.* **9**, 182–189 (2017).
53. 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).
54. Shridhar, M. et al. Alfred: a benchmark for interpreting grounded instructions for everyday tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10740–10749 (2020).
55. Ahn, M. et al. Do as I can, not as I say: grounding language in robotic affordances. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.01691 (2022).
56. Raad, M. A. et al. Scaling instructable agents across many simulated worlds. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2404.10179 (2024).
57. Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S. & Kambhampati, S. PlanBench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In *Proc. of the 37th International Conf. on Neural Information Processing Systems (NIPS '23)*, 38975–38987 (Curran Associates, 2024).
58. Momennejad, I. et al. Evaluating cognitive maps and planning in large language models with cogeval. *Adv. Neural Inf. Process. Syst.* **36** (2024).
59. Goodman, N. D. & Frank, M. C. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* **20**, 818–829 (2016).
60. Sumers, T. R., Ho, M. K., Griffiths, T. L. & Hawkins, R. D. Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychol. Rev.* (2023).
61. Jeon, H. J., Milli, S. & Dragan, A. Reward-rational (implicit) choice: a unifying formalism for reward learning. *Adv. Neural Inf. Process. Syst.* **33**, 4415–4426 (2020).
62. Kollar, T. et al. Generalized grounding graphs: a probabilistic framework for understanding grounded language. *J. Artif. Intell. Res.* 1–35 (2013).
63. Bratman, M. E. *Shared Agency: A Planning Theory of Acting Together* (Oxford Univ. Press, 2013).
64. Stacy, S. et al. Modeling communication to coordinate perspectives in cooperation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2106.02164 (2021).
65. Wu, S. A. et al. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Top. Cogn. Sci.* **13**, 414–432 (2021).
66. Reddy, S., Dragan, A. D. & Levine, S. Where do you think you're going? Inferring beliefs about dynamics from behavior. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1805.08010 (2018).
67. Alanqary, A. et al. Modeling the mistakes of boundedly rational agents within a Bayesian theory of mind. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2106.13249 (2021).
68. Dragan, A. D., Lee, K. C. & Srinivasa, S. S. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human–Robot Interaction (HRI)* 301–308 (IEEE, 2013).
69. Miura, S. & Zilberstein, S. A unifying framework for observer-aware planning and its complexity. In *Uncertainty in Artificial Intelligence* 610–620 (PMLR, 2021).
70. Flower, L. & Hayes, J. R. A cognitive process theory of writing. *Coll. Compos. Commun.* **32**, 365–387 (1981).
71. Hayes, J. R. Modeling and remodeling writing. *Writ. Commun.* **29**, 369–388 (2012).
72. Lee, M., Liang, P. & Yang, Q. CoAuthor: designing a human–AI collaborative writing dataset for exploring language model capabilities. In *Proc. 2022 CHI Conference on Human Factors in Computing Systems* 1–19 (2022).
73. Lee, M. et al. A design space for intelligent and interactive writing assistants. *Proc. of the CHI Conference on Human Factors in Computing Systems*, 1–35 (2024).
74. Ippolito, D., Yuan, A., Coenen, A. & Burnam, S. Creative writing with an AI-powered writing assistant: perspectives from professional writers. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.05030 (2022).
75. Gero, K. I., Liu, V. & Chilton, L. Sparks: inspiration for science writing using language models. In *Proc. 2022 ACM Designing Interactive Systems Conference* 1002–1019 (2022).
76. Gero, K. I., Long, T. & Chilton, L. B. Social dynamics of AI support in creative writing. In *Proc. 2023 CHI Conference on Human Factors in Computing Systems* 1–15 (2023).

77. Dell'Acqua, F. et al. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* Working Paper (Harvard Business School Technology & Operations Management Unit, 2023).
78. Porter, J., Boyd, C., Skandari, M. R. & Laiteerapong, N. Revisiting the time needed to provide adult primary care. *J. Gen. Intern. Med.* **38**, 147–155 (2023).
79. Dewa, C. S., Loong, D., Bonato, S. & Trojanowski, L. The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review. *BMJ Open* **7**, e015141 (2017).
80. Chowdhery, A. et al. Palm: scaling language modeling with pathways. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.02311 (2022).
81. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
82. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
83. Tu, T. et al. Towards conversational diagnostic AI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2401.05654 (2024).
84. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med*. https://doi.org/10.1001/jamainternmed.2023.1838 (2023).
85. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **32**, 18069–18083 (2020).
86. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
87. Ghassemi, M. et al. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl. Sci. Proc.* **2020**, 191–200 (2020).
88. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol* **157**, 1362–1369 (2021).
89. Cabitza, F. & Zeitoun, J.-D. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Ann. Transl. Med.* **7** https://doi.org/10.21037/atm.2019.04.07 (2019).
90. Puig, X. et al. Watch-and-help: a challenge for social perception and human–AI collaboration. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.09890 (2020).
91. Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J. & Tenenbaum, J. Inferring the future by imagining the past. *Adv. Neural Inf. Process. Syst.* **36**, 21196–21216 (2024).
92. Fisac, J. F. et al. Generating plans that predict themselves. In *Algorithmic Foundations of Robotics XII: Proc. 12th Workshop on the Algorithmic Foundations of Robotics* 144–159 (Springer, 2020).
93. Grice, H. P. in *Speech* Acts (eds. Cole, P. & Morgan, J. L.) 41–58 (Brill, 1975).
94. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1702.08608 (2017).
95. Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019).
96. Smith, B. C. *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press, 2019).
97. Sucholutsky, I. & Griffiths, T. L. Alignment with human representations supports robust few-shot learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.11990 (2023).
98. Sucholutsky, I. et al. Getting aligned on representational alignment. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.13018 (2023).
99. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18327–18332 (2013).
100. Ho, M. K. & Griffiths, T. L. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annu. Rev. Control Robot. Auton. Syst.* **5**, 33–53 (2022).
101. Yang, S. C.-H., Folke, T. & Shafto, P. The inner loop of collective human–machine intelligence. *Top. Cogn. Sci*. https://doi.org/10.1111/tops.12642 (2023).
102. Steyvers, M. & Kumar, A. Three challenges for AI-assisted decision-making. *Perspect. Psychol. Sci*. https://doi.org/10.1177/17456916231181102 (2023).
103. Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. in *The Cambridge Handbook of Computational Psychology* (ed. Sun, R.) 59–100 (Cambridge Univ. Press, 2008).
104. Chater, N. & Manning, C. D. Probabilistic models of language processing and acquisition. *Trends Cogn. Sci.* **10**, 335–344 (2006).
105. Oaksford, M. & Chater, N. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning* (Oxford Univ. Press, 2007).
106. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
107. Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007).
108. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
109. Yildirim, I., Belledonne, M., Freiwald, W. & Tenenbaum, J. Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).
110. Allen, K. R., Smith, K. A. & Tenenbaum, J. B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl Acad. Sci. USA* **117**, 29302–29310 (2020).
111. Zhang, C. E., Wong, L., Grand, G. & Tenenbaum, J. B. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 45*, https://escholarship.org/uc/item/7018f2ss (2023).
112. Tenenbaum, J. Bayesian modeling of human concept learning. *Adv. Neural Inf. Process. Syst.* **11**, 59–65 (1998).
113. Goodman, N. D., Tenenbaum, J. B., Feldman, J. & Griffiths, T. L. A rational analysis of rule-based concept learning. *Cogn. Sci.* **32**, 108–154 (2008).
114. Piantadosi, S. T., Tenenbaum, J. B. & Goodman, N. D. The logical primitives of thought: empirical foundations for compositional cognitive models. *Psychol. Rev.* **123**, 392–424 (2016).
115. Griffiths, T., Steyvers, M., Blei, D. & Tenenbaum, J. Integrating topics and syntax. *Adv. Neural Inf. Process. Syst.* **17**, 537–544 (2004).
116. Goodman, N. D. & Lassiter, D. in *The Handbook of Contemporary Semantic Theory* (eds Lapin, S. & Fox, C.) 655–686 (John Wiley & Sons, 2015).
117. Yang, Y. & Piantadosi, S. T. One model for the learning of language. *Proc. Natl Acad. Sci. USA* **119**, e2021865119 (2022).
118. Schulz, L. E., Bonawitz, E. B. & Griffiths, T. L. Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Dev. Psychol.* **43**, 1124–1139 (2007).
119. Gopnik, A. et al. A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* **111**, 3–32 (2004).
120. Kirfel, L., Icard, T. & Gerstenberg, T. Inference from explanation. *J. Exp. Psychol. Gen.* **151**, 1481–1501 (2022).
121. Lagnado, D. A., Gerstenberg, T. & Zultan, R. Causal responsibility and counterfactuals. *Cogn. Sci.* **37**, 1036–1073 (2013).
122. Hemmer, P. & Steyvers, M. A Bayesian account of reconstructive memory. *Top. Cogn. Sci.* **1**, 189–202 (2009).

123. Ullman, T. D. & Tenenbaum, J. B. Bayesian models of conceptual development: learning as building models of the world. *Annu. Rev. Dev. Psychol.* **2**, 533–558 (2020).
124. Griffiths, T. L. & Tenenbaum, J. B. Theory-based causal induction. *Psychol. Rev.* **116**, 661–716 (2009).
125. Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. *Cogn. Sci.* **38**, 599–637 (2014).
126. Ho, M. K., Saxe, R. & Cushman, F. Planning with theory of mind. *Trends Cogn. Sci.* **26**, 959–971 (2022).
127. Hay, N., Russell, S., Tolpin, D. & Shimony, S. E. Selecting computations: theory and applications. In *Proc. of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* 346–355 (2012).
128. Tomov, M. S., Yagati, S., Kumar, A., Yang, W. & Gershman, S. J. Discovery of hierarchical representations for efficient planning. *PLoS Comput. Biol.* **16**, e1007594 (2020).
129. Baker, C. L. & Tenenbaum, J. B. in *Plan, Activity, and Intent Recognition: Theory and Practice* (eds Sukthankar, G. et al.) 177–204 (Morgan Kaufmann, 2014).
130. Callaway, F. et al. Rational use of cognitive resources in human planning. *Nat. Hum. Behav.* **6**, 1112–1125 (2022).
131. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).
132. Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J. & Mansinghka, V. Online Bayesian goal inference for boundedly rational planning agents. *Adv. Neural Inf. Process. Syst.* **33**, 19238–19250 (2020).
133. Ying, L. et al. The Neuro-Symbolic Inverse Planning Engine (NIPE): modeling probabilistic social inferences from linguistic inputs. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.14325 (2023).
134. Jara-Ettinger, J., Gweon, H., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn. Sci.* **20**, 589–604 (2016).
135. Johnson-Laird, P. N. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* 6 (Harvard Univ. Press, 1983).
136. Byrne, R. M. Mental models and counterfactual thoughts about what might have been. *Trends Cogn. Sci.* **6**, 426–431 (2002).
137. Shafto, P., Goodman, N. D. & Griffiths, T. L. A rational account of pedagogical reasoning: teaching by, and learning from, examples. *Cogn. Psychol.* **71**, 55–89 (2014).
138. Sumers, T. R., Ho, M. K., Hawkins, R. D., Narasimhan, K. & Griffiths, T. L. Learning rewards from linguistic feedback. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 35, 6002–6010 (2021).
139. Liquin, E. G., Luzuriaga, N. & Gureckis, T. M. Teaching and learning through pedagogical environment design. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 45*, https://escholarship.org/uc/item/9xq3w7rc (2023).
140. Kumar, A., Smyth, P. & Steyvers, M. Differentiating mental models of self and others: a hierarchical framework for knowledge assessment. *Psychol. Rev.* **130**, 1566–1591 (2023).
141. Hawkins, R. D. et al. From partners to populations: a hierarchical Bayesian account of coordination and convention. *Psychol. Rev.* **130**, 977–1016 (2023).
142. Hawkins, R. D. et al. Flexible social inference facilitates targeted social learning when rewards are not observable. *Nat. Hum. Behav.* **7**, 1767–1776 (2023).
143. Frank, M. C. & Goodman, N. D. Predicting pragmatic reasoning in language games. *Science* **336**, 998 (2012).
144. Goodman, N. D. & Stuhlmüller, A. Knowledge and implicature: modeling language understanding as social cognition. *Top. Cogn. Sci.* **5**, 173–184 (2013).
145. Ho, M. K., Cushman, F., Littman, M. L. & Austerweil, J. L. Communication in action: planning and interpreting communicative demonstrations. *J. Exp. Psychol. Gen.* **150**, 2246–2272 (2021).
146. Griffiths, T. L. Understanding human intelligence through human limitations. *Trends Cogn. Sci.* **24**, 873–883 (2020).
147. Tversky, A. & Kahneman, D. Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* **5**, 207–232 (1973).
148. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131 (1974).
149. Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N. & Sanborn, A. N. The autocorrelated Bayesian sampler: a rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychol. Rev.* **131**, 456–493 (2023).
150. Van Rooij, I. The tractable cognition thesis. *Cogn. Sci.* **32**, 939–984 (2008).
151. Icard, T. & Goodman, N. D. A resource-rational approach to the causal frame problem. In *CogSci* https://cocolab.stanford.edu/papers/IcardGoodman2015-Cogsci.pdf (2015).
152. Icard, T. Resource rationality. Preprint at https://philpapers.org/archive/ICARRT.pdf (2023).
153. Lieder, F. & Griffiths, T. L. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2020).
154. Anderson, J. R. *The Adaptive Character of Thought* (Psychology Press, 1990).
155. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The child as hacker. *Trends Cogn. Sci.* **24**, 900–915 (2020).
156. Cheyette, S. J., Callaway, F., Bramley, N. R., Nelson, J. D. & Tenenbaum, J. People seek easily interpretable information. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 45*, https://escholarship.org/uc/item/5sm2b484 (2023).
157. Saad, F. A. K. *Scalable Structure Learning, Inference, and Analysis with Probabilistic Programs*. PhD thesis, Massachusetts Institute of Technology (2022).
158. Saad, F. A., Cusumano-Towner, M. F., Schaechtle, U., Rinard, M. C. & Mansinghka, V. K. Bayesian synthesis of probabilistic programs for automatic data modeling. *Proc. ACM Program. Lang.* **3**, 37 (2019).
159. Li, M. Y., Fox, E., & Goodman, N. D. Automated statistical model discovery with language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.17879 (2024).
160. Lew, A. K., Tessler, M. H., Mansinghka, V. K. & Tenenbaum, J. B. Leveraging unstructured statistical knowledge in a probabilistic language of thought. In *Proc. Annual Conference of the Cognitive Science Society*, https://cognitivesciencesociety.org/cogsci20/papers/0520/0520.pdf (2020).
161. Lew, A., Agrawal, M., Sontag, D. & Mansinghka, V. Pclean: Bayesian data cleaning at scale with domain-specific probabilistic programming. In *International Conference on Artificial Intelligence and Statistics* 1927–1935 (PMLR, 2021).
162. Gothoskar, N. et al. Bayes3d: fast learning and inference in structured generative models of 3D objects and scenes. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2312.08715 (2023).
163. Mansinghka, V. K. et al. Probabilistic programming with programmable inference. In *Proc. 39th ACM SIGPLAN Conference on Programming Language Design and Implementation* 603–616 (2018).
164. Lew, A. K., Huot, M., Staton, S. & Mansinghka, V. K. Adev: sound automatic differentiation of expected values of probabilistic programs. *Proc. ACM Program. Lang.* https://doi.org/10.1145/3571198 (2023).

165. Becker, M. R. et al. Probabilistic programming with programmable variational inference. *Proc. ACM Program. Lang.* **8**, 2123–2147 (2024).
166. Saad, F. A., Rinard, M. C. & Mansinghka, V. K. Sppl: probabilistic programming with fast exact symbolic inference. In *Proc. 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation* 804–819 (2021).
167. Lew, A. K., Ghavamizadeh, M., Rinard, M. C. & Mansinghka, V. K. Probabilistic programming with stochastic probabilities. *Proc. ACM Program. Lang.* **7**, 1708–1732 (2023).
168. Guggenberger, T. M., Möller, F., Haarhaus, T., Gür, I. & Otto, B. Ecosystem types in information systems. In *Proc. 28th European Conference on Information Systems (ECIS2020)* https://aisel.aisnet.org/ecis2020_rp/45/ (2020).
169. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Mag.* **38**, 50–57 (2017).
170. Wachter, S. & Mittelstadt, B. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Bus. Law Rev.* **2**, 494–620 (2019).
171. Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K. & Chen, L. Generative AI and ChatGPT: applications, challenges, and AI–human collaboration. *J. Inform. Techn. Case Appl. Res.* **25**, 277–304 (2023).
172. Norman, D. *Design of Everyday Things* (Basic Books, 1988).
173. Chemero, A. in *How Shall Affordances Be Refined?* (ed. Jones, K. S.) 181–195 (Routledge, 2018).
174. Zerilli, J., Bhatt, U. & Weller, A. How transparency modulates trust in artificial intelligence. *Patterns* **3**, 100455 (2022).
175. Messeri, L. & Crockett, M. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
176. 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).
177. Steyvers, M., Tejeda, H., Kerrigan, G. & Smyth, P. Bayesian modeling of human–AI complementarity. *Proc. Natl Acad. Sci. USA* **119**, e2111547119 (2022).
178. Chandra, K., Li, T.-M., Nigam, R., Tenenbaum, J. & Ragan-Kelley, J. Watchat: explaining perplexing programs by debugging mental models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2403.05334 (2024).
179. Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J. & Tenenbaum, J. Cooperative explanation as rational communication. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 46*, https://escholarship.org/uc/item/8bf5g4h6 (2024).
180. Hadfield-Menell, D., Russell, S. J., Abbeel, P. & Dragan, A. Cooperative inverse reinforcement learning. *Adv. Neural Inf. Process. Syst.* **29**, 3916–3924 (2016).
181. Chandra, K., Li, T.-M., Tenenbaum, J. & Ragan-Kelley, J. Acting as inverse inverse planning. In *ACM SIGGRAPH 2023 Conference Proceedings* 1–12 (2023).
182. Chen, T., Houlihan, S. D., Chandra, K., Tenenbaum, J. & Saxe, R. Intervening on emotions by planning over a theory of mind. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 46*, https://escholarship.org/uc/item/4gz7c85c (2024).
183. Chandra, K., Li, T.-M., Tenenbaum, J. B. & Ragan-Kelley, J. Storytelling as inverse inverse planning. *Top. Cogn. Sci.* **16**, 54–70 (2024).
184. Blei, D. M., Jordan, M. I., Griffiths, T. L. & Tenenbaum, J. B. Hierarchical topic models and the nested Chinese restaurant process. In *Proc. of the 16th International Conf. on Neural Information Processing Systems (NIPS'03)*, 17–24 (MIT Press, 2003).
185. Loula, J. et al. Learning generative population models from multiple clinical datasets via probabilistic programming. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery* https://openreview.net/forum?id=Sm1KnFlx0H (2024).
186. Cabitza, F., Rasoini, R. & Gensini, G. F. Unintended consequences of machine learning in medicine. *JAMA* **318**, 517–518 (2017).
187. Mozannar, H. & Sontag, D. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning* 7076–7087 (PMLR, 2020).
188. Dvijotham, K. et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Med.* **29**, 1814–1820 (2023).
189. Tsvetkova, M., Yasseri, T., Pescetelli, N. & Werner, T. Human–machine social systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.14410 (2024).
190. Schneiders, E., Cheon, E., Kjeldskov, J., Rehm, M. & Skov, M. B. Non-dyadic interaction: a literature review of 15 years of human–robot interaction conference publications. *ACM Trans. Hum. Robot Interact.* **11**, 13 (2022).
191. Hornecker, E., Krummheuer, A., Bischof, A. & Rehm, M. Beyond dyadic HRI: building robots for society. *Interactions* **29**, 48–53 (2022).
192. Yadav, A. & Mehta, R. Beyond dyadic interactions: assessing trust networks in multi-human–robot teams. In *Companion of the 2024 ACM/IEEE International Conference on Human–Robot Interaction* 1153–1157 (2024).
193. Sucholutsky, I. et al. Representational alignment supports effective machine teaching. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2406.04302 (2024).
194. Li, L. et al. Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery. *Cyborg Bionic Syst.* **4**, 0042 (2023).
195. Boyce, V., Hawkins, R. D., Goodman, N. D. & Frank, M. C. Interaction structure constrains the emergence of conventions in group communication. *Proc. Natl Acad. Sci. USA* **121**, e2403888121 (2024).
196. Trouille, L., Lintott, C. J. & Fortson, L. F. Citizen science frontiers: efficiency, engagement, and serendipitous discovery with human–machine systems. *Proc. Natl Acad. Sci. USA* **116**, 1902–1909 (2019).
197. Hornbæk, K. & Oulasvirta, A. What is interaction? In Proc. *2017 CHI Conference on Human Factors in Computing Systems* 5040–5052 (2017).
198. Lee, M. et al. Evaluating human–language model interaction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.09746 (2022).
199. Allen, K. et al. Using games to understand the mind. *Nat. Hum. Behav.* **8**, 1035–1043 (2024).
200. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2304.03442 (2023).
201. Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. *Science* **365**, 885–890 (2019).
202. Bakhtin, A. et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science* **378**, 1067–1074 (2022).
203. Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **151**, 90–103 (2019).
204. Green, B. & Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum. Comput. Interact.* **3**, 50 (2019).
205. Inuwa-Dutse, I., Toniolo, A., Weller, A. & Bhatt, U. Algorithmic loafing and mitigation strategies in human–AI teams. *Comput. Hum. Behav. Artif. Hum.* **1**, 100024 (2023).
206. Hofman, J. M., Goldstein, D. G. & Rothschild, D. M. Steroids, sneakers, coach: the spectrum of human–AI relationships. *SSRN* https://doi.org/10.2139/ssrn.4578180 (2023).

207. Buschek, D., Zürn, M. & Eiband, M. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native English writers. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1-13 (Association for Computing Machinery, 2021).
208. 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. *Proc. ACM Hum. Comput. Interact.* **5**, 188 (2021).
209. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
210. Dietvorst, B. J., Simmons, J. P. & Massey, C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* **64**, 1155–1170 (2018).
211. Zerilli, J., Bhatt, U. & Weller, A. Transparency modulates trust in artificial intelligence. *Patterns* **3**, 100455 (2022).
212. Mumford, L. *Technics and Civilization* (Routledge & Kegan Paul, 1936).
213. Weizenbaum, J. *Computer Power and Human Reason: From Judgment to Calculation* (W. H. Freeman & Co., 1976).
214. Weidinger, L. et al. Taxonomy of risks posed by language models. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 214–229 (2022).
215. Zhuang, S. & Hadfield-Menell, D. Consequences of misaligned AI. *Adv. Neural Inf. Process. Syst.* **33**, 15763–15773 (2020).
216. Kalai, A. T. & Vempala, S. S. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing* 160-171 (2024).
217. Amodei, D. et al. Concrete problems in AI safety. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1606.06565 (2016).
218. Russell, S. *Human Compatible: AI and the Problem of Control* (Viking, 2019).
219. Russell, S. in *Perspectives on Digital Humanism* (eds Werthner, H. et al.) 19–24 (Springer Cham, 2021).
220. Carroll, M., Chan, A., Ashton, H. & Krueger, D. Characterizing manipulation from AI systems. In *Proc. 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* 1–13 (2023).
221. Lazar, S. & Nelson, A. AI safety on whose terms? *Science* **381**, 138 (2023).
222. Roncone, A., Mangin, O. & Scassellati, B. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* 1014–1021 (IEEE, 2017).
223. Carroll, M. et al. On the utility of learning about humans for human–AI coordination. *Adv. Neural Inf. Process. Syst.* **32**, 5174–5185 (2019).
224. Macindoe, O., Kaelbling, L. P. & Lozano-Pérez, T. Pomcop: belief space planning for sidekicks in cooperative games. In *Proc. AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* Vol. 8, 38–43 (2012).
225. Lin, J., Fried, D., Klein, D., & Dragan, A. Inferring rewards from language in context. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 8546–8560 (2022).
226. Keuning, H., Jeuring, J. & Heeren, B. A systematic literature review of automated feedback generation for programming exercises. *ACM Trans. Comput. Educ.* **19**, 3 (2018).
227. Sarsa, S., Denny, P., Hellas, A. & Leinonen, J. Automatic generation of programming exercises and code explanations using large language models. In *Proc. 2022 ACM Conference on International Computing Education Research* Vol. 1, 27–43 (2022).
228. Head, A., Appachu, C., Hearst, M. A. & Hartmann, B. Tutorons: generating context-relevant, on-demand explanations and demonstrations of online code. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* 3–12 (IEEE, 2015).
229. Rafferty, A. N., Jansen, R. A. & Griffiths, T. L. Assessing mathematics misunderstandings via Bayesian inverse planning. *Cogn. Sci.* **44**, e12900 (2020).
230. Poesia, G. & Goodman, N. D. Peano: learning formal mathematical reasoning. *Philos. Trans. R. Soc. A* **381**, 20220044 (2023).
231. Slonim, N. et al. An autonomous debating system. *Nature* **591**, 379–384 (2021).
232. Jarrett, D. et al. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop* https://openreview.net/forum?id=sv7KZcUqu1 (2023).
233. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.14325 (2023).
234. Bakker, M. et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Adv. Neural Inf. Process. Syst.* **35**, 38176–38189 (2022).
235. Small, C., Bjorkegren, M., Erkkilä, T., Shaw, L. & Megill, C. Polis: scaling deliberation by mapping high dimensional opinion spaces. *Recerca* **26**, https://doi.org/10.6035/recerca.5516 (2021).
236. Huot, M. et al. Gensql: a probabilistic programming system for querying generative models of database tables. *Proc. ACM Program. Lang.* **8**, 790–815 (2024).
237. Steinruecken, C., et al.) 161–173 (Springer Cham, 2019).
238. Davies, A. et al. Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
239. Cranmer, M. et al. Discovering symbolic models from deep learning with inductive biases. *Adv. Neural Inf. Process. Syst.* **33**, 17429–17442 (2020).
240. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
241. Ashkinaze, J., Mendelsohn, J., Qiwei, L., Budak, C. & Gilbert, E. How AI ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. Preprint at *arXiv*, https://doi.org/10.48550/arXiv.2401.13481 (2024).
242. Suri, S. et al. *The Use of Generative Search Engines for Knowledge Work and Complex Tasks* Tech. Rep. MSR-TR-2024-9 (Microsoft, 2024).
243. Vartiainen, H. & Tedre, M. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digit. Creat.* **34**, 1–21 (2023).
244. Gafni, O. et al. Make-a-scene: scene-based text-to-image generation with human priors. In *European Conference on Computer Vision* 89–106 (Springer, 2022).
245. Fan, J. E., Dinculescu, M. & Ha, D. Collabdraw: an environment for collaborative sketching with an artificial agent. In *Proc. 2019 Conference on Creativity and Cognition* 556–561 (2019).
246. Ge, S., Goswami, V., Zitnick, C. L. & Parikh, D. Creative sketch generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2011.10039 (2020).
247. Dvorožňák, M. et al. Monster mash: a single-view approach to casual 3D modeling and animation. *ACM Trans. Graph.* **39**, 214 (2020).
248. Chater, N. & Oaksford, M. (eds) *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (Oxford Univ. Press, 2008).
249. Spelke, E. S. Core knowledge. *Am. Psychol.* **55**, 1233–1243 (2000).
250. Piantadosi, S. T. The computational origin of representation. *Minds Mach.* **31**, 1–58 (2021).

251. Quilty-Dunn, J., Porot, N. & Mandelbaum, E. The best game in town: the reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behav. Brain Sci.* **46**, e261 (2023).
252. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl Acad. Sci. USA* **105**, 10687–10692 (2008).
253. Ellis, K. et al. Dreamcoder: bootstrapping inductive program synthesis with wake-sleep library learning. In *Proc. 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation* 835–850 (2021).
254. Lieder, F., Chen, O. X., Krueger, P. M. & Griffiths, T. L. Cognitive prostheses for goal achievement. *Nat. Hum. Behav.* **3**, 1096–1106 (2019).
255. Newell, A. & Simon, H. A. *Human Problem Solving* (Prentice-Hall, 1972).
256. Mattar, M. G. & Lengyel, M. Planning in the brain. *Neuron* **110**, 914–934 (2022).
257. Ho, M. K. et al. People construct simplified mental representations to plan. *Nature* **606**, 129–136 (2022).
258. Baker, C., Saxe, R. & Tenenbaum, J. Bayesian theory of mind: modeling joint belief–desire attribution. In *Proc. Annual Meeting of the Cognitive Science Society Vol. 33*, https://escholarship.org/content/qt5rk7z59q/qt5rk7z59q.pdf (2011).
259. Degen, J. The rational speech act framework. *Annu. Rev. Linguist.* **9**, 519–540 (2023).
260. Binz, M. et al. Meta-learned models of cognition. *Behav. Brain Sci.* **47**, e147 (2024).
261. Grant, E., Finn, C., Levine, S., Darrell, T. & Griffiths, T. Recasting gradient-based meta-learning as hierarchical Bayes. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1801.08930 (2018).
262. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).

Acknowledgements

We thank R. Turner, L. Schulz, T. Brooke-Wilson, V. Chen, A. Rote, L. Ying, T. Chen, M. Ashman, M. Walmsley, A. Jiang, M. Jamnik, D. Dvijotham, J. Ragan-Kelley, W. Crichton, A. Lew, T. O'Donnell, J. Loula, M. Tenenbaum, M. McNaughton-Collins and J. Collins for valuable conversations that informed this work. K.M.C. acknowledges support from the Marshall Commission and the Cambridge Trust. U.B. acknowledges support by European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement no. 101070617. I.S. acknowledges funding from an NSERC fellowship (no. 567554-2022). K.C. is supported by the Hertz Foundation, the Paul and Daisy Soros Fellowship and an NSF Graduate Research Fellowship under grant no. 1745302. M.L. acknowledges funding from MSR. T.Z.-X. acknowledges support from the OpenPhilanthropy AI Fellowship. V.M. acknowledges a gift from the Siegel Family Foundation. A.W. acknowledges support from a Turing AI Fellowship under grant no. EP/V025279/1, the Alan Turing Institute and the Leverhulme Trust via CFI. T.L.G. acknowledges support from ONR grant no. N00014-22-1-2813. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the institutions listed above. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors contributed to the writing of the work. K.M.C., I.S., U.B., T.L.G., A.W. and J.B.T. conceived the work. K.M.C., I.S., U.B., K.C. and L.W. contributed equally. M.L., C.E.Z., T.Z.-X. and M.H. contributed equally. V.M., A.W., J.B.T. and T.L.G. contributed equally to senior advising.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materialsshould be addressed to Katherine M. Collins.

Peer review information *Nature Human Behaviour* thanks Anthony Chemero, Tim Miller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.