

Building Machines that Learn and Think *with* People

Katherine M. Collins^{*1}, Ilia Sucholutsky^{*2}, Umang Bhatt^{*3,4}, Kartik Chandra^{*5}, Lionel Wong^{*5}, Mina Lee^{†6,7}, Cedegao E. Zhang^{‡5}, Tan Zhi-Xuan^{‡5}, Mark Ho^{‡3}, Vikash Mansinghka^{‡5}, Adrian Weller^{‡1,4}, Joshua B. Tenenbaum^{‡5}, and Thomas L. Griffiths^{‡2}

¹University of Cambridge

²Princeton University

³NYU

⁴The Alan Turing Institute

⁵MIT

⁶Microsoft Research

⁷University of Chicago

Abstract

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 (AI) 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 AI thought partners can engage and 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.

1 Introduction

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, thirty 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 AI technologies, particularly language models, have catalyzed this transition. 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, at least communicated often through the medium of language³. We long have – from developing new modes of thinking through questioning and debate to teaching and learning through language. The apparent power of these new systems – getting closer to the kind of artificial intelligence (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

*Contributed equally.

†Contributed equally.

‡Equal senior role.

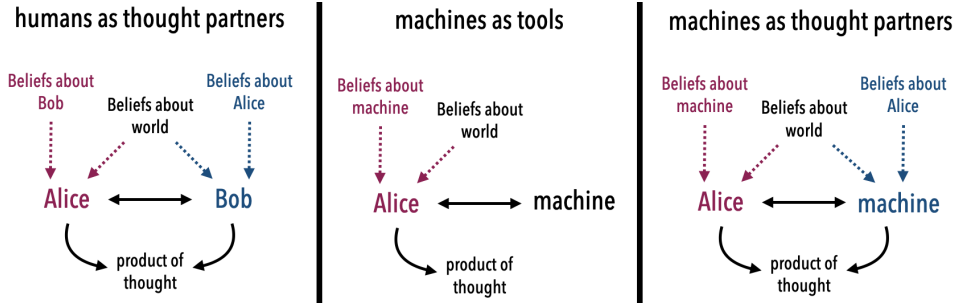


Figure 1: Examples of ecosystems for thinking. Humans have long thought together. Machines expanded the efficiency of human thinking. Now, machines – powered by AI – open up new realms of computational thought partnership with humans.

systems that truly act as effective thought partners. We argue that good thought partners are systems (1) which can understand us, (2) which we can understand, and (3) which 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 (e.g., LLMs¹⁰) with large amounts of human demonstrations and feedback, along with “traces” of human thought scraped from web-scale data^{11–13}. While such an approach has produced systems that accurately mimic human *behavior* (e.g., producing fluent text), these machines do not robustly simulate human *cognition* (e.g., 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 task, world, and human (where these models are structured²¹, 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 behavioral sciences provide valuable insights for designing human-centric, uncertainty-aware thought partners. Drawing on such research, we argue that effective thought partners are those which *build models of the human and the world*.

This toolkit includes foundation models^{22–24}, but is not limited to them. Indeed, foundation models like LLMs are fueling new motifs for thinking about human minds in computational terms (e.g., “rational meaning construction”¹⁶) interleaved *alongside* techniques from probabilistic programming^{25–29}, goal-directed search^{30–32}, and other explicit, structured representations, e.g., of agents thinking about other agents^{33–35}. We already have tools that help us build machines that learn and think *like* people³⁶. We propose applying that toolkit to collaborative cognition – to build machines that learn and think *with* people.

2 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 ones³⁷. And we *think together*. For generations³⁷, 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 fueled by engaging with diverse partners who offer new ideas yet share our values³⁸.

2.1 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.

2.2 Example Domains

We next outline a few diverse domains in which the development of AI thought partners able to truly collaborate with humans (Figure 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 program^{41–45}. New “programming assistant” tools like GitHub Copilot have rapidly gained enormous popularity and attention; however, these tools are often unreliable^{46–48}, e.g., failing to understand users’ intentions⁴⁹ and generating bugs that may be particularly risky alongside beginner programmers⁵⁰. Programming 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 need partners who can answer our questions – like *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 do and do not know.

Thought Partners for Embodied Assistance. Ensuring 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 cook⁵¹, to working with someone with different physical abilities⁵², or carrying out a high-stakes search-and-rescue mission⁵³. While 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 possibilities⁶², while also understanding that these can be *shared* across multiple minds^{63–65}. In addition, effective partners account for each others’ limitations in perception, planning, and world modeling, correcting for possible mistakes^{66,67}, and acting so as 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 from helping 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 sensemake, 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 patient⁷⁸, driving high rates of burnout with knock-on effects on patient care quality⁷⁹. 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 patients⁸⁴. 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}).

2.3 Desiderata

What then do we want from thought partners? There are many criteria for tools for thought that are of course relevant: efficiency, accuracy, robustness, fairness, cost, scalability, etc. But the domains above illuminate that what is distinctive about a thought *partner* is its relationship to the *user*⁸⁹. Looking to ideas the behavioral sciences motivates three desiderata to guide the design of human-centered 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 in order 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 reality⁹⁶. This means being accurate and knowledgeable, but also working with a shared representation of the world, domain, or task^{97–99}. Further, our use of ‘*we*’ emphasizes that thought partnerships are fundamentally about *synergy*, moving beyond the sum of its parts.

3 Engineering Human-Centered 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 ends²⁰. 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 behavior. 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^{22,155,156}.

3.1 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 – and 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

Mode	Ongoing Challenges	Sampling of Existing Systems
Collaborative planning <ul style="list-style-type: none"> • Joint decision-making • Decentralized cooperation • Goal and task assistance 	Reliable goal inference Value and intent alignment Scalable multi-agent planning	Collaborative robots ^{68,100} Video game sidekicks ^{101,102} Language-based assistants ^{35,103}
Collaborative learning <ul style="list-style-type: none"> • Pair & team problem-solving • Identification of knowledge gaps • New problem construction 	Strong & robust problem-solving abilities Personalized curriculum pacing Problem construction of targeted difficulty	Programming learning aids ^{104–107} Mathematics tutors ^{15,108,109}
Collaborative deliberation <ul style="list-style-type: none"> • Debate & argumentation • Critical review & discussion • Consensus formation 	Opinion diversity Verifiable reasoning Formation of common ground	Machine-assisted debating ^{110–112} Consensus writing & opinion mapping ^{113,114}
Collaborative sensemaking <ul style="list-style-type: none"> • Explanation • Visualization • Data Analytics 	Exponential increases in data produced Accessible communication Fidelity of insights to the world	Probabilistic data modeling ^{115–119} Machine-assisted theory discovery ^{120–122}
Collaborative creation & ideation <ul style="list-style-type: none"> • Co-design • Idea critiquing • Brainstorming 	Generation diversity Style consistency Modular customizability	Machine-assisted writing ^{72,74,123} Prompted image creation ^{124–126} Collaborative sketching ^{127–129}

Table 1: Modes of collaborative thought. Settings in which human-human and human-AI thought partners can engage.

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 behavior^{18,157}. 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.

3.2 Computational Cognitive Science Motifs

We now (non-exhaustively) spotlight several key insights about modeling humans, modeling humans modeling humans, and modeling humans modeling the world from computational cognitive science – “motifs” for reverse engineering the mind (Table 2) – that we believe can inform engineering of human-centered thought partners. While we acknowledge that there are communities within cognitive science that 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.

Motif	Description	Sample References
Probabilistic Mental Models and Inference	Humans update beliefs and draw inferences consistent with probabilistic generative models representing the world.	21,130,131
Structured Knowledge Representations	Humans have abstract, highly structured conceptual representations that include causality, agents, and physical representations.	132–134
Hierarchical Models	Humans construct and update <i>hierarchical</i> representations that separate concrete knowledge and belief from abstract ones.	135–137
Theory Learning as Program Synthesis	Humans minds can be viewed as growing and editing theories of the world, expressed as programs, to “improve” their codebase (world models).	138–140
Resource-Rationality	Humans make rational choices about how to allocate finite computational resources, including time and memory.	141–143
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.	144–146
Bayesian Theory of Mind (BToM)	Humans represent <i>other</i> agents as intentional, intelligent actors; and probabilistically infer their mental states from observations of actions.	147–149
Rational Speech Acts (RSA)	Humans reason about language as an intentional, communicative action to infer a speakers’ underlying goals.	59,150,151
Learning to Learn	Humans <i>meta-learn</i> (improve our overarching ability to learn) jointly with learning new concrete concepts and skills.	36,152–154

Table 2: Bayesian Thought Partner Toolkit. 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.

Probabilistic Models of Cognition. Decades of work in computational cognitive science have demonstrated the power of modeling aspects of human cognition as Bayesian inference through structured probabilistic generative world models^{21,131,137,158,159}. Such approaches have found empirical success in capturing a diversity of facets of human cognition from early word learning¹⁶⁰, to visual perception^{161,162}, physical reasoning^{99,163,164}, concept learning^{165–167}, language processing and acquisition^{158,168–170}, causal inference in children^{171,172} and adults^{173,174}, memory reconstruction¹⁷⁵, and theory formation^{176,177}, among many others. Probabilistic models of cognition, particularly those built using a Bayesian approach, have offered principled formalisms in capturing rapid belief updating¹⁷⁸ and how we may integrate our commonsense world knowledge with new evidence to inform the actions and decisions we take in the world¹⁴⁹. Probabilistic inference over structured representations, particularly drawing on Bayesian modeling and tools like meta-level Markov Decision Processes¹⁷⁹, has provided a computational account of how humans plan so flexibly, with the capability of forming rich hierarchical goals and subgoals, across varied timescales^{149,155,180–182}.

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 (ToM) have offered strong accounts for how we may rapidly reason about each others’ beliefs, desires, goals, and intentions^{33,147,183–185}. We may build mental models^{186,187} of our thought partners, which can in turn be used to support communication and collaboration, informing the way we teach^{188–190}, infer whether to rely on a partner for help¹⁹¹, and support rapid, flexible adaptation to new conversation partners^{192,193}. We call particular attention to the Rational Speech Act (RSA) framework^{59,150}, which models communicative partners as recursively reasoning about each others’ 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 modeled with the same computational toolbox^{194,195}.

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 others¹⁹⁶. Thus, we sometimes make systematically biased inferences^{197,198}. Such “erroneous” judgments can be captured by modeling humans as making rational use of our finite resources; e.g., via approximate inference^{178,199} or bounded planning⁶⁷. Crucially, human cognition is tractable²⁰⁰. 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^{17,142,143,182,196,201,202}. One approach to modeling 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^{138,176}. Such representations – bolstered by tools like program synthesis – help explore suboptimal behavior²⁰³.

3.3 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 programming²⁶ 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^{116,204}, and synthesized via LLMs that encode rich priors^{16,118,205}. Probabilistic programs also enable fast approximate inference in world models that cohere with human common-sense knowledge and domain expertise^{115,206}, where the learned models are themselves amenable to modular inspection and editing by humans. Modern probabilistic programming languages^{25,27,207} offer not just generic inference but *programmable* inference, that is, they automate the math for hybrids of optimization^{208,209}, dynamic programming²¹⁰, and Monte Carlo inference²¹¹. While 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.

3.4 Infrastructure around Thought Partners

The design of systems that learn and think with people necessitates not only careful construction of the thought partner (i.e., the machine itself), but also the *infrastructure* within which human and computational thought partners collaborate¹⁵⁷. Questions like “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^{73,212–215}, and crucially informed by studies of human behavior. For example,

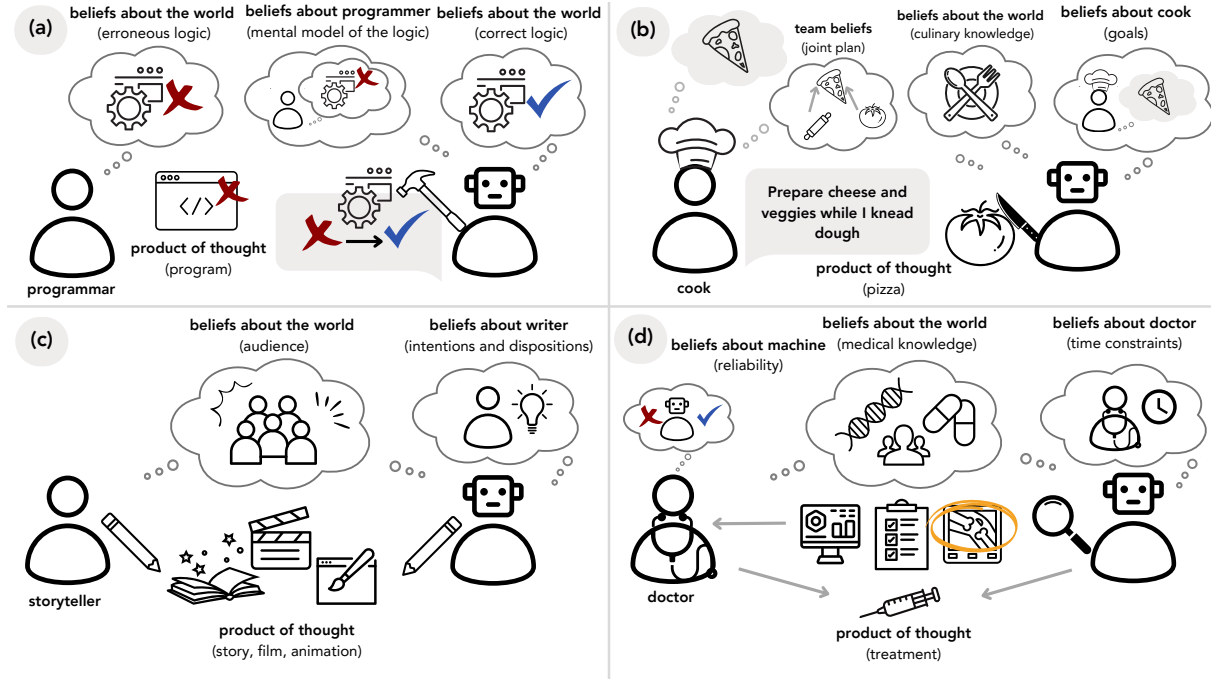


Figure 2: Case Study Depictions. (a) WatChat infers the user’s buggy mental model of the programming environment and interactively helps “patch” bug(s) in their understanding; (b) CLIPS reasons explicitly about agents’ goals, integrating (culinary) world knowledge and the human’s utterances to infer 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 a human doctor across modalities, a shared understanding of biology and patient needs, and a model of others’ limitations.

Article 14 of the EU 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 begets not only careful design of thought partners (e.g., that we can understand), but demands careful design of the system of affordances^{216,217} 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 behavior, e.g. overreliance²¹⁸ and “illusions of understanding”²¹⁹. Bayesian modeling has already found success in inferring humans’ reliance strategies²²⁰ and regions of the task space where a human versus machine can complement one another²²¹.

4 Case Studies in Engineering Thought Partners

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

4.1 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.¹⁰⁶, is to apply the Bayesian toolkit to *explain surprising behavior* of computer programs in a human-like way. Chandra et al., apply

Bayesian models of mental state inference and rational communication²²² to design a system called “WatChat” that answers questions like “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 (see Table 2). Such a framework can also be inverted to help design new questions for teachers or self-driven learners to identify misconceptions.

4.2 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 modeled in a Bayesian architecture called Cooperative Language-Guided Inverse Plan Search (CLIPS)³⁵. By modeling humans as cooperative planners who use language to communicate *joint plans* to achieve their goals⁶⁵, 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 *pro-actively* assisting with the goals that underlie the instruction. For example, CLIPS can understand the likely intentions behind an instruction like “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 infrastructure²⁵ to unite algorithms for Bayesian inverse planning^{33,184} and human-AI alignment^{51,61,223} 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.

4.3 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 ToM can predict and even design interventions on the audience’s experience of a story^{224,225}. Chandra et al. conceive of storytelling as “inverse inverse planning”: that is, starting with human social cognition, modeled as Bayesian inverse planning³³, 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}.

4.4 Thought Partners for Medicine

Finally, we envision medical thought partners 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. While no system yet meets our desiderata for these criteria, we believe a range of motifs and tools from the Bayesian thought partner toolkit here can support the development of such systems for collaborative sensemaking and deliberation. We imagine Bayesian thought partners that can update their medical world knowledge in light of new insights in biology, e.g., editing a code snippet of the underlying probabilistic world model¹⁶ or growing the representation

in a non-parametric hierarchical Bayesian model¹³⁵. Such a model can then, similar to WatChat, synthesize new questions to ensure the human doctor’s own medical world model is sound. Early work demonstrates that we can employ elements of our toolkit, specifically probabilistic programming, to learn rich generative models for oncology and support efficient user queries²²⁷. Yet, effective medical thought partners beckon a broader view of the ecosystem in which they are deployed^{89,228}. If a doctor is over-relying on the output of the thought partner, or overburdened amidst 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 deemed necessary of collaborative planning²²⁹. Systems for routing based on probabilistic modeling are already proving successful in simulation²³⁰.

5 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-centered program of building machines that learn and think *with* people.

5.1 Non-Dyad Settings

While there is substantial work to be done characterizing the space of possibilities for a single human and single AI thought partner (“dyadic”), we envision a future where *many* humans and *many* machines engage (“non-dyadic”), across roles and specialties in increasingly complex social systems²³¹, engage in the realm of thought^{232–234}. 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 partner²³⁵ and medical robot collision avoidance systems that need to account for multiple humans²³⁶. As in the dyad setting, extensions to non-dyadic settings can be bolstered by a deepening understanding of human behavior in groups – expanding the Bayesian thought partner toolkit – as is already underway in the study of convention formation^{192,237}. 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 impact human participation and performance with outcomes including both iterative scientific progress and serendipitous scientific discovery²³⁸.

5.2 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*²³⁹. Recent works have highlighted deficits in static evaluation of foundation models^{15,240}, 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 to study 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^{241–244}.

5.3 Risks and Important Considerations

Computational thought partners are by no means a guaranteed nor universal good and come with certain risks. We call out three such spheres of risk: (i) reliance, critical thinking, and access, (ii) anthropomorphization, and (iii) misalignment.

First, AI thought partners could induce over-reliance and impair the development of critical thinking skills^{219,245–247}, potentially acting as “steroids” for the mind²⁴⁸. We are concerned about these risks; our emphasis on the *infrastructure* around thought partner use is explicitly intended to help practitioners take steps to address these challenges, motivating further design of infrastructure modifications like cognitive forcing functions^{249,250}. Conversely, it is possible that 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^{251–253}. Already, research has found that the kinds of queries people make of AI systems can be informed by the amount of prior experience they have interacting with chatbots¹⁵ meaning 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 partners²⁵⁴. Our desiderata “I understand you” advocates for thought partners whose behavior we understand; while this could draw on how we understand other humans, however, we should be careful about *interpreting* such machine thought partners *as* we do humans. As Weizenbaum⁶ 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 intelligence²⁵⁵. Human-like thought partners should maintain categorical delineation between humans and machines to prevent overreliance^{245,256} and promote human dignity without encroaching on any partner’s self-worth²⁵⁷. The term used to refer to a thought partner can affect the assumptions made about their capabilities (e.g., *teammate* implies the machine and human are on equal footing) or can detract from a partner’s human-like nature (e.g., *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 goals²⁵⁸, provide wrong or misleading information²⁵⁹, or violate safety constraints²⁶⁰. A Bayesian approach to thought partnership can address some of these issues, enabling uncertainty-aware decision-making that avoids overconfidence^{223,261,262}. Yet, while inferring human thoughts and behavior can be used to design better collaborators, models of humans are inherently dual-use and can also be used to mislead, surveil, or manipulate²⁶³. It is crucial to consider whether thought partners are aligned with society at large, or merely superficially aligned with users while serving more powerful interests²⁶⁴.

6 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 amongst behavioral scientists, AI practitioners, domain experts, and related disciplines is crucial as we strive to build machines that truly learn and think *with* people.

Glossary of main terms

- Collaborative cognition: the process by which two or more agents work together in some aspect(s) of thinking (e.g., planning together, learning together, 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 which 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 (e.g., human-human, human-AI, AI-AI).
- Resource-rationality: the idea that human behavior and cognition can be viewed as *rational* under bounded constraints (e.g., under limited working memory).
- Probabilistic generative model: a model of how the data one observes about the world is 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 (e.g. 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.

Acknowledgments

We thank Richard Turner, Laura Schulz, Tyler Brooke-Wilson, Valerie Chen, Alena Rote, Lance Ying, Tony Chen, Matt Ashman, Mike Walmsley, Albert Jiang, Mateja Jamnik, Dj Dvijotham, Jonathan Ragan-Kelley, Will Crichton, Alex Lew, Tim O’Donnell, Joao Loula, Marty Tenenbaum, Mary McNaughton-Collins, and Jim Collins for valuable conversations that informed this work. KMC gratefully acknowledges support from the Marshall Commission and the Cambridge Trust. UB acknowledges support by ELSA (European Lighthouse on Secure and Safe AI) funded by the European Union under grant agreement No. 101070617; IS acknowledges funding from an NSERC fellowship (567554-2022); KC is supported by the Hertz Foundation, the Paul and Daisy Soros Fellowship, and an NSF Graduate Research Fellowship under grant #1745302.; ML acknowledges funding from MSR; TZX acknowledges support from the OpenPhilanthropy AI Fellowship. VM acknowledges a gift from the Siegel Family Foundation. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust

via CFI. TLG acknowledges support from ONR grant N00014-22-1-2813. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the institutions listed above.

References

- [1] GitHub Copilot · Your AI pair programmer. <https://github.com/features/copilot>.
- [2] Copilot for Microsoft 365 – Microsoft Adoption. <https://adoption.microsoft.com/en-us/copilot/>.
- [3] Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.
- [4] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433, 1950.
- [5] Manfred E Clynes and Nathan S Kline. Cyborgs and space. *Astronautics*, 14(9):26–27, 1960.
- [6] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [7] Ben Shneiderman. . In *Human-Centered AI*. Oxford University Press, 01 2022. ISBN 9780192845290. doi: 10.1093/oso/9780192845290.001.0001.
- [8] Alan Bundy. The computer modelling of mathematical reasoning. 1983.
- [9] John R Anderson, C Franklin Boyle, Albert T Corbett, and Matthew W Lewis. Cognitive modeling and intelligent tutoring. 1990.
- [10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, and Simran Arora et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.
- [11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, and Carroll et al Wainwright. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, and Shane et al Legg. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- [13] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, and Yuqing et al Du. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- [14] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.
- [15] Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, and Miri et al Zilka. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.
- [16] Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, and Vikash K et al Mansinghka. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, pages arXiv–2306, 2023.
- [17] Cedegao E Zhang, Katherine M Collins, Adrian Weller, and Joshua B Tenenbaum. Ai for mathematics: A cognitive science perspective. *arXiv preprint arXiv:2310.13021*, 2023.
- [18] Hyowon Gweon, Judith Fan, and Been Kim. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023.
- [19] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, and Joshua B et al Tenenbaum. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.
- [20] R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

- [21] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.
- [22] Thomas L. Griffiths, Jian-Qiao Zhu, Erin Grant, and R. Thomas McCoy. Bayes in the age of intelligent machines, 2023.
- [23] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.
- [24] Marcel Binz and Eric Schulz. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*, 2023.
- [25] Marco F Cusumano-Towner, Feras A Saad, Alexander K Lew, and Vikash K Mansinghka. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th acm sigplan conference on programming language design and implementation*, pages 221–236, 2019.
- [26] Noah D Goodman, Vikash K Mansinghka, Daniel Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 220–229, 2008.
- [27] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, and Neeraj et al Pradhan. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.
- [28] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International conference on artificial intelligence and statistics*, pages 1682–1690. PMLR, 2018.
- [29] Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. Technical report, Center for Brains, Minds and Machines (CBMM), 2014.
- [30] Bas van Opheusden, Ionatan Kuperwajs, Gianni Galbiati, Zahy Bnaya, and Yunqi et al Li. Expertise increases planning depth in human gameplay. *Nature*, 618(7967):1000–1005, 2023.
- [31] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.
- [32] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, and Tom et al Griffiths. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [33] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.
- [34] Julian Jara-Ettinger, Laura E Schulz, and Joshua B Tenenbaum. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123: 101334, 2020.
- [35] Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2094–2103, 2024.
- [36] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40: e253, 2017. doi: 10.1017/S0140525X16001837.
- [37] Junyi Chu and Laura E Schulz. Play, curiosity, and cognition. *Annual Review of Developmental Psychology*, 2(1):317–343, 2020.
- [38] Itai Yanai and Martin J Lercher. It takes two to think. *Nature Biotechnology*, pages 1–2, 2024.
- [39] Keith J Holyoak and Robert G Morrison. *The Cambridge handbook of thinking and reasoning*. Cambridge University Press, 2005.

- [40] Keith J Holyoak and Robert G Morrison. *The Oxford handbook of thinking and reasoning*. Oxford University Press, 2012.
- [41] Amy J Ko and Brad A Myers. Designing the whyline: a debugging interface for asking questions about program behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 151–158, 2004.
- [42] Amy J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, and Margaret et al Burnett. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)*, 43(3):1–44, 2011.
- [43] Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.
- [44] John R Anderson and Brian J Reiser. The lisp tutor. *Byte*, 10(4):159–175, 1985.
- [45] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.
- [46] Saki Imai. Is github copilot a substitute for human pair-programming? an empirical study. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, pages 319–321, 2022.
- [47] Nhan Nguyen and Sarah Nadi. An empirical evaluation of github copilot’s code suggestions. In *Proceedings of the 19th International Conference on Mining Software Repositories*, pages 1–5, 2022.
- [48] Michel Wermelinger. Using github copilot to solve simple programming problems. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 172–178, 2023.
- [49] Shraddha Barke, Michael B James, and Nadia Polikarpova. Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1):85–111, 2023.
- [50] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, and Michel C et al Desmarais. Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software*, 203:111734, 2023.
- [51] Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, and Dylan et al Hadfield-Menell. Pragmatic-pedagogic value alignment. In *Robotics Research: The 18th International Symposium ISRR*, pages 49–57. Springer, 2020.
- [52] Fabian Ranz, Vera Hummel, and Wilfried Sihm. Capability-based task allocation in human-robot collaboration. *Procedia Manufacturing*, 9:182–189, 2017.
- [53] Jennifer Casper and Robin R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3):367–385, 2003.
- [54] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, and Winson et al Han. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.
- [55] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, and Omar et al Cortes. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. Last Accessed: 2024-07-07.
- [56] Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, and Andrew et al Bolt. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.
- [57] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022.

- [58] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, and Nebojsa et al Jojic. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36, 2024.
- [59] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.
- [60] Theodore R Sumers, Mark K Ho, Thomas L Griffiths, and Robert D Hawkins. Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review*, 2023.
- [61] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.
- [62] Thomas Kollar, Stefanie Tellex, Matthew R Walter, Albert Huang, and Abraham et al Bachrach. Generalized grounding graphs: A probabilistic framework for understanding grounded language. *Journal of Artificial Intelligence Research*, pages 1–35, 2013.
- [63] Michael E Bratman. *Shared agency: A planning theory of acting together*. Oxford University Press, 2013.
- [64] Stephanie Stacy, Chenfei Li, Minglu Zhao, Yiling Yun, and Qingyi et al Zhao. Modeling communication to coordinate perspectives in cooperation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.
- [65] Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, and David C et al Parkes. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.
- [66] Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you’re going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems*, 31, 2018.
- [67] Arwa Alanqary, Gloria Z Lin, Joie Le, Tan Zhi-Xuan, and Vikash K et al Mansinghka. Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*, 2021.
- [68] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.
- [69] Shuwa Miura and Shlomo Zilberstein. A unifying framework for observer-aware planning and its complexity. In *Uncertainty in Artificial Intelligence*, pages 610–620. PMLR, 2021.
- [70] Linda Flower and John R. Hayes. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387, 1981. ISSN 0010096X.
- [71] John R Hayes. Modeling and remodeling writing. *Written communication*, 29(3):369–388, 2012.
- [72] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.
- [73] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, and Vipul et al Raheja. A design space for intelligent and interactive writing assistants. *CHI*, 2024.
- [74] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. Creative writing with an ai-powered writing assistant: Perspectives from professional writers, 2022.
- [75] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019, 2022.
- [76] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. Social dynamics of ai support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

- [77] Fabrizio Dell’Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, and Katherine et al Kellogg. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.
- [78] Justin Porter, Cynthia Boyd, M Reza Skandari, and Neda Laiteerapong. Revisiting the time needed to provide adult primary care. *Journal of general internal medicine*, 38(1): 147–155, 2023.
- [79] Carolyn S Dewa, Desmond Loong, Sarah Bonato, and Lucy Trojanowski. The relationship between physician burnout and quality of healthcare in terms of safety and acceptability: a systematic review. *BMJ open*, 7(6):e015141, 2017.
- [80] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, and Gaurav et al Mishra. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [81] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, and Jason et al Wei. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [82] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
- [83] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, and Jan Freyberg et al. Towards conversational diagnostic ai, 2024.
- [84] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, and Zechariah et al Zhu. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 04 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.1838.
- [85] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24): 18069–18083, 2020.
- [86] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.
- [87] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Irene Y et al Chen. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.
- [88] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology*, 157(11):1362–1369, 2021.
- [89] Federico Cabitza and Jean-David Zeitoun. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of translational medicine*, 7(8), 2019.
- [90] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, and Yuan-Hong et al Liao. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.
- [91] Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Josh Tenenbaum. Inferring the future by imagining the past. *Advances in Neural Information Processing Systems*, 36, 2024.
- [92] Jaime F Fisac, Chang Liu, Jessica B Hamrick, Shankar Sastry, and J Karl et al Hedrick. Generating plans that predict themselves. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, pages 144–159. Springer, 2020.
- [93] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.
- [94] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

- [95] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
- [96] Brian Cantwell Smith. *The promise of artificial intelligence: reckoning and judgment*. MIT Press, 2019.
- [97] Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. *NeurIPS*, 2023.
- [98] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, and Andreea Bobu et al. Getting aligned on representational alignment, 2023.
- [99] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.
- [100] Alessandro Roncone, Olivier Mangin, and Brian Scassellati. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1014–1021. IEEE, 2017.
- [101] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, and Sanjit et al Seshia. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.
- [102] Owen Macindoe, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Pomcop: Belief space planning for sidekicks in cooperative games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 8, pages 38–43, 2012.
- [103] Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. Inferring rewards from language in context. *arXiv preprint arXiv:2204.02515*, 2022.
- [104] Hieke Keuning, Johan Jeuring, and Bastiaan Heeren. A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE)*, 19(1):1–43, 2018.
- [105] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.
- [106] Kartik Chandra, Tzu-Mao Li, Rachit Nigam, Joshua Tenenbaum, and Jonathan Ragan-Kelley. Watchat: Explaining perplexing programs by debugging mental models, 2024.
- [107] Andrew Head, Codanda Appachu, Marti A Hearst, and Björn Hartmann. 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)*, pages 3–12. IEEE, 2015.
- [108] Anna N Rafferty, Rachel A Jansen, and Thomas L Griffiths. Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive science*, 44(10):e12900, 2020.
- [109] Gabriel Poesia and Noah D Goodman. Peano: learning formal mathematical reasoning. *Philosophical Transactions of the Royal Society A*, 381(2251):20220044, 2023.
- [110] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, and Ben et al Bogin. An autonomous debating system. *Nature*, 591(7850):379–384, 2021.
- [111] Daniel Jarrett, Miruna Pislari, Michiel A Bakker, Michael Henry Tessler, and Raphael et al Koster. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.
- [112] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
- [113] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, and Lucy et al Campbell-Gillingham. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

- [114] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2), 2021.
- [115] Alexander Lew, Monica Agrawal, David Sontag, and Vikash Mansinghka. Pclean: Bayesian data cleaning at scale with domain-specific probabilistic programming. In *International Conference on Artificial Intelligence and Statistics*, pages 1927–1935. PMLR, 2021.
- [116] Feras A Saad, Marco F Cusumano-Towner, Ulrich Schaechtle, Martin C Rinard, and Vikash K Mansinghka. Bayesian synthesis of probabilistic programs for automatic data modeling. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–32, 2019.
- [117] Mathieu Huot, Matin Ghavami, Alexander K Lew, Ulrich Schaechtle, and Cameron E et al Freer. Gensql: A probabilistic programming system for querying generative models of database tables. *Proceedings of the ACM on Programming Languages*, 8(PLDI):790–815, 2024.
- [118] Michael Y Li, Emily B Fox, and Noah D Goodman. Automated statistical model discovery with language models. *arXiv preprint arXiv:2402.17879*, 2024.
- [119] Christian Steinruecken, Emma Smith, David Janz, James Lloyd, and Zoubin Ghahramani. The automatic statistician. *Automated machine learning: Methods, systems, challenges*, pages 161–173, 2019.
- [120] Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, and Daniel et al Zheng. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.
- [121] Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, and Kyle et al Cranmer. Discovering symbolic models from deep learning with inductive biases. *Advances in neural information processing systems*, 33:17429–17442, 2020.
- [122] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, and M Pawan et al Kumar. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.
- [123] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How ai ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment, 2024.
- [124] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, and Mengting et al Wan. The use of generative search engines for knowledge work and complex tasks. Technical Report MSR-TR-2024-9, Microsoft, March 2024.
- [125] Henriikka Vartiainen and Matti Tedre. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity*, 34(1):1–21, 2023.
- [126] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, and Devi et al Parikh. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.
- [127] Judith E Fan, Monica Dinculescu, and David Ha. Collabdraw: an environment for collaborative sketching with an artificial agent. In *Proceedings of the 2019 Conference on Creativity and Cognition*, pages 556–561, 2019.
- [128] Songwei Ge, Vedanuj Goswami, C Lawrence Zitnick, and Devi Parikh. Creative sketch generation. *arXiv preprint arXiv:2011.10039*, 2020.
- [129] Marek Dvorožňák, Daniel Šykora, Cassidy Curtis, Brian Curless, and Olga et al Sorkine-Hornung. Monster mash: a single-view approach to casual 3d modeling and animation. *ACM Transactions on Graphics (ToG)*, 39(6):1–12, 2020.
- [130] Nick Chater and Mike Oaksford, editors. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press, Oxford, UK, 2008.
- [131] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge handbook of computational psychology*, chapter 3, pages 59–100. Cambridge University Press, 2008.

- [132] Elizabeth S Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000.
- [133] Steven T Piantadosi. The computational origin of representation. *Minds and machines*, 31(1):1–58, 2021.
- [134] Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261, 2023.
- [135] Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16, 2003.
- [136] Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.
- [137] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- [138] Joshua S Rule, Joshua B Tenenbaum, and Steven T Piantadosi. The child as hacker. *Trends in cognitive sciences*, 24(11):900–915, 2020.
- [139] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, and Lucas et al Morales. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pages 835–850, 2021.
- [140] Tomer D Ullman and Joshua B Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558, 2020.
- [141] Falk Lieder, Owen X Chen, Paul M Krueger, and Thomas L Griffiths. Cognitive prostheses for goal achievement. *Nature human behaviour*, 3(10):1096–1106, 2019.
- [142] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.
- [143] Thomas Icard. Resource rationality. Book manuscript, 2023.
- [144] Allen Newell and Herbert A. Simon. *Human problem solving*. Prentice-Hall, 1972.
- [145] Marcelo G Mattar and Máté Lengyel. Planning in the brain. *Neuron*, 110(6):914–934, 2022.
- [146] Mark K. Ho, David Abel, Carlos G. Correa, Michael L. Littman, and Jonathan D. et al Cohen. People construct simplified mental representations to plan. *Nature*, 606(7912): 129–136, 2022.
- [147] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016.
- [148] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.
- [149] Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971, 2022.
- [150] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.
- [151] Judith Degen. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540, 2023.
- [152] Marcel Binz, Ishita Dasgupta, Akshay K Jagadish, Matthew Botvinick, and Jane X et al Wang. Meta-learned models of cognition. *Behavioral and Brain Sciences*, pages 1–38, 2023.
- [153] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

- [154] Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.
- [155] Mark K Ho and Thomas L Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:33–53, 2022.
- [156] Scott Cheng-Hsin Yang, Tomas Folke, and Patrick Shafto. The inner loop of collective human-machine intelligence. *Topics in cognitive science*, 2023.
- [157] Mark Steyvers and Aakriti Kumar. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, page 17456916231181102, 2023.
- [158] Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.
- [159] Mike Oaksford and Nick Chater. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007.
- [160] Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.
- [161] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004.
- [162] Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science advances*, 6(10):eaax5979, 2020.
- [163] Kelsey R. Allen, Kevin A. Smith, and Joshua B. Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020.
- [164] Cedegao E Zhang, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.
- [165] Joshua Tenenbaum. Bayesian modeling of human concept learning. *Advances in neural information processing systems*, 11, 1998.
- [166] Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154, 2008.
- [167] Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392, 2016.
- [168] Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. Integrating topics and syntax. *Advances in neural information processing systems*, 17, 2004.
- [169] Noah D Goodman and Daniel Lassiter. Probabilistic semantics and pragmatics uncertainty in language and thought. *The handbook of contemporary semantic theory*, pages 655–686, 2015.
- [170] Yuan Yang and Steven T Piantadosi. One model for the learning of language. *Proceedings of the National Academy of Sciences*, 119(5):e2021865119, 2022.
- [171] Laura E Schulz, Elizabeth Baraff Bonawitz, and Thomas L Griffiths. Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers’ causal inferences. *Developmental psychology*, 43(5):1124, 2007.
- [172] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, and Tamar et al Kushnir. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.
- [173] Lara Kirfel, Thomas Icard, and Tobias Gerstenberg. Inference from explanation. *Journal of Experimental Psychology: General*, 151(7):1481, 2022.
- [174] David A Lagnado, Tobias Gerstenberg, and Ro’i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.
- [175] Pernille Hemmer and Mark Steyvers. A bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1):189–202, 2009.

- [176] Tomer D. Ullman and Joshua B. Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558, 2020. doi: 10.1146/annurev-devpsych-121318-084833.
- [177] Thomas L Griffiths and Joshua B Tenenbaum. Theory-based causal induction. *Psychological review*, 116(4):661, 2009.
- [178] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637, 2014.
- [179] Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. *arXiv preprint arXiv:1408.2048*, 2014.
- [180] Momchil S. Tomov, Samyukta Yagati, Agni Kumar, Wanqian Yang, and Samuel J. Gershman. Discovery of hierarchical representations for efficient planning. *PLOS Computational Biology*, 16(4):1–42, 04 2020. doi: 10.1371/journal.pcbi.1007594.
- [181] Chris L Baker and Joshua B Tenenbaum. Modeling human plan recognition using bayesian theory of mind. *Plan, activity, and intent recognition: Theory and practice*, 7:177–204, 2014.
- [182] Frederick Callaway, Bas van Opheusden, Sayan Gul, Priyam Das, and Paul M et al Krueger. Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 6(8):1112–1125, 2022.
- [183] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.
- [184] Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33:19238–19250, 2020.
- [185] Lance Ying, Katherine M Collins, Megan Wei, Cedegao E Zhang, and Tan et al Zhi-Xuan. The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*, 2023.
- [186] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.
- [187] Ruth MJ Byrne. Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences*, 6(10):426–431, 2002.
- [188] Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89, 2014.
- [189] Theodore R Summers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010, 2021.
- [190] Emily G Liquin, Nicole Luzuriaga, and Todd M Gureckis. Teaching and learning through pedagogical environment design. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.
- [191] Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Differentiating mental models of self et al: A hierarchical framework for knowledge assessment. *Psychological Review*, 2023.
- [192] Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, and Kenny et al Smith. From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4):977, 2023.
- [193] Robert D Hawkins, Andrew M Berdahl, Alex ‘Sandy’ Pentland, Joshua B Tenenbaum, and Noah D et al Goodman. Flexible social inference facilitates targeted social learning when rewards are not observable. *Nature Human Behaviour*, pages 1–10, 2023.
- [194] Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

- [195] Mark K Ho, Fiery Cushman, Michael L Littman, and Joseph L Austerweil. Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, 150(11):2246, 2021.
- [196] Thomas L Griffiths. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883, 2020.
- [197] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.
- [198] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.
- [199] Jian-Qiao Zhu, Joakim Sundh, Jake Spicer, Nick Chater, and Adam N Sanborn. The autocorrelated bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*, 2023.
- [200] Iris Van Rooij. The tractable cognition thesis. *Cognitive science*, 32(6):939–984, 2008.
- [201] Thomas Icard and Noah D Goodman. A resource-rational approach to the causal frame problem. In *CogSci*, 2015.
- [202] John R Anderson. *The adaptive character of thought*. Psychology Press, 1990.
- [203] Samuel J Cheyette, Frederick Callaway, Neil R Bramley, Jonathan D Nelson, and Josh Tenenbaum. People seek easily interpretable information. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.
- [204] Feras Ahmad Khaled Saad. *Scalable Structure Learning, Inference, and Analysis with Probabilistic Programs*. PhD thesis, Massachusetts Institute of Technology, 2022.
- [205] Alexander K Lew, Michael Henry Tessler, Vikash K Mansinghka, and Joshua B Tenenbaum. Leveraging unstructured statistical knowledge in a probabilistic language of thought. In *Proceedings of the annual conference of the cognitive science society*, 2020.
- [206] Nishad Gothoskar, Matin Ghavami, Eric Li, Aidan Curtis, and Michael et al Noseworthy. Bayes3d: fast learning and inference in structured generative models of 3d objects and scenes. *arXiv preprint arXiv:2312.08715*, 2023.
- [207] Vikash K Mansinghka, Ulrich Schaechtle, Shivam Handa, Alexey Radul, Yutian Chen, and Martin Rinard. Probabilistic programming with programmable inference. In *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 603–616, 2018.
- [208] Alexander K. Lew, Mathieu Huot, Sam Staton, and Vikash K. Mansinghka. Adev: Sound automatic differentiation of expected values of probabilistic programs. *Proc. ACM Program. Lang.*, 7(POPL), jan 2023. doi: 10.1145/3571198. URL <https://doi.org/10.1145/3571198>.
- [209] McCoy R Becker, Alexander K Lew, Xiaoyan Wang, Matin Ghavami, Mathieu Huot, Martin C Rinard, and Vikash K Mansinghka. Probabilistic programming with programmable variational inference. *Proceedings of the ACM on Programming Languages*, 8(PLDI):2123–2147, 2024.
- [210] Feras A Saad, Martin C Rinard, and Vikash K Mansinghka. Sppl: probabilistic programming with fast exact symbolic inference. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pages 804–819, 2021.
- [211] Alexander K Lew, Matin Ghavamizadeh, Martin C Rinard, and Vikash K Mansinghka. Probabilistic programming with stochastic probabilities. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1708–1732, 2023.
- [212] Tobias Moritz Guggenberger, Frederik Möller, Tim Haarhaus, Inan Gür, and Boris Otto. Ecosystem types in information systems. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, 2020.

- [213] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3):50–57, 2017.
- [214] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494, 2019.
- [215] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration, 2023.
- [216] Don Norman. *Design Of Everyday Things*. New York: Basic Books. Olins, W.(2005). A Marca. Lisboa: Verbo. Packard, V . . . , 1988.
- [217] Anthony Chemero. An outline of a theory of affordances. In *How Shall Affordances Be Refined?*, pages 181–195. Routledge, 2018.
- [218] John Zerilli, Umang Bhatt, and Adrian Weller. How transparency modulates trust in artificial intelligence. *Patterns*, page 100455, 2022.
- [219] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.
- [220] Heliodoro Tejada, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 5(4):1–18, 2022.
- [221] Mark Steyvers, Heliodoro Tejada, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.
- [222] Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Josh Tenenbaum. Cooperative explanation as rational communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.
- [223] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- [224] Kartik Chandra, Tzu-Mao Li, Joshua Tenenbaum, and Jonathan Ragan-Kelley. Acting as inverse inverse planning. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023.
- [225] Tony Chen, Sean Dae Houlihan, Kartik Chandra, Josh Tenenbaum, and Rebecca Saxe. Intervening on emotions by planning over a theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.
- [226] Kartik Chandra, Tzu-Mao Li, Joshua B Tenenbaum, and Jonathan Ragan-Kelley. Storytelling as inverse inverse planning. *Topics in Cognitive Science*, 16(1):54–70, 2024.
- [227] João Loula, Katherine M Collins, Ulrich Schaechtle, Joshua B Tenenbaum, and Adrian et al Weller. Learning generative population models from multiple clinical datasets via probabilistic programming. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.
- [228] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.
- [229] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.
- [230] Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, and Robert et al Stanforth. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, pages 1–7, 2023.
- [231] Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. Human-machine social systems. *arXiv preprint arXiv:2402.14410*, 2024.
- [232] Eike Schneiders, EunJeong Cheon, Jesper Kjeldskov, Matthias Rehm, and Mikael B Skov. Non-dyadic interaction: A literature review of 15 years of human-robot interaction conference publications. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2): 1–32, 2022.

- [233] Eva Hornecker, Antonia Krummheuer, Andreas Bischof, and Matthias Rehm. Beyond dyadic hri: Building robots for society. *interactions*, 29(3):48–53, 2022.
- [234] Aakash Yadav and Ranjana Mehta. Beyond dyadic interactions: Assessing trust networks in multi-human-robot teams. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1153–1157, 2024.
- [235] Iliia Sucholutsky, Katherine M Collins, Maya Malaviya, Nori Jacoby, and Weiyang et al Liu. Representational alignment supports effective machine teaching. *arXiv Preprint arXiv:2406.04302*, 2024.
- [236] Ling Li, Xiaojian Li, Bo Ouyang, Hangjie Mo, and Hongliang et al Ren. Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery. *Cyborg and Bionic Systems*, 4:0042, 2023.
- [237] Veronica Boyce, Robert D Hawkins, Noah D Goodman, and Michael C Frank. Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, 121(28), 2024.
- [238] Laura Trouille, Chris J Lintott, and Lucy F Fortson. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences*, 116(6):1902–1909, 2019.
- [239] Kasper Hornbæk and Antti Oulasvirta. What is interaction? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5040–5052, 2017.
- [240] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, and Esin et al Durmus. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023.
- [241] Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E Fan, and Samuel J et al Gershman. Using games to understand the mind. *Nature Human Behaviour*, pages 1–9, 2024.
- [242] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, and Percy Liang et al. Generative agents: Interactive simulacra of human behavior, 2023.
- [243] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.
- [244] Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, and Colin et al Flaherty. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.
- [245] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.
- [246] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.
- [247] Isa Inuwa-Dutse, Alice Toniolo, Adrian Weller, and Umang Bhatt. Algorithmic loafing and mitigation strategies in human-ai teams. *Computers in Human Behavior: Artificial Humans*, 1(2):100024, 2023.
- [248] Jake M Hofman, Daniel G Goldstein, and David M Rothschild. Steroids, sneakers, coach: The spectrum of human-ai relationships. *Available at SSRN 4578180*, 2023.
- [249] Daniel Buschek, Martin Zürn, and Malin Eiband. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445372.
- [250] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

- [251] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
- [252] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170, 2018.
- [253] John Zerilli, Umang Bhatt, and Adrian Weller. Transparency Modulates Trust in Artificial Intelligence. *Patterns*, 2022.
- [254] Lewis Mumford. *Technics and civilization*. 1936.
- [255] Joseph Weizenbaum. Computer power and human reason: From judgment to calculation. 1976.
- [256] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, and Po-Sen et al Huang. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.
- [257] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.
- [258] Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.
- [259] Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.
- [260] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, and John et al Schulman. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [261] Stuart Russell. *Human compatible: AI and the problem of control*. Viking, 2019.
- [262] Stuart Russell. Artificial intelligence and the problem of control. *Perspectives on Digital Humanism*, pages 19–24, 2021.
- [263] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.
- [264] Seth Lazar and Alondra Nelson. Ai safety on whose terms? *Science*, 381(6654):138–138, 2023.