Meaningful Long-Term Thought Partnerships of Minds and Machines

Katherine M. Collins*[1,2], Lionel Wong*[3], Joshua B. Tenenbaum[2], and Judith E. Fan[3]

[1]University of Cambridge
[2]Massachusetts Institute of Technology
[3]Stanford University

Abstract

Many innovations have come from people working together as partners in thought. These partnerships, however, are not restricted to single encounters. Some of the most meaningful collaborations evolve over weeks, months, or even lifetimes. What are the core computations that enable long-term thought partnerships? Prior work in cognitive science has made initial progress by investigating how people construct mental models of their partners on-the-fly, establish common ground using language and other modalities, and generate joint plans that lead to successful outcomes. However, it remains unknown what cognitive mechanisms enable such interactions to evolve into genuine partnerships over longer timescales, especially under measures of success that extend beyond task performance. Theoretical and empirical progress on these issues could be instrumental for defining and designing AI systems that may even be capable of establishing long-term thought partnerships with humans. This paper lays out several promising avenues for leveraging approaches from cognitive science and AI to study enriching intellectual partnerships.

Introduction

We all have a choice: how to make use of our limited lifetime. Some may choose to spend their days caring for others, helping those who are sick; others, may choose to pursue the adventures of science and navigating into the unknown, or building something new. Many of these pursuits are not done in isolation. We regularly engage with others in the quest for our dreams, be it asking for advice on a business decision, consulting another doctor on a tricky case, or co-designing the next experiment with a fellow scientist. Our "thought partnering" (Collins et al., 2024; Yanai and Lercher, 2024) is not limited to a single instance, or even a few interactions. If we are lucky, we find and develop deep, long-term collaborations. Some of the best innovations in science and the arts have come from multi-year thought partnerships, from Amos Tversky and Daniel Khaneman, to Esther Duflo and Abhijit Banerjee, and Steven Spielberg and John Williams.

Cognitive science has made substantial progress in how we think about how people collaborate with each other. This article reflects on this progress and asks how we may move beyond models of short-term collaboration towards modeling long-term meaningful thought partnership. We start by discussing three bodies of work: modeling each other, modeling joint action, and modeling communication (Figure 1). We discuss challenges in capturing the core computations that empower collaboration over longer time horizons. We then turn to how — and whether — ideas from human-human partnerships can extend to possible meaningful long-term human-AI thought partnerships.

*Contributed equally.

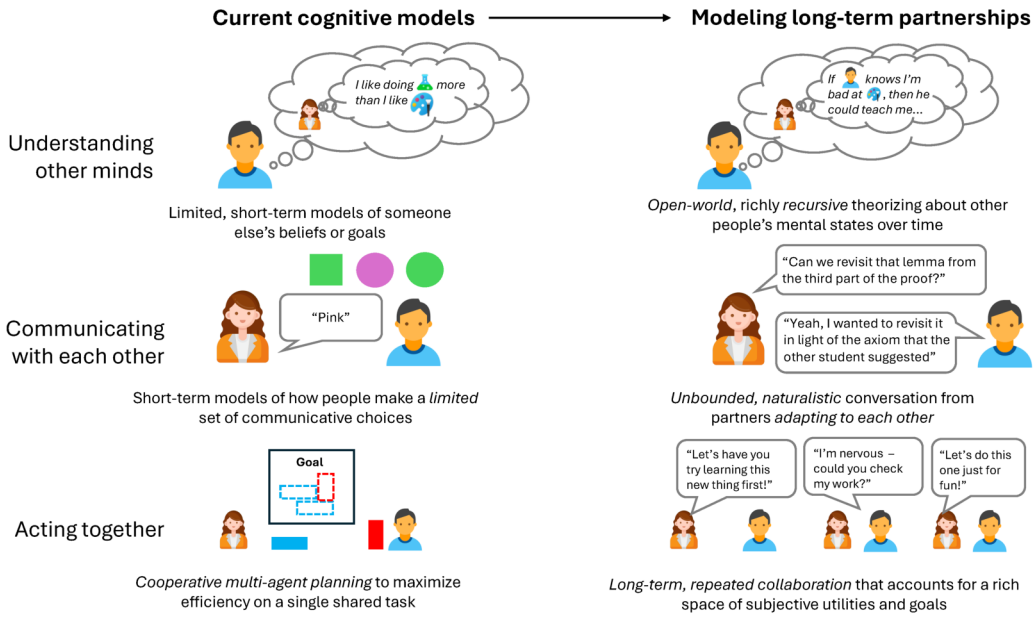

Figure 1: Three foundational lines of work in cognitive science for modeling human collaboration, and key open questions raised by our *current cognitive models* (*left*) as we seek to capture *long-term partnership* (*right*). These include questions about how we scale today's cognitive models of how people **understand other minds** (*top*) by reasoning about their mental states; models of how people **communicate with each other** (*middle*) by choosing language to convey information effectively to a partner; and models of how people **act together** (*bottom*) to jointly achieve towards shared goals.

Laying the theoretical foundations of partnership: understanding, communicating, and acting together

We begin by reviewing three foundational lines of work that underlie our discussion of partnership: formal theories of how people **understand other minds**, how they **communicate with each other**, and ultimately, what it means for them to **act together** to achieve a shared goal. However, each raises several significant *challenges* for scaling these theories towards rich models of *long-term partnership*. We highlight these open questions here.

Understanding other minds

When people enter into a collaboration, they often know that each person has their own rich inner mind, with their own individual beliefs, goals, and feelings. Potential collaborators can understand a lot about each other, just from watching how a partner acts. Computational models of *theory-of-mind* formalize this knowledge of how other peoples' mental states shape their actions (Baker et al., 2009; Jara-Ettinger et al., 2020). Models like Baker et al. (2009) formalize a simple set of intuitions that most people have about others: people expect that others are not merely acting at random; people's actions are likely based on their goals and own beliefs about what is happening.

In practice, however, contemporary theory-of-mind models used in cognitive science face several fundamental challenges that come up especially in the case of long-term collaboration. First, the ways that people reason about each other over the course of an extended collaboration are often nested. A mentor might want to help their junior colleague in the long run, while also mitigating any embarrassment their colleague might feel for needing a hand. This kind of reasoning is often highly uncertain and *recursive*, in that it involves thinking about how another person might think about oneself, several layers deep. More generally, any extended

collaboration involves navigating a changing landscape of new considerations that come up over time, often unexpectedly. Existing computational models do not easily explain how people can reason about an enormous and changing space of possible factors under the computational time and memory than human minds have at their disposal (Griffiths, 2020). Understanding how to scale formal models of theory-of-mind towards longer time scales and more realistic settings where anything might come up at any time is an open question.

Communicating with each other

But human collaborators do not just try to read each other's minds by watching each other's actions, as in Baker et al. (2011). Instead, they can communicate with each other. Successful collaboration relies on some shared knowledge. Potential collaborators may not start knowing the same things nor even agree on their shared goals. Communication lets collaborators close those gaps (Clark and Brennan, 1991). A collaborator can tell their partner key details necessary to work effectively on a task (e.g., how to use a tool) and information about themselves (what they prefer to work on; what they need more experience in order to learn). Rational models of communication, like the Rational Speech Acts framework Frank and Goodman (2012) frame communication as a thoughtful, intentional action that requires reasoning about what a partner may already believe, and how they are likely to interpret (or misinterpret) communicative choices.

However, rational models of communication pose several computational challenges to explain real communication between long-term intellectual partners. First, many rational models of communication are idealized models that cannot be directly applied to explain how people reason about unbounded, arbitrary natural language in general. Most cognitive models built on this framework operate over a predefined set of possible communicative actions, like a set of phrases and their interpretations. Second, these models do not generally explain the dramatic ways in which collaborators can change their communicative channels over long periods of time. Real close collaborators often develop specialized vocabularies to capitalize on their shared knowledge and talk more precisely about key concepts in their domain (Hawkins et al., 2023). Modeling long-term collaboration requires scaling rational models of communication to account for the richness of real natural language, and the ways partners change that language to suit their needs.

Acting together

Ultimately, what makes something a collaboration is the act of working jointly towards a shared goal, wherein partners act together in ways that complement and enrich each other, yielding something greater than the mere sum of two individuals working on their own. Theories of *cooperative multi-agent action and planning* formalize how several people can work together to efficiently achieve a shared goal, often under uncertainty (Grosz and Kraus, 1996). Many of these theories build on notions of individual rational planning and goal-directed action, which describe how any one agent should plan their actions in order to optimally achieve a goal. The Bayesian Delegation model in Wu et al. (2021), for instance, uses this general idea to formalize a basic, but fundamental, aspect of cooperation. It describes how collaborators should reason about what small subtask their partners are currently on track to achieve, so that the team can avoid redundant effort or accidentally slowing each other down as they work (Mieczkowski et al., 2025).

As models of true long-term collaboration, however, these computational approaches raise several open questions. Long-term planning of any kind poses significant challenges for current cognitive theories. Much like theory-of-mind, existing theories of rational action often suggest that planning and acting requires much more time and computational effort than people have at their disposal. Moreover, existing theories of multi-agent action usually focus narrowly on how

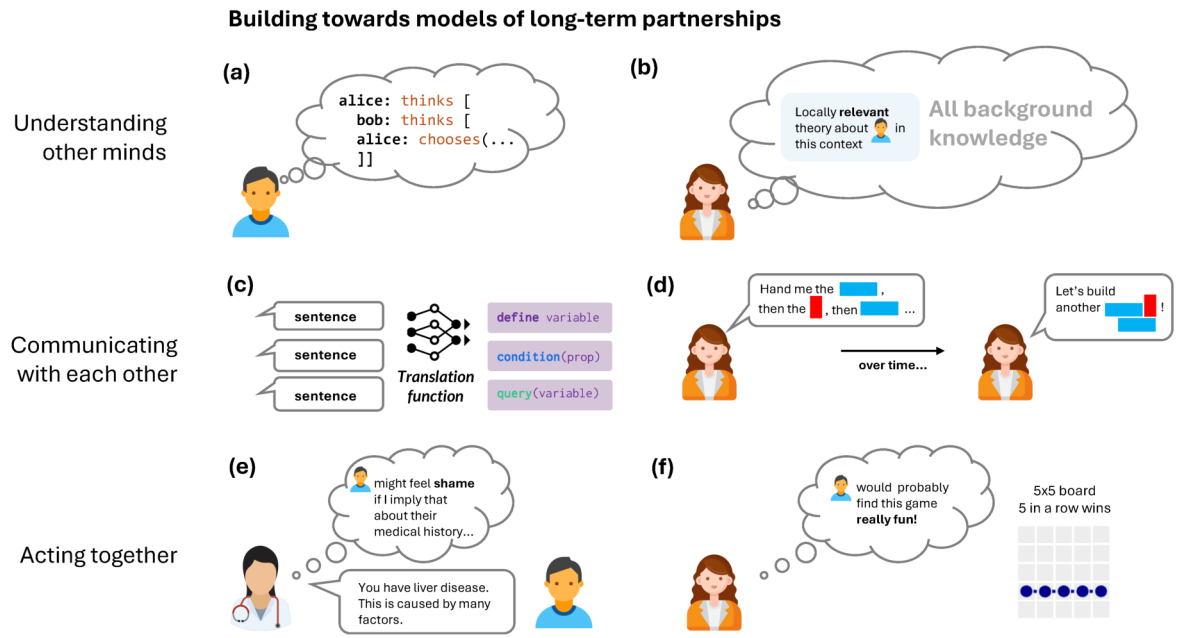

Figure 2: Examples of recent work that addresses key questions in modeling *long-term partnership*. To build towards long-term models of **understanding other minds** (*top*), **(a)** the `memo` programming language Chandra et al. (2025) formalizes richly *recursive* models of mental states, and **(b)** recent work Wong et al. (2025); Zhang et al. (2025) formalizes how people might propose *small, relevant* theories of someone else's mental states in a given context. To build towards long-term models of **communicating with others**, **(c)** recent work Loula et al. (2025); Wong et al. (2023) formalizes how people might turn *naturalistic language* into structured mental representations of conceptual content; and **(d)** other work Ellis et al. (2021); O'Donnell (2015) formalizes how people might *adapt their vocabulary* to talk about more complex, shared concepts over time. To build towards long-term models of **acting together**, **(e)** recent work Chen et al. (2024); Collins et al. (2025a); Houlihan et al. (2023) models how people might act to take into account someone else's *emotions*; or **(f)** reason about subjective and experiential sources of reward, like what someone else would find *fun* Chu et al. (2025); Collins et al. (2025b).

two people might achieve any one goal, as quickly as possible. Real collaborators take a much richer and wider set of considerations into account, both for themselves and for their partners: people who choose to collaborate over the long-term might continue working together because the partnership feels fair; they can always learn something new and surprising from the other person; they find it rewarding to teach and mentor another person; or, because the work itself and the act of working with their partner brings them meaning and joy. Real people — when collaborating well — strive to act towards each other with dignity, and empathy, and mutual respect. All of these go well beyond a simple notion of cooperative action, and yet lie at the very heart of what it means to collaborate well.

Building towards theories of *meaningful, long-term* partnership

We next lay out a few recent works which have started to make progress towards understanding, formalizing, and empirically studying computations potentially underlying the cognitive mechanisms driving rich human-human long-term thought partnerships (Figure 2).

Modeling how people understand each other in real-world, long-term partnerships

How do collaborators make sense of the complexities of any real person, juggling what their partner might want from the task at hand, alongside what that other person might be thinking

about them? And how do they even begin to know what the other person is thinking — even imperfectly — in the messy dynamics of a real partnership? Modeling any real long-term partnership will require scaling today's limited theory-of-mind models to handle richly recursive reasoning about other minds, and reasoning about sets of actions and mental states that someone cannot know in advance. These two computational problems might be linked to explain how we feasibly reason about and understand each other without an inordinate amount of computational effort. For instance, modeling how collaborators might consider richly recursive theories about what another collaborator has in mind, but assume that what their partner is thinking is probably limited by their own available memory and effort (Griffiths et al., 2015; Zhi-Xuan et al., 2020). Or more broadly, explaining real long-term human collaboration might require building on recent work suggesting that people reason in general, including about each other, by constantly coming up with and revising *small* but sensible theories that capture just enough detail to understand the situation at hand Brooke-Wilson (2023); Wong et al. (2025); Ying et al. (2025); Zhang et al. (2025). These small, approximate theories of another person, built just to capture what seems relevant and useful to make sense of what they are doing, might explain how people can ever understand and reason about each other in an ever-changing collaboration; all of which may be advanced by new computational frameworks for modeling social agents (Chandra et al., 2025).

Modeling communication over real language and changing language

How do long-term collaborators reason about what to say, out of the enormous space of things they could say? And how do they change the language they use over time to suit their needs, inventing the kinds of rich and specialized vocabularies that are used in real academic work and in extended collaborations? Two exciting research directions for modeling communication between long-term partners involve explaining how rational communication can integrate with general language models, and redefining rational communication to account for how people invent new concepts and change their vocabularies accordingly. In the first direction, recent work that integrates structured rational models for inference and planning with large statistical language models Lew et al. (2020); Loula et al. (2025); Wong et al. (2023) suggests one starting point for integrating rational communicative frameworks like those Frank and Goodman (2012). These approaches suggest complimentary directions for bringing together recent advances in statistical language modeling with existing cognitive models of human reasoning. The Rational Meaning Construction framework in Wong et al. (2023) proposes that people make sense of language in general because they can translate generally from natural language into structured mental "programming languages" for representing and reasoning about the world. The language model probabilistic programming framework in Lew et al. (2020) and Loula et al. (2025) suggests that people might also have ways for defining general inference algorithms over small, composable primitive distributions learned from linguistic experience. Neither of these frameworks yet incorporates the rich mentalizing about other minds that characterizes real human communication. Future work could build on each approach towards rational communicative models that explain how agents reason about each other, but that can operate over the domain of general natural language. In the second direction, a growing body of work which suggest that people might invent new conceptual abstractions that help them reason and plan, much as software engineers construct growing libraries of abstractions defined over the primitives of simpler programming languages Ellis et al. (2021); O'Donnell (2015). These frameworks propose how people might use and change the language they speak over time, by coining new words that let them efficiently talk about these new conceptual abstractions. Most work in this vein has been largely theoretical, or explored in limited experimental settings; future work could build on these approaches to explain how people fundamentally change the communicative channels at their disposal to suit a deepening intellectual partnership.

Modeling meaningful collaboration

Why do people choose to collaborate with each other at all? How do people try to work towards all of the things that make any real collaboration meaningful, or even begin to assess whether a collaboration has been valuable beyond the mere fact of whether any one job got done? These are deep questions beyond the reach of existing computational models — and perhaps often beyond the reach of people trying to get any would-be collaboration off of the ground. Still, several intersecting lines of recent work suggest how we might begin to build towards these kinds of rich, subjective evaluative criterion within formal computational theories. One recent line of work takes steps towards formalizing how people understand each others' *emotions* Houlihan et al. (2023). This work points out that even emotions as nuanced as pride, envy, joy, and respect are not completely arbitrary nor ineffable can be formally modeled in theory-of-mind frameworks. Other recent work also takes steps to formalize how people evaluate what it felt like to *experience* working with another person, beyond any obvious external reward. (Chu et al., 2025; Collins et al., 2025b), for instance, ask how people evaluate whether playing a game with someone else will actually be fun — separate from whether they think they will win or lose. They propose that these subjective evaluations of experience are fundamentally rooted in the same underlying cognitive capacities that let us plan how to act in general, or reason about other people's mental states and actions, and point to aspects of what makes an experience rewarding — whether it feels joyful — beyond simply achieving a task as efficiently as possible. Other recent works look at how people might plan around the kinds of subjective and experiential goals that we mention here — modeling, for instance, how people might adjust their actions to be empathetic as well as efficient (Collins et al., 2025a), or how people might specifically act to intervene on someone's emotions in order to save them from embarrassment or cheer them up (Chen et al., 2024). Each of these models only carves off a handful of subjective considerations, like empathy or embarrassment, and describes how people might plan around them with respect to a fairly small set of possible actions and outcomes. Real partnerships involve many of these considerations at once. Understanding how people come to understand, in a general sense, what feels rewarding to them and others — and how that shapes their plans and actions in the long-term — requires answering many of the formal questions we raise throughout this section, and breaking new ground.

Studying long-term partnership with new tasks and datasets

A cognitive science of long-term collaboration requires not just model developments, but new experimental methods and evaluative measures, in which we can actually probe the kinds of theoretical and computational questions we raise here. Work like (McCarthy et al., 2025), in which two people cooperate on a realistic computer-aided design task, or Vélez et al. (2024), in which groups of people develop and pass on miniature technological innovations on a sped-up timescale, suggests the kinds of tasks we might use to study complex collaboration even in a controlled laboratory setting. Future work should also collect richer and more subjective evaluative measure within any of these tasks, like whether people expect or found the act of working together to be respectful, dignified, or fair.

Imagining meaningful long-term human-AI thought partnerships

This article focuses on three core computations that may underlie long-term human thought partnerships — the ability to grow and change one's understanding of other agents; to flexibly adapt communication between agents; and to form and evolve shared goals that incorporate subjective utilities, like having fun. To what extent do these concepts extend, or break down, when applied to understand asymmetric human-AI interaction? Despite immense progress in

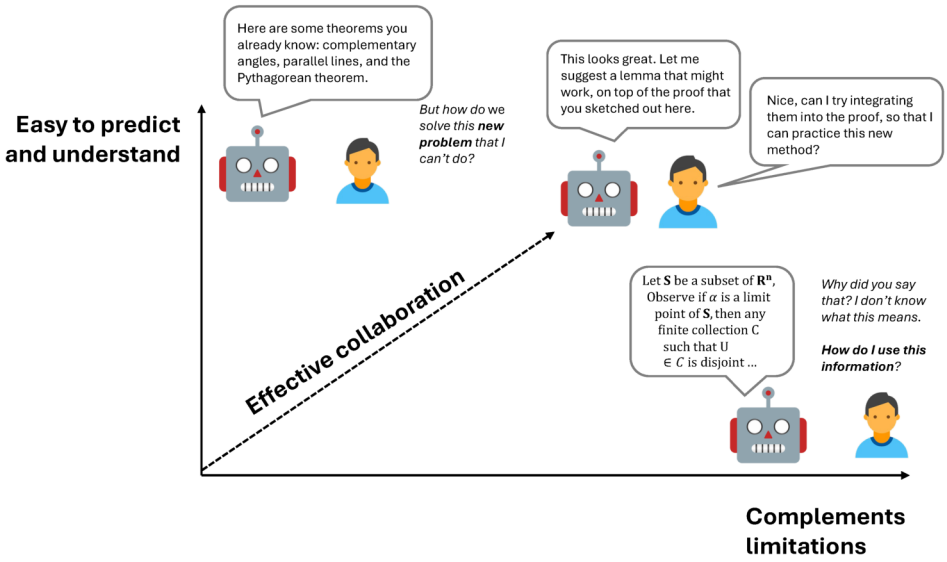

Figure 3: Systems that **meet our expectations** (but do not **complement our limitations**) may be so simple or human-like that we do not get comparative benefit from engaging with AI. Conversely, systems that are optimized exclusively to complement people's limitations may end up being so complex that people struggle to predict when and how they should be used. Human-centric AI systems may require a middle-ground, ideally move towards the upper right quadrant, potentially requiring trading off one or the other axis. Yet, it is an open question whether balancing these two axes alone may not be enough for a system to warrant the label thought *"partner"* (Collins et al., 2024).

AI, recent advances have often been decoupled from collaborative efficiency (Challapally et al., 2025); the most performant AI systems may not be the most effective collaborators (Bansal et al., 2021; Carroll et al., 2019). Indeed, many current systems sit at one of two unsatisfying ends of a collaborative section (Figure 3). An effective human-centric AI system may be one that both meets our expectations yet also is sufficiently unpredictable — as many good human thought partnerships are. While good collaborators often construct a common lexicon (as discussed above), collaborators also teach and push one another, bringing to bear distinct knowledge bases. Herein, AI systems offer tremendous potential as new kinds of collaborators that can shape new strategies for thinking, like how AI gameplay is influencing human-human gameplay strategies (Shin et al., 2023).

However, even if we build AI systems that meet our expectations and complement our limitations — does that mean that they warrant being called a thought *partner* (Collins et al., 2024)? Or, is there a tacit assumption of something else underlying the collaborations for it to be worthy of the label "partnership"? Perhaps that the collaboration itself is not only an effective means to an end, but a relationship that grows in value as it develops precisely because of the assumed *personhood* of everyone involved? Something like personhood and the belief that both partners have a valuable and individual agency of their own, and yet have chosen nonetheless to each *renew* the terms of the partnership over time, might well be fundamental. It may be this assumption of agency, in the fullest sense, that leads people to dramatically reshape their communication to reflect both parties or bend the collaboration to the needs and individual desires of each other. Perhaps human partners do not seek to understand each other's minds' only to better achieve their shared goals, but because they also simply believe each partner to have a mind worthy of being understood. Cognitive scientists are poised to help define, extend, and reshape the foundations of durable long-term thought partnerships and inform how we design the next-generation of human-facing AI systems. Theories about such partnerships (human-human and potentially, human-AI) may in turn, offer mirrors for what we truly expect and seek out in human partnerships with one another.

Conclusion

Some of the richest human experiences take place with the meeting of two minds. Lasting collaborations define many of the relationships that color our individual experience, and the partnerships whose intellectual contributions shape modern society. In this paper, we review several computational foundations that already exist to study long-term partnerships within cognitive science and outline the horizons ahead if we hope to scale towards the complexity and timescales of real, extended collaborations. In doing so, we also hope to lay the groundwork for imagining what kind of AI systems may be worth the name "thought partner" — systems that can be durable and meaningful enough that we would want to invite them into our lives over longer horizons: collaborations that look beyond myopic productivity on short-term tasks and towards truly complex partnerships that are fulfilling, dignified, and even fun.

Recommended Reading

- (Collins et al., 2024) (see References) explores the idea of human-compatible AI thought partners that engages deeply with cognitive science.
- (Yanai and Lercher, 2024) (see References) discusses the value of specifically dyadic (two person) collaborations in science.
- (Hawkins et al., 2023) (see References) studying computations underlying cooperative language change between partners over time.
- (McCarthy et al., 2025) (see References) develops new tasks for studying collaboration, in the context of computer-aided design.

Acknowledgments

We thank Kartik Chandra for many valuable comments on this manuscript. We also thank Ilia Sucholutsky, Tom Griffiths, Umang Bhatt, Adrian Weller, Valerie Chen, Lance Ying, Tan Zhi Xuan, Kerem Oktar, Alex Lew, Tyler Brooke-Wilson, Kaya Stechly, Elizabeth Mieczkowski, Junior Okoroafor, Max Kleiman-Weiner, Neil Lawrence, Eric Horvitz, and Ced Zhang for many conversations around thought partnerships that informed this work. Figures use icons from Icons8. We thank Rob Goldstone and our Reviewers for invaluable comments and questions that fundamentally shaped our thinking around human-AI thought partners.

References

Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3):329–349.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414.

Brooke-Wilson, T. (2023). Bounded rationality as a strategy for cognitive science. *Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA. Retrieved from https://philosophy. mit. edu/wp-content/uploads/brookewilson dissertation. pdf*.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., and Seshia, S. e. a. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Challapally, A., Pease, C., Raskar, R., and Chari, P. (2025). The GenAI divide: State of AI in business 2025. Project report, MIT NANDA.

Chandra, K., Chen, T., Tenenbaum, J. B., and Ragan-Kelley, J. (2025). A domain-specific probabilistic programming language for reasoning about reasoning (or: a memo on memo). *psyarxiv preprint*.

Chen, T., Houlihan, S. D., Chandra, K., Tenenbaum, J., and Saxe, R. (2024). Intervening on emotions by planning over a theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Chu, J., Zheng, K., and Fan, J. E. (2025). What makes people think a puzzle is fun to solve? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Collins, K. M., Chandra, K., Weller, A., and Reagan-Kelley, Jonathan Tenenbaum, J. (2025a). Emotions in explanations. *CogSci*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature Human Behavior*.

Collins, K. M., Zhang, C. E., Wong, L., Barba, M., Todd, G., Weller, A., Cheyette, S., Griffiths, T. L., and Tenenbaum, J. B. (2025b). People use fast, flat goal-directed simulation to reason about novel problems. In preparation.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., and Morales, L. e. a. (2021). 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.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229.

Grosz, B. J. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., and Goodman, N. D. (2023). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4):977.

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., and Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, 381(2251):20220047.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334.

Lew, A. K., Tessler, M. H., Mansinghka, V. K., and Tenenbaum, J. B. (2020). Leveraging unstructured statistical knowledge in a probabilistic language of thought. In *Proceedings of the annual conference of the cognitive science society*.

Loula, J., LeBrun, B., Du, L., Lipkin, B., Pasti, C., Grand, G., Liu, T., Emara, Y., Freedman, M., Eisner, J., et al. (2025). Syntactic and semantic control of large language models via sequential monte carlo. *arXiv preprint arXiv:2504.13139*.

McCarthy, W. P., Vaduguru, S., Willis, K. D., Matejka, J., Fan, J. E., Fried, D., and Pu, Y. (2025). mrCAD: Multimodal refinement of computer-aided designs. *arXiv preprint arXiv:2504.20294*.

Mieczkowski, E., Turner, C., Vélez, N., and Griffiths, T. L. (2025). People evaluate idle collaborators based on their impact on task efficiency. *Cognition*, 264:106200.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Shin, M., Kim, J., van Opheusden, B., and Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120.

Vélez, N., Wu, C. M., Gershman, S. J., and Schulz, E. (2024). The rise and fall of technological development in virtual communities.

Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., O'Donnell, T., Lew, A. K., Andreas, J. D., Tenenbaum, J. B., et al. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *CogSci*.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., and Mansinghka, V. K. e. a. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, pages arXiv–2306.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., and Parkes, D. C. e. a. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432.

Yanai, I. and Lercher, M. J. (2024). It takes two to think. *Nature Biotechnology*, pages 1–2.

Ying, L., Truong, R., Collins, K. M., Zhang, C. E., Wei, M., Brooke-Wilson, T., Zhi-Xuan, T., Wong, L., and Tenenbaum, J. B. (2025). Language-informed synthesis of rational agent models for grounded theory-of-mind reasoning on-the-fly. *arXiv preprint arXiv:2506.16755*.

Zhang, Z., Jin, C., Jia, M. Y., and Shu, T. (2025). Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. *arXiv preprint arXiv:2502.15676*.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., and Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33:19238–19250.