AI for Mathematics: A Cognitive Science Perspective

Cedegao E. Zhang*
MIT BCS
cedzhang@mit.edu

Katherine M. Collins*
University of Cambridge
kmc61@cam.ac.uk

Adrian Weller
University of Cambridge
Alan Turing Institute
aw665@cam.ac.uk

Joshua B. Tenenbaum
MIT BCS
jbt@mit.edu

Abstract

Mathematics is one of the most powerful conceptual systems developed and used by the human species. Dreams of automated mathematicians have a storied history in artificial intelligence (AI). Rapid progress in AI, particularly propelled by advances in large language models (LLMs), has sparked renewed, widespread interest in building such systems. In this work, we reflect on these goals from a *cognitive science* perspective. We call attention to several classical and ongoing research directions from cognitive science, which we believe are valuable for AI practitioners to consider when seeking to build truly human (or superhuman)-level mathematical systems. We close with open discussions and questions that we believe necessitate a multi-disciplinary perspective—cognitive scientists working in tandem with AI researchers and mathematicians—as we move toward better mathematical AI systems which not only help us push the frontier of the mathematics, but also offer glimpses into how we as humans are even capable of such great cognitive feats.

1 Introduction

Building computational systems that understand and practice mathematics at the level of human mathematicians has been a long-standing aspiration of artificial intelligence (AI) [1–15]. The rise of large language models (LLMs) has sparked imaginations that we are closer than ever to attaining, or surpassing, human-level performance on a range of tasks [16–18]. Yet, simultaneously, something seems amiss: despite these models achieving tremendous performance in many realms of human expertise (e.g., medicine, law, creative writing), the performance of these models on *mathematics* specifically lags behind [19–22]. There are many efforts to improve the mathematical problem-solving capabilities of LLMs, such as adjusting the training data and feedback strategies [23–27], equipping models with expanded background knowledge at inference-time [28], or composing LLMs with existing computational mathematics systems [29–33]. Recent efforts to build in principles from cognitive science, e.g., the importance of learning abstractions, have also seen success [34, 35]; however, we believe that the broader AI-mathematics community still has much to draw from cognitive science: in the questions we ask and the methods by which we approach such challenges.

First, we believe it is essential to reflect on what goals we are even trying to pursue. What does it *mean* for AI systems to excel at mathematics *at or beyond* a human level? Is simply excelling at a suite of standard benchmark datasets sufficient? While there is no doubt value in benchmarks to spur progress, humans—and human mathematicians—are capable of so much more than what can be captured in a static benchmark. We are capable of *intuitions* and *judgments* [36], of reasoning

*These authors contributed equally to this work.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Workshop on MATH-AI.

about the world *as world* [37], of seeking deeper explanations and understandings of results [38], of flexibly developing new problem solving tactics and not just solving new problems, but *posing* them too [39, 40]. How then should we proceed to develop human-level AI mathematicians? In the rest of this position paper, we argue that perspectives from cognitive science have a lot to offer in this new age of LLMs. Cognitive scientists, AI researchers, and mathematicians can productively contribute together to this vision towards growing flexible, automated mathematicians that help us push the frontiers of mathematical knowledge and reflect back on how we are even capable of remarkable achievements of mathematical cognition [41].

2 Looking to cognitive science

We now call attention to several classical and active research directions within cognitive science which we believe hold value for those building mathematical AI systems.

2.1 Sample-efficient learning

One of the hallmarks of human cognition is our ability to learn new concepts, knowledge, and problem-solving strategies, from little data [42–47]. In mathematics, data paucity is a particular conundrum, e.g., it is costly and difficult to obtain high-quality data on advanced topics, and few texts may exist on the cutting-edge or more obscure branches of mathematics. On the other hand, human mathematicians, from early learners to expert-level mathematicians, do not need millions of examples to learn mathematical concepts and problem-solving strategies. Yet, even though the rote *number* of examples developing mathematicians may be exposed to is small, that does not mean that a concept is grasped *immediately* upon exposure. It make take a human multiple encounters with an example, extended time sitting and thinking—squeezing out a tremendous amount from a handful of examples, e.g., through active engagement like self-explanation (see below)—to master a concept or strategy, after which such knowledge can be readily generalized to new situations [44, 48–51].

2.2 Concepts, representations, and world models

It is inspiring to reflect on the sample-efficiency of human learning. If we are to obtain or surpass such capabilities in AI systems, it is important to examine *how* humans may achieve such efficiency in the first place. Towards this end, we point to the rich cognitive science literature on concepts, their representations, and how the human mind builds rich models of the world out of concepts [52–55].

In cognitive science, much research in cognitive science points to powerful inductive biases gleaned through evolution: "core knowledge" [56, 57]. It has been speculated that a core "number sense" [58] forms the foundation upon which our mathematical prowess is built. Strong evidence points to two core number systems—for reasoning about numerosity exactly, and approximately [49, 58, 59]. From these core knowledge systems, we can develop *concepts* [42, 54, 60]. Notably in mathematics, concepts have *precise definitions*, unlike other abstract concepts such as justice and knowledge or everyday concepts like chair. At the same time, mathematicians think about concepts more than in terms of definitions; they can give examples and counterexamples, draw out relationships between concepts, and so on—this type of conceptual richness is compatible with the psychological theory of conceptual-role semantics [52].

So, what are the *form(s)* of these conceptual representations? Contemporary cognitive science has provided strong evidence for that conceptual representations may be modeled by "languages of thought" [61–64], which in mathematics, may be built over core geometric primitives [65]. Closely linked with "languages of thought" is the notion of a "world model". In AI, many have highlighted the importance of world models, although researchers disagree about how to build such models within AI systems [66–68, 64]. It is generally accepted that a world model should support simulation of possibilities, causal and counterfactual reasoning, and calibrated judgements about belief and truth [43, 67–74]. We hypothesize that the intuitions that mathematicians acquire over years of practice can be seen as forming world models of the mathematical universe. Here, we use the famous "$\mathsf{P} = \mathsf{NP}$?" problem as an illustrative example. Most people believe that $\mathsf{P} \neq \mathsf{NP}$. It seems that much evidence of such strong beliefs over an unproven statement comes from simulating what would happen if $\mathsf{P} = \mathsf{NP}$ or $\mathsf{P} \neq \mathsf{NP}$. If the former is true, many counter-intuitive consequence would follow, whereas we would not need to heavily adjust our other beliefs about computation if the latter is true [75–77]. This kind of simulation and argumentation, we suggest, may be powered by world models.

2.3 Goals, planning, and agency

Today, the dominant paradigm for large language models is a passive one: a (very large) training corpus is provided to the model, and the model optimizes some given objective function [78, 16]. At inference-time, a model is presented with a problem (e.g., a translation or reasoning task) and tries to make good predictions. However, this is not how humans think about or perform problem-solving. Humans are planning agents with goals spanning across different communities and timescales [79, 80]. When planning to achieve a goal, we can flexibly divide a task into sub-goals, form and leverage simplified abstract representations to inform planning, and replan [81–85]. Planning is crucial to success in mathematical reasoning. Consider when a teacher gives a student a problem to solve; the student needs to generate sub-goals and come up with strategies, such as looking up definitions, consider examples, examine different cases, or simply look for help. Moreover, mathematical cognition is not just about planning for set goals, but *inventing* new goals, problems, and concepts [39–41, 51]. How do some mathematicians *form the goal* of inventing new mathematics, and how do they achieve it? Engineering and scientific insights on these questions—drawing on cognitive science, AI, and mathematics—may drive a huge leap forward towards creative AI mathematicians.

2.4 Cognitive limitations and resource-rationality

However, humans are far imperfect planners, and they may fail to execute the plans we do embark upon. Mathematicians may become wedded to a particular proof strategy only to realize it was misguided and need to backtrack, or worse, could fall prey to functional fixedness [86, 87] and the sunk cost fallacy [88]. Such instances put a damper in the notion that humans are rational reasoners [89]. Cognitive scientists here too have developed rich frameworks to reconcile such challenges. Rather, we may be viewed as rational *in light of resource constraints*, i.e., "resource-rational" [90–93]. This notion finds particular importance when thinking about humans and AI systems. Fundamentally, humans and computational systems have different resource limitations: computers are able to make calculations extremely fast, are not constrained to the same limitations on working memory, and do not succumb to daily inevitable fatigue that we humans do. When building mathematical AI systems then, it is prudent to question whether we should be designing AI systems to mimic human resource constraints [92]. If trying to build a computational "thought partner" to complement humans and enable us to explore greater mathematical depths than we have so far, for instance, by making more calculations and proposing possible new patterns in troves of data [15], then perhaps we do not want to curtail a model's resources. However, one could argue that perhaps, such resource limitations are not a failing, but rather an *advantage*: for instance, empowering us to judiciously *select* which problems to solve in the first place. Indeed, mathematics communities (generally) do not waste too much time on problems that people believe to be out of reach. Studying under what settings resource limitations on mathematical cognition are advantageous, and when they are not, is a ripe space for collaboration across cognitive science, mathematics, and AI, particularly when thinking about making sensible use of limited resources even present in large-scale AI systems [94, 95].

2.5 Communication and explanation

We close our tour of cognitive science insights to spark the imaginations of those seeking to build mathematical AI by reiterating that mathematics is a *group activity* consisting of communities, and development of knowledge in any intellectual community depends on effective *communication*. We argue that a cognitive perspective on communication is valuable for the math-AI community for two core reasons. First, the *output* of our communication amongst each other forms the bedrock of the data used to train LLMs. Second, insights from cognitive science reveal that communication can spur learning for the communicator [96]. We start by reflecting on the latter.

Ample evidence in cognitive science reveals the power of self-explanation for improving learning and generalization [96–102]. Explanations can help the explainer identify abstractions to inform induction [96, 99] and reveal gaps in one's own knowledge [97], motivating information-seeking to resolve such gaps [100]. At first glance, recent LLM research such as chain-of-thought-prompting [103], "self-taught reasoning" [104], "self-reflection" [105] could be viewed as self-explanation to improve reasoning, but we encourage ruminating on the cognitive underpinnings. In fact, we argue that these are *not* instances of self-explanation in the way that humans self-explain. For humans, self-explanation is something that we *want* to do, because understanding is intrinsically valuable [99, 106]. Thus, it is desirable to not just have new prompting strategies leveraging explanations, but systems designed with explanations at their core.

And what about communication to others? We externalize many of our inner thoughts, whether that be writing out the steps of a new proof, drawing diagrams to convey a concept, or debating with a friend what the largest possible number is. These externalized thoughts and interactions increasingly form the bedrock of training data for AI systems. Nonetheless, humans do not communicate *all* of our inner thoughts; rather, we communicate what we believe is essential to convey [107]—often requiring the listener to make inferences about what the communicator *intends* to communicate (which may differ from what they *actually* produced) [108–110]. Such communication frameworks may be important for building mathematical AI systems that can adequately "read between the lines" in the data available and recognize that when providing mathematical assistance to humans, humans *are capable* of such inferences (e.g., we do not always require overly verbose responses and in fact may find it less helpful in mathematics [19]).

3 Concluding remarks

Catalyzing community cross-talk As we highlight, the cognitive science community has been studying topics deeply relevant to mathematical AI. We hope our piece helps further expose AI practitioners and mathematicians to what we believe are valuable terminology and conceptual structures from cognitive science. Cognitive scientists too can sharpen our theories from further exchanges across communities; we lay out a few strategies to facilitate such conversations. First, accessibility of higher-level mathematics is perhaps one of the most pernicious barriers to effective collaboration across cognitive scientists and AI practitioners in the space. Convenings designed to engage not just the AI and mathematics community but also cognitive scientists would aid in building a shared vocabulary across these communities. Second, there is a need for improved *research tools* to empower the study of mathematics across our communities. For more than a decade, cognitive scientists and AI practitioners alike have benefited enormously from crowdsourcing platforms such as Amazon Mechanical Turk [111] and Prolific [112]. However, at present, it is hard to find targeted domain practitioners on such sites. We suggest that it would be extremely valuable for the community to discuss the idea of a "Mechanical Turk for mathematics"; i.e., a platform where AI and cognitive scientists can post studies, questions, data gathering attempts about mathematics and mathematicians and students can participate in them. Ideally, such an effort could benefit all parties involved. Third, we note that a strong catalyst for collaboration can be a shared goal [113]. We point to *games* as a sensible playground which may appeal to mathematicians, cognitive scientists, and AI practitioners. Games have been ripe grounds for study in both AI [114–116] and cognitive science [81, 117–119], and as recently exposed by Poesia [120], aspects of mathematics itself may be cast in the language of games. We see this as a particularly exciting framing that allows us to better understand many aspects of mathematics with the help of mathematicians.

Looking forward With the resurgence in interest around AI and mathematics, we emphasize the value of engaging with the cognitive science community in the quest towards more powerful automated mathematicians. Engaging across the mathematics, cognitive science, and AI communities is paramount in even defining what this quest is and where we intend to go. To close, we propose several directions of inquiry that we think the nexus of the cognitive science, AI, and mathematics communities are poised to address. For instance, advances in AI can serve as tools to help us better understand the relationship between mathematical problem-solving capabilities and the *modalities* of mathematical data—language (natural and formal) alongside figures and diagrams; what makes a problem easy or hard (and how this differs across humans and AI systems); what kinds of prior knowledge, including human commonsense knowledge, is necessary to learn mathematics; and what are the computational foundations of mathematical insights. We believe that steps along these directions, taken across our communities, can not only spur the development of truly powerful AI mathematicians, but also shed light on what is so special about *humans'* feats of mathematical cognition—sparking efforts to improved tailored mathematical education and push the boundaries of what we, jointly with AI systems, understand about the wonderful world of mathematics.

Acknowledgements

We thank Timothy Gowers, Gabriel Poesia, and Roger Levy for comments on earlier drafts. We thank Noah Goodman, Raymond Wang, and Lionel Wong for discussions related to this work. We also thank Albert Jiang, Mateja Jamnik, the Human-Oriented Automated Theorem Proving System Team at Cambridge, and the Spring 2023 GPS Seminar at MIT for conversations that inspired aspects of this work. KMC acknowledges funding from the Marshall Commission and Cambridge Trust. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. JBT acknowledges funding from AFOSR Grant #FA9550-22-1-0387 and the MIT-IBM Watson AI Lab.

References

[1] Alan M. Turing. Intelligent machinery. Technical report, National Physical Laboratory, 1948.

[2] Allen Newell and Herbert A. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

[3] A. Newell, J. C. Shaw, and H. A. Simon. Empirical explorations of the logic theory machine: A case study in heuristic. In *Papers Presented at the February 26-28, 1957, Western Joint Computer Conference: Techniques for Reliability*, IRE-AIEE-ACM '57 (Western), page 218–230, New York, NY, USA, 1957. Association for Computing Machinery.

[4] Hao Wang. Toward mechanical mathematics. *IBM Journal of Research and Development*, 4(1):2–22, 1960.

[5] Martin Davis, George Logemann, and Donald Loveland. A machine program for theorem-proving. *Communications of the ACM*, 5(7):394–397, 1962.

[6] Alfred Tarski. Truth and proof. *Scientific American*, 220(6):63–77, 1969.

[7] Douglas B. Lenat. *AM: An artificial intelligence approach to discovery in mathematics as heuristic search*. PhD thesis, Stanford University, Stanford, CA, 1976. Ph.D. thesis.

[8] W. W. Bledsoe. Non-resolution theorem proving. *Artificial Intelligence*, 9:1–35, 1977.

[9] Alan Bundy. *The Computer Modelling of Mathematical Reasoning*. Academic Press, 1983.

[10] Alan Bundy. The use of explicit plans to guide inductive proofs. In *9th International Conference on Automated Deduction*, 1988.

[11] Alan Bundy, Andrew Stevens, F. V. Harmelen, Andrew Ireland, and Alan Smaill. Rippling: A heuristic for guiding inductive proofs. *Artificial Intelligence*, 62:185–253, 1993.

[12] Stephan Schulz. E - a brainiac theorem prover. *AI Communications*, 15:111–126, 2002.

[13] Leonardo Mendonça de Moura and Nikolaj S. Bjørner. Z3: An efficient smt solver. In *International Conference on Tools and Algorithms for Construction and Analysis of Systems*, 2008.

[14] Mohan Ganesalingam and William Timothy Gowers. A fully automatic problem solver with human-style output. *arXiv preprint arXiv:1309.4501*, 2013.

[15] Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[17] OpenAI. Gpt-4 technical report, 2023.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[19] Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*, 2023.

[20] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

[21] Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*, 2023.

[22] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*, 2023.

[23] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[24] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

[25] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[26] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

[27] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[28] Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Naturalprover: Grounded mathematical proof generation with language models. *Advances in Neural Information Processing Systems*, 35:4913–4927, 2022.

[29] Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. *arXiv preprint arXiv:2210.12283*, 2022.

[30] Albert Q. Jiang, Wenda Li, Szymon Tworkowski, Konrad Czechowski, Tomasz Odrzygóźdź, Piotr Miłoś, Yuhuai Wu, and Mateja Jamnik. Thor: Wielding hammers to integrate language models and automated theorem provers. *Advances in Neural Information Processing Systems*, 35:8360–8373, 2022.

[31] Ernest Davis and Scott Aaronson. Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on math and science problems. *arXiv preprint arXiv:2308.05713*, 2023.

[32] Emily First, Markus N. Rabe, Talia Ringer, and Yuriy Brun. Baldur: Whole-proof generation and repair with large language models. *arXiv preprint arXiv:2303.04910*, 2023.

[33] Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. Solving math word problems by combining language models with symbolic solvers. *arXiv preprint arXiv:2304.09102*, 2023.

[34] Zhening Li, Gabriel Poesia, Omar Costilla-Reyes, Noah Goodman, and Armando Solar-Lezama. LEMMA: Bootstrapping high-level mathematical reasoning with learned symbolic abstractions. *arXiv preprint arXiv:2211.08671*, 2022.

[35] Gabriel Poesia and Noah D. Goodman. Peano: learning formal mathematical reasoning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220044, 2023.

[36] Stanislas Dehaene. Origins of mathematical intuitions: The case of arithmetic. *Annals of the New York Academy of Sciences*, 1156(1):232–259, 2009.

[37] Brian Cantwell Smith. *The promise of artificial intelligence: Reckoning and judgment*. The MIT Press, 2019.

[38] Paolo Mancosu. Mathematical explanation: Problems and prospects. *Topoi*, 20(1):97–117, 2001.

[39] Edward A. Silver and Jinfa Cai. An analysis of arithmetic problem posing by middle school students. *Journal for research in mathematics education*, 27(5):521–539, 1996.

[40] Laura Schulz. Finding new facts; thinking new thoughts. *Advances in child development and behavior*, 43:269–94, 12 2012.

[41] James L. McClelland. Capturing advanced human cognitive abilities with deep neural networks. *Trends in Cognitive Sciences*, 26(12):1047–1050, 2022.

[42] Susan Carey. *The Origin of Concepts*. Oxford University Press, 2009.

[43] 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.

[44] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[45] Alison Gopnik. The theory theory as an alternative to the innateness hypothesis. *Chomsky and his critics*, pages 238–254, 2003.

[46] Jennifer A. Kaminski, Vladimir M. Sloutsky, and Andrew F. Heckler. The advantage of abstract examples in learning math. *Science*, 320(5875):454–455, 2008.

[47] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[48] Xinming Zhu and Herbert A. Simon. Learning mathematics from examples and by doing. *Cognition and instruction*, 4(3):137–166, 1987.

[49] Stanislas Dehaene. Varieties of numerical abilities. *Cognition*, 44(1-2):1–42, 1992.

[50] Lara Alcock, Mark Hodds, Somali Roy, and Matthew Inglis. Investigating and improving undergraduate proof comprehension. *Notices of the AMS*, 62(7):742–752, 2015.

[51] Frank K. Lester and Jinfa Cai. Can mathematical problem solving be taught? Preliminary answers from 30 years of research. In Patricio Felmer, Erkki Pehkonen, and Jeremy Kilpatrick, editors, *Posing and Solving Mathematical Problems: Advances and New Perspectives*, pages 117–135. Springer International Publishing, Cham, 2016.

[52] Ned Block. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10(1), 1986.

[53] Gregory Murphy. *The big book of concepts*. The MIT press, 2004.

[54] Noah D. Goodman, Joshua B. Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. In Eric Margolis and Stephen Laurence, editors, *The Conceptual Mind: New Directions in the Study of Concepts*. The MIT Press, 2015.

[55] Brenden M. Lake and Gregory L. Murphy. Word meaning in minds and machines. *Psychological Review*, 130(2):401–431, 2023.

[56] Elizabeth S. Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000.

[57] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

[58] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press, 2011.

[59] Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314, 2004.

[60] Eric Margolis and Stephen Laurence. *The Conceptual Mind: New Directions in the Study of Concepts*. The MIT Press, 2015.

[61] Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.

[62] 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):392–424, 2016.

[63] Stanislas Dehaene, Fosca Al Roumi, Yair Lakretz, Samuel Planton, and Mathias Sablé-Meyer. Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 26(9):751–766, 2022.

[64] Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, 2023.

[65] Mathias Sablé-Meyer, Kevin Ellis, Josh Tenenbaum, and Stanislas Dehaene. A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, 139:101527, 2022.

[66] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

[67] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

[68] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: A survey. *arXiv preprint arXiv:2302.07842*, 2023.

[69] Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[70] Tomer D. Ullman and Joshua B. Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2:533–558, 2020.

[71] Philip N. Johnson-Laird. *Mental models*. The MIT Press, 1989.

[72] Joshua S. Rule, Joshua B. Tenenbaum, and Steven T. Piantadosi. The child as hacker. *Trends in cognitive sciences*, 24(11):900–915, 2020.

[73] 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.

[74] Judea Pearl and Dana Mackenzie. *The book of why: The new science of cause and effect*. Basic books, 2018.

[75] W. V. O. Quine. Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

[76] Scott Aaronson. $\mathsf{P} \stackrel{?}{=} \mathsf{NP}$. In John Forbes Nash, Jr. and Michael Th. Rassias, editors, *Open Problems in Mathematics*, pages 1–122. Springer International Publishing, Cham, 2016.

[77] Timothy Gowers. What makes mathematicians believe unproved mathematical statements? *Annals of Mathematics and Philosophy*, 1(1), 2023.

[78] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[79] Michael Bratman. *Intention, plans, and practical reason*. CSLI Publications, 1987.

[80] Michael Tomasello. *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press, 2022.

[81] Allen Newell and Herbert A. Simon. *Human problem solving*. Prentice-Hall, 1972.

[82] 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):e1007594, 2020.

[83] Mark K. Ho, David Abel, Carlos G. Correa, Michael L. Littman, Jonathan D. Cohen, and Thomas L. Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[84] Mark K. Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971, 2022.

[85] Carlos G. Correa, Mark K. Ho, Frederick Callaway, Nathaniel D. Daw, and Thomas L. Griffiths. Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, 19(6):e1011087, 2023.

[86] Robert E Adamson. Functional fixedness as related to problem solving: a repetition of three experiments. *Journal of experimental psychology*, 44(4):288, 1952.

[87] Mark K. Ho, Jonathan D. Cohen, and Thomas L. Griffiths. Rational simplification and rigidity in human planning. *PsyArXiv*, 2023.

[88] Hal R. Arkes and Catherine Blumer. The psychology of sunk cost. *Organizational behavior and human decision processes*, 35(1):124–140, 1985.

[89] Eldar Shafir and Robyn A. LeBoeuf. Rationality. *Annual Review of Psychology*, 53(1):491–517, 2002.

[90] Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349:273–278, 2015.

[91] 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.

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

[93] Thomas Icard. Resource rationality. Book manuscript, 2023.

[94] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[95] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Commun. ACM*, 63(12):54–63, 2020.

[96] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[97] Michelene T.H. Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182, 1989.

[98] Michelene T.H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

[99] Joseph J. Williams and Tania Lombrozo. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, 34(5):776–806, 2010.

[100] Elizabeth Baraff Bonawitz, Tessa J.P. van Schijndel, Daniel Friel, and Laura Schulz. Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, 64(4):215–234, 2012.

[101] Mark Hodds, Lara Alcock, and Matthew Inglis. Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, 45(1):62–101, 2014.

[102] Bethany Rittle-Johnson. Developing mathematics knowledge. *Child Development Perspectives*, 11(3):184–190, 2017.

[103] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[104] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[105] Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

[106] Tania Lombrozo. The instrumental value of explanations. *Philosophy Compass*, 6(8):539–551, 2011.

[107] Herbert P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[108] Stephen C. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983.

[109] Dan Sperber and Wilson Deirdre. *Relevance: Communication and cognition*. Blackwell Publishing, 2 edition, 1995.

[110] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

[111] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[112] Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[113] Muzafer Sherif, O.J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif. *The Robbers Cave experiment: Intergroup conflict and cooperation*. Wesleyan University Press, 1988.

[114] Murray Campbell, A Joseph Hoane Jr., and Feng-hsiung Hsu. Deep Blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[115] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[116] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[117] Fernand Gobet, Jean Retschitzki, and Alex de Voogt. *Moves in mind: The psychology of board games*. Psychology Press, 2004.

[118] Pedro A. Tsividis, Thomas Pouncy, Jaqueline L. Xu, Joshua B. Tenenbaum, and Samuel J. Gershman. Human learning in Atari. In *2017 AAAI Spring Symposium Series*, 2017.

[119] Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E. Fan, Samuel J. Gershman, Alison Gopnik, Thomas L. Griffiths, Joshua K. Hartshorne, Tobias U. Hauser, Mark K. Ho, Joshua R. de Leeuw, Wei Ji Ma, Kou Murayama, Jonathan D. Nelson, Bas van Opheusden, Thomas Pouncy, Janet Rafner, Iyad Rahwan, Robb Rutledge, Jacob Friis Sherson, Ozgur Simsek, Hugo Spiers, Christopher Summerfield, Mirko Thalmann, Natalia Vélez, Andrew J. Watrous, Joshua B. Tenenbaum, and Eric Schulz. Using games to understand the mind. *PsyArXiv*, 2023.

[120] Gabriel Poesia. Research agenda. https://gpoesia.com/research/. Accessed: 2023-10-02.