Language Is Not All You Need … but Language, Probabilistic Programs & Bayesian Models of Cognition Will Get You Pretty Far

Joshua B. Tenenbaum

Since their origins in the 1950s, cognitive science and artificial intelligence have made slow but steady progress together toward a functional understanding of human intelligence. The arrival of large language models (LLMs) has upended this dynamic, with unprecedented commercial investment driven by the bet that superhuman AI could emerge simply from learning patterns in language at sufficient scale. While language is a singular tool for human thinking, there is far more to intelligence than language, as evidenced by how young children and nonhuman animals learn quickly and robustly even without language, and by the continuing jagged frontier of successes and failures in LLM-based AI. This essay considers another route to intelligent machines, grounded in principled theories and well-tested models of how minds and brains think before language and how learning language transforms thinking. By deploying AI breakthroughs in LLMs as models of language use – rather than as end-to-end models of intelligence – and connecting them to models of the thinking and learning our minds do prior to and independent of language, we have the opportunity not only to build more robust and efficient AI systems, but to rebuild the bridge between cognitive science and AI. This is a route to answering the biggest open questions about how human minds work, and with that understanding, making AI that positively impacts mental health, education, and society in ways unlikely to come from machine learning alone.

Language is the human superpower. Animals without language, including prelinguistic human children, can learn about the world, make reasonable inferences and decisions in scenarios they've never experienced, and even understand each other's actions and goals. But with language, these capacities vastly expand and deepen: We can learn from others, including people who lived long before we were born, and not only from our direct experience. We can cooperate and collaborate with others much more flexibly and effectively because we

https://doi.org/10.1162/DAED.a.980

can communicate our thinking with words. Language is also a medium of thought, for long-term planning and problem solving, and a substrate for associative memory and analogical retrieval. It lets us connect facts, stories, conversations, sights, sounds, tastes, and emotions – everything that we can talk about, however vaguely – to marshal from a lifetime of experience the knowledge and understanding we most need to be thinking with in any given moment.

The potential to capture these capabilities, and more, is what makes recent breakthroughs in AI language models so enticing. Artificial intelligence appears close to achieving the field's oldest dream of an intelligent machine that you interact with as you would interact with another person – by talking to it. This is how AI has always presented itself to the broader world, dating back to Alan Turing's 1950 paper on "Computing Machinery and Intelligence" – in which he famously proposed a chatbot-like imitation game (now known as the "Turing Test") as a way to operationalize his vision of a thinking machine – or in science fiction films from *2001: A Space Odyssey* (1968) to *Star Wars* (1977), *Star Trek* (2009), *Her* (2013), and many others.[1] Given their sci-fi-like capabilities, it's no wonder that large language models (LLMs) are generating such excitement and enthusiasm in the general public, and that companies, governments, and investors of all kinds are putting significant capital and energy into them. We seem to be on the precipice of developing true thinking machines.

And yet, despite the many billions invested in LLM development over the last several years, and remarkable successes in building systems that can solve problems we used to think only humans could solve, there is something missing from this view of intelligence. LLM-based AI has been called "weird," "alien," "blurry," and most tellingly "jagged": that is, the frontiers of its capability do not advance smoothly in ways we have come to expect from how human intelligence develops; rather, we see world-class or even superhuman levels of AI expertise in some very challenging tasks and contexts, combined with surprisingly poor performance in much easier tasks.[2] When OpenAI released its first "reasoning-trained" model, o1, in December 2024, it scored as high as the top 0.1 percent of American high school students on the AIME, one of the hardest nationwide math contests, but it could not successfully play tic-tac-toe. A year later in December 2025, OpenAI's latest GPT-5.2 model achieved a perfect (100 percent) score on the AIME but got less than 75 percent correct on a math contest designed for third and fourth graders.

As a cognitive scientist, I appreciate both why these systems seem so promising and why they fall short at achieving what we think of as human-like intelligence. The brain's capacity for language is one of the most crucial, and most obvious, aspects of our species' advanced cognition. But precisely because it's so central to how we think, we tend to conflate language fluency with intelligence itself, leading to predictable biases against people with accents and people who stutter, and biases for people and tools we perceive as fluent.[3] It's no wonder, then, that we might tend to

think of LLMs, which can generate fluid prose, poetry, and code many times faster than humans, as exceptionally intelligent.

When you dig deeper from a cognitive science perspective, what you discover is more complicated. In a series of papers, my colleagues Evelina Fedorenko, Roger Levy, and their labs (in language and computational psycholinguistics) at MIT have shown that even small language models – such as OpenAI's GPT-2, trained on more than a human lifetime's worth of data but orders of magnitude less data than GPT-5 – can quantitatively predict language-specific neural responses in the human brain and behavioral signatures of sentence comprehension (such as the extra time it takes you to read a sentence with a surprising twist).[4] These models capture "formal linguistic competence," what you implicitly know about the rules of language that lets you interpret and produce sentences, but fall far short on tasks of "functional linguistic competence," or any kind of thinking that requires synthesizing language facility with extralinguistic capabilities, like logical and causal reasoning, world modeling, and social cognition. These results, and other converging findings from LLMs trained on child-sized datasets and fMRI studies of how language in the brain is distinct from and interacts with other nonlinguistic brain systems, suggest that while language models are useful as models of human language processing, language alone is not all we need to build thinking machines or an accurate theory of how our minds work.[5]

I have spent the last twenty-five years building computational models of cognition, and specifically Bayesian models, which represent human thinking as a form of rational, approximate probabilistic inference, updating beliefs about the world by combining prior knowledge with newly observed evidence. So it will be no surprise that I believe developments in Bayesian cognitive science might hold the keys to building such a theory and a new generation of more human-like intelligence in AI. But I also believe that both cognitive science and artificial intelligence are at their best when they work in dialogue with each other.

These two fields grew up together in the 1950s and, for most of their history, made progress toward understanding intelligence by growing in conversation with each other. In recent years, this tradition of bidirectional and diverse exchanges has taken a back seat to Silicon Valley–driven scaling of a single architecture based on a single idea from machine learning: the transformer-based language model. While this will surely continue to produce economic and productivity gains, I believe that making further progress toward understanding the mind, and using that understanding to produce more rational and human-like forms of AI, is more likely to be achieved if we can rebuild the historical bridge between cognitive science and computer science.

As a step in this direction, my colleagues and I have been exploring how to think about the truly impressive advances in language modeling offered by LLMs through

a cognitive science lens. We are finding that language models can be integrated with both classic approaches to Bayesian cognitive modeling and recent developments in scalable Bayesian inference from probabilistic programming to open up a new and deeper understanding of human intelligence, and perhaps more intelligent AI.

Our starting point is the idea that intelligence starts with the capacity to model a changing world, to make inferences about the world's state and the likely effects of your actions in it, and to choose actions flexibly that you expect could achieve your goals. These capacities don't simply emerge over time as the result of learning to predict patterns in image sequences, speech streams, and forms of sensory data. Rather, they are core functions of brains that exist from the start and are the basis for how we can learn so much from so little information. This is true not just for humans but for animals whose brains are many times smaller than our own. Mice, for example, figure out how to move their bodies and manipulate objects in space to achieve their goals, despite their brains being about a thousand times smaller than ours and despite having no massive corpus of training data to draw on and no language capability. Like our brains, theirs are designed from the beginning to build internal models of the physical world that generalize to new situations beyond their direct experience, and to quickly update those models in response to their experiences. Mice can be tricked or trapped by humans thinking deeper and harder about physical mechanisms to catch them; but humans know very well that building a good mousetrap isn't easy. That's the point.

Human brains are, of course, capable of far more flexible thinking and rapid learning than mouse brains. This adaptive thinking begins well before children acquire language; it precedes and grounds language, and – together with our drive to understand and be understood by others – is partly what allows us to learn language so quickly and robustly. When children learn to read and write and begin to access all the knowledge that has accumulated culturally over generations, their intelligence and ability to learn accelerate in a reinforcing singularity that is the source of our species' linguistic superpower. But the seeds of these capabilities exist well before that. Intelligence, in other words, starts not with experience, data, or language, but with brains and minds built for rationality: constructing mental models to make good guesses and bets about the world and what one should do in it.

This is why I am skeptical that the "scaling thesis" for building AI, dominant in Silicon Valley today, is the best route to an understanding of the brain or to building machines with truly human-like intelligence. Unlike the human approach to *growing* intelligence from a seed of rationality, attempts to *scale* intelligence by scaling up data and compute start only with simple, "dumb-by-design" mechanisms of learning predictive patterns in data, and then hope everything else – world models, thinking and reasoning, social cognition, morality, and so on – emerges from it. We shouldn't be surprised when such models turn out to be jagged or brittle, require carefully curated training data to be effective across different scenarios,

generalize only weakly and unpredictably to settings outside of their training data ("out of distribution"), and become increasingly costly to update in response to new experiences.

These observations aren't meant as a criticism of the scaling route – it's a bet worth pursuing, at least if you have a trillion dollars. But the differences between the "growing up" and "scaling up" approaches might explain why LLMs can feel head-scratching even when they are working, and why it is worth also making bets on building systems that grow into intelligence the way a human being does.

This means starting with a computational architecture, as all brains do, for agency based on rational world modeling, and then introducing language in the way the human brain does, as a way of supercharging this architecture to make it vastly more flexible, generalizable, shareable, and scalable.

What precisely is an architecture for "rational world modeling"? This means constructing a mental model of our environment and – as a function of the state of the world we perceive, the actions we can take, and our goals – assessing the rewards of possible future states, the potential costs of actions, and the probabilities of our actions achieving those future states. From these components, we can compute the value of taking different actions by estimating which action sequences are likely to lead to high utility (rewards minus costs). We make these probabilistic inferences based on some combination of memories – how did similar actions in similar states turn out in the past? – and mental simulation, imagining based on our understanding of causal mechanisms how novel actions in novel states might unfold.

This architecture is not new and it is not uniquely human. It derives from the classical picture of rational agents as expected-utility maximizers, and in a deep sense it characterizes how all animals' brains are designed to make good bets to achieve their goals.[6] But in human minds, these processes are massively enriched and extended. Our world models include not just mental representations of the physical world but models of other agents in the world around us and, recursively, our approximations of their inner world models. Our utilities include not just basic survival needs, but our values, pleasures, and preferences, and those of other agents around us, including social norms and moral frameworks. And our decision and planning processes unfold not just in immediate fight-or-flight scenarios but over hours, days, and even years. When you add thinking and meta-reasoning to this architecture, it gets even more complex and powerful. We're not just deploying basic world models anymore, we're constructing and manipulating new models – including recursively and hierarchically composed models, and reduced models to help us plan efficiently – and extending our rational decision-making to our own internal computations, making good bets about how to use our computational resources.[7]

Language is partly what enables these distinctively human forms of world modeling, and what also transforms our thinking and learning in other ways. It gives our minds tools for marshaling, analogizing, recombining, and reinventing world

models from all the mental pieces we've ever built; tools for learning how to build new models, not only based on direct experience but, socially and culturally, from interactions with others and from accessing the vast stores of human knowledge accumulated over generations; and tools for imagining models of hypothetical or counterfactual ways the world could be or could have been but that have never appeared in any human being's experience.

How close are cognitive scientists to being able to implement this kind of architecture in a machine-executable model or a practical AI system? Not nearly as far along as the industry scaling route, but a lot closer than most of the world recognizes. The key pieces are in place and being prototyped on small scales, and we are ready to try larger-scale experiments that could be carried out with even a tiny fraction of the investment currently being poured into the scaling route.

This has become possible by building on three phases of research from the past twenty-five years of computational cognitive science. Starting in the late 1990s, under the banner of "Bayesian models of cognition," researchers began developing an approach to reverse-engineering the mind as an architecture for general-purpose and flexible rational world modeling. This work is reviewed in a 2024 book by Thomas Griffiths, Nick Chater, and me, written with many students, collaborators, and colleagues who have contributed to this research program and made it their own.[8] Through their efforts, we now have a robust paradigm for building theoretically grounded, quantitatively predictive models of human learning and thinking that works in virtually every domain of cognition that behavioral scientists have studied, including causal reasoning, language acquisition and use, intuitive physics, decision-making and action understanding, social and moral cognition, and much more.

The most sophisticated of these Bayesian models have been built over the last decade, enabled by the advent of modern "probabilistic programming" tools starting around 2015. Probabilistic programs provide expressive languages for writing computational models of rational agents (natural or artificial) and their world models, as well as the inference algorithms agents use to perceive, reason, plan, learn, and communicate with their models.[9] These expressive languages bring together the most powerful ideas from three influential paradigms for understanding intelligence as computation that have emerged over the history of AI and cognitive science: *symbolic* approaches, with their strengths of precision, abstraction, modularity, and composition; *neural-network* (or *differentiable-programming*) approaches, with their strengths of distributed representations, efficient parallelism, and gradient-based optimization; and *probabilistic* (or *Bayesian*) approaches with their strengths of rational inference and decision-making under uncertainty and risk. By allowing Bayesian modelers to take full advantage of symbolic and neural-network approaches and tools, probabilistic programming languages (PPLs) have made it possible to rapidly prototype, test, and scale cognitive models we could barely conceive of before 2010.

Probabilistic programming as a discipline is young but growing. In the last few years, we have seen these frameworks integrated with the same graphics processing unit (GPU) backends and transformer-based architectures that have driven the neural network revolution, and a proliferation of new PPLs are already being used to make Bayesian models of cognition more flexible, more efficient, or both.[10] The arrival of code-trained LLMs and agentic AI coding environments (such as Anthropic's Claude Code) that can understand and integrate with PPLs will only accelerate this trend. Over the next two years, I expect we will increasingly see Bayesian models that used to take weeks to program manually, and were painfully slow to compute with, now being programmed in a day or less and tested almost instantly – much closer to the speed of thought.

More deeply, these tools are starting to make it possible to build the full capacities of natural language into our rational cognitive architectures. Though not standalone models of human intelligence, LLMs can be integrated with Bayesian cognitive models and probabilistic programs to build the first broad-coverage models of human thinking as it is extended and transformed by language. We can now test hypotheses about the many functional contributions language makes to cognition in ways that have not been possible before and, in the rest of this essay, I will briefly survey two ongoing efforts: one formalizes how language communicates structured thoughts in conversation, to build and use a world model jointly between two or more agents; the second formalizes how language can be used within a single mind (human or AI) to build flexible, bespoke world models for a new situation, by synthesizing, composing, and recombining pieces of other models marshaled from the agent's memory.

One long-standing view of the function of language is as an external medium for communicating human thoughts. Essentially, there is a mapping from an internal representation of thought into the external symbol systems that we call language, and learning language is learning this mapping. Everyone may have a different internal representational code, but presumably there is enough shared structure among minds that every child can learn an appropriate mapping. If this is right, we should be able to describe this process in a general way that starts with an underlying internal language designed for thinking – that is, making good guesses and bets about the world using the probabilistic models that are core features of the brain – and connecting this language of thought to externalizable symbol systems (spoken words or signs) via a probabilistic model of communication. This approach supports bidirectional mappings for both understanding the language of others and externalizing one's own thoughts.

In the paper "From Word Models to World Models," by the cognitive science and AI researchers Lio Wong, Gabriel Grand, Alexander Lew, and others, we explore a method to do just that, which Wong calls "rational meaning construction."[11] Our

work uses LLMs, but not as end-to-end thinking systems. Instead, they are used to model this superpower that language gives human minds: a tool for mapping between the external language of a conversation with other minds and the mind's own internal languages of thought.

More technically, Wong and colleagues show how to integrate LLMs with Bayesian models written in a PPL to give an architecture for "language-informed thinking," a proposal that contrasts dramatically with today's AI industry scaling route. Rather than attempting to learn all of thinking from patterns in language, the focus is an agent architecture that starts with a substrate designed for rational thinking: a probabilistic language of thought (PLoT), which we formally instantiate using probabilistic programs that express an AI agent's foundational knowledge (or priors) over possible worlds. Then we use LLMs to translate natural language into these probabilistic programs, which in turn are used to perform context-sensitive Bayesian inference over distributions of possible worlds to infer what is likely to be true or likely to be a good answer to any question or decision, updated from an agent's priors in light of what they have observed or been told in a conversation.[12] Relative to today's frontier AI models, we can use much smaller, more data- and energy-efficient LLMs to implement this mapping – perhaps closer to the resources of the human brain – because the LLM is only being asked to understand language, not to implement (or imitate) reasoning.

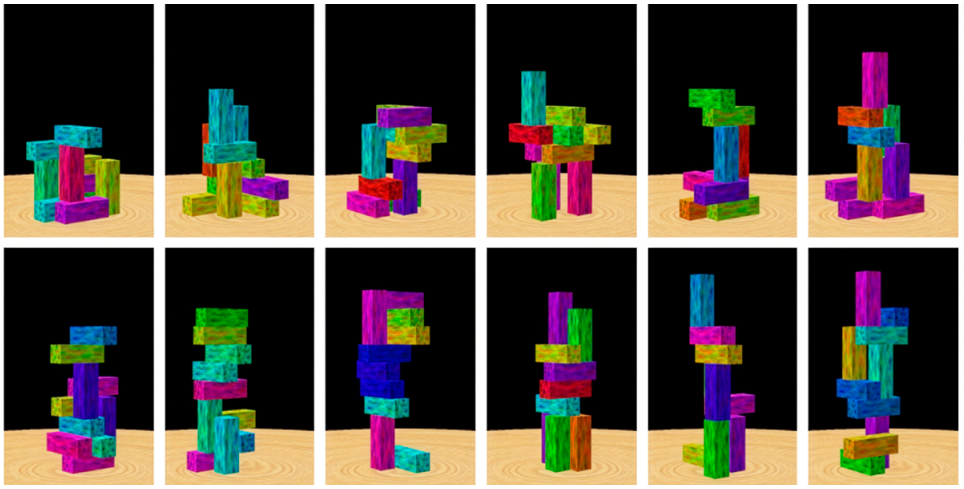
I don't have space here to cover all the recent developments that demonstrate the promise of this approach, but I will highlight just one. In a study published in January 2025, psychology and machine-learning researcher Luca Schulze Buschoff and colleagues evaluated how well state-of-the-art multimodal (vision-trained) LLMs captured the judgments of human participants in simple commonsense settings: intuitive physics, causal reasoning, and intuitive psychology.[13] The models struggled to match human patterns of reasoning, such as predicting whether a stack of blocks shown in an image is stable, even though they were able to perform basic visual operations on the same scenes, such as recognizing the colors of the blocks.

These results are telling because cognitive scientists have long known how to capture human judgments in these tasks using Bayesian models written as probabilistic programs, and indeed the tasks studied were designed originally to test these models. For example, in intuitive physics, Peter Battaglia and Jessica Hamrick built a Bayesian model for how people make inferences about similar block-stacking scenarios like those depicted in Figure 1.

Based on reconstructing 3D scenes from images and simulating their next few moments in a video-game-style approximate physics engine, their model captured people's judgments about the stability of towers and many other scenarios like these, with high quantitative accuracy that has yet to be matched by AI systems.[14]

It is worth emphasizing how much this approach differs from traditional machine learning. The Bayesian model requires no task-specific training with

Figure 1
Example Block-Stacking Scenarios Designed to Study Human Intuitive Judgments of Physical Stability

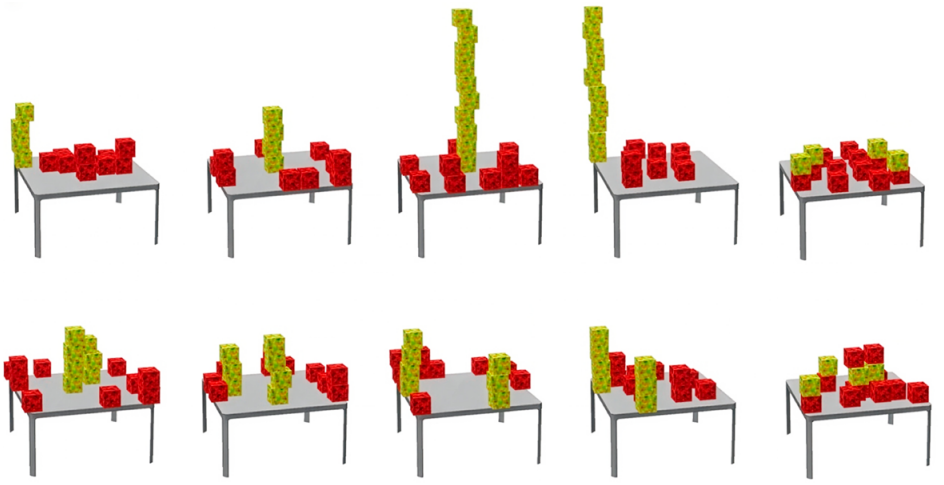


Source: Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum, "Simulation as an Engine of Physical Scene Understanding," *Proceedings of the National Academy of Sciences* 110 (45) (2013): 18327–18332.

external feedback signals to judge stability. It makes predictions purely from its own internal simulations – a form of "imagination." This allows it to generalize immediately, as people can, to tasks very different from those encountered before. For example, given tabletop scenes of red and yellow blocks like those in Figure 2, we can ask people to judge how likely it is that bumping the table will lead to more red blocks or yellow blocks falling onto the floor. The same probabilistic simulation approach captures these judgments almost as well as it does the much more familiar task of judging stability.

This is also where we can see the power of bringing language into the picture, as well as the idea of rational meaning construction at work. What if instead of being asked to make an inference based on the images above, a human or AI agent was given only a verbal description of a scene? If I tell you that there are several tall stacks of red blocks near the center of a table, and a few short stacks at the table's edge, and less than half of the blocks in the short stacks are yellow, you would probably predict that a bump to the table is likely to cause more red blocks to fall off. If instead I told you that only one of the tall stacks had red blocks, you might instead predict more yellow blocks would fall off. You make these inferences not

Figure 2
Example Scenarios Designed to Test Probabilistic Simulation Models of Novel Intuitive Physics Judgments



Source: Illustration, produced by Peter Battaglia, courtesy of the author.

by inspecting the actual visual scene in front of you but by constructing an imagined scene in your mind's eye.

AI and cognitive science researchers Cedegao Zhang, Lio Wong, Gabriel Grand, and I captured this kind of language-based thinking in our first empirical demonstration of the rational meaning construction approach. We used an LLM to translate text descriptions of red and yellow block scenes into code for probabilistic simulation models, which can then be run in the same simulation-based inference mode as before.[15] Our experiments included both short, vague descriptions such as "a tall stack of red blocks," as well as more complex ones, with multiple stacks of blocks or an exact number stacked in more precise locations. By averaging the results across a few of these translations and a few simulations from each synthesized probabilistic model, we could quantitatively predict people's judgments almost as well as in the original, image-based study.

Recall that the Schulze Buschoff study showed that cutting-edge vision-language transformers performed far worse at much simpler tasks.[16] And yet, when paired with domain-specific Bayesian world models, even small language models were able to translate complex descriptions of scenes into reasonably accurate code simulation models that make similar judgments to people about the same scenarios. To

me this shows both the power of language to let us think flexibly about new situations, and the value of not identifying language with all of thinking.

This work is still very preliminary, and intuitive physics is just one domain of cognition. But we are already seeing similar advantages for rational meaning construction in other settings of complex language-informed thinking: for instance, in modeling how people understand agents' beliefs, desires, and preferences from observing their interactions and communication, or how they form new causal world models in gameplay from combinations of experience and advice from previous players.[17] At this point, I'd say it's a reasonable bet that somehow combining Bayesian inference in probabilistic programs as the rational core of human thinking with code-trained language models to capture the way people spin words into these world models can take us far in both computational cognitive science and AI.

And yet if we are looking for a full theory of human cognitive architecture, any such approach still faces major challenges of scaling and generality. Most immediately, in each domain where we have built highly predictive quantitative models of human reasoning, the probabilistic code for world modeling has been written at least partly by human cognitive scientists. But in the real world, no external designer is there to program our mind's world models for us. How would all this code (or its analog) get into people's heads? And could we engineer an AI system that generates these capabilities for itself without needing to be programmed even partially from the outside?

Put another way, could we build cognitive architectures with the same inputs and outputs as today's LLMs – which, like human minds, can extend generally and flexibly to any scenario describable in words – but with internal processes that truly mirror the mind's operations? Such a system would have to be able to construct world models on its own, at least in all the situations and at the level of depth and explicitness that humans can imagine, and then reason coherently, rationally, and efficiently at least as well as humans do.

Confronting this challenge is the goal of the second line of work I want to talk about briefly. An effort led by cognitive scientists Tyler Brooke-Wilson, Lio Wong, and Katie Collins has begun exploring the hypothesis that the mind operates with what we call a "model synthesis architecture" (MSA). Brooke-Wilson originally proposed and motivated the MSA hypothesis in his doctoral dissertation at MIT, and this team (which I am part of) is now trying to implement it computationally and test it behaviorally.[18] We posit that human minds, rather than walking around with a giant monolithic probabilistic codebase expressing everything they know about the world, construct small bespoke world models on the fly, as needed, incorporating only the variables and causal dependencies that might be relevant to a given situation and the inference problem they are currently trying to solve. These ad hoc, situation-specific world models take the form of probabilistic programs that

support efficient probabilistic inference, just as in all the examples described earlier. But this code is now synthesized online by the agent via a process of "marshalling," or assembling and adapting fragments of previously constructed models based on similar problems and situations recalled from memory. This architecture thus allows minds that are computationally limited (in processing, memory, time, and other resources) to reason coherently and rationally in local contexts while drawing on globally relevant considerations from across their experiences, knowledge, and beliefs.

We have implemented a simple first prototype of an MSA using a combination of LLMs and PPLs.[19] As in rational meaning construction, where we used LLMs to capture how people use language as a tool for *externalizing* and communicating structured thoughts, here we use code-trained LLMs to capture how minds use language in a different way – as an *internal* tool for general-purpose associative memory, analogical retrieval, and mapping. These associations and analogies span multiple forms of mental representation: general, often vague, background knowledge; informal "word models" of specific situations; and more formal symbolic (PPL-based) world models. These are just the capacities needed in a system for bespoke model construction that supports broad coverage yet locally coherent rational inference.

Our MSA prototype is being tested in a classic domain of probabilistic causal reasoning: making inferences about the outcomes and traits of players in team sports matches, such as tug-of-war, doubles ping-pong, or canoe racing.[20] For instance, you might be told the outcomes of several races between teams of participants ("In the first race, Val and Gale lost to Sam and Kay. Val and Harper beat Sam and Ness in the second race…") and be asked to judge how strong you think Sam is, or how hard they tried in the first race. People can make graded probabilistic judgments like these systematically and reliably – and the MSA can explain how they do so.

Reasoning in this architecture proceeds in several stages, moving from a natural-language problem description to PPL code generation to probabilistic inference in the generated world model. Each stage is implemented through a combination of language model proposals and probabilistic reweighting of sampled model traces. First, the problem description is parsed into fragments of code expressing the observed facts and the questions to be answered. The observation "In the first race, Val and Gale lost to Sam and Kay" might be parsed into the expression

```
condition(lost({team1: ['val', 'gale'], team2: ['sam', 'kay'], race: 1}))
```

and the question "How hard did Sam try in the first race?" into

```
query: effort_level_in_race({athlete: 'sam', race: 1}).
```

This step is effectively the same as in rational meaning construction, and similar to some linguistic views of "understanding" meaning in language as a mapping into a formal predicate-based representation, but with mappings that are more pragmatic and situation-specific. The predicates have names in English, with all

the associations those carry, but they are not yet meaningfully grounded in any world model. They become formally meaningful only in the context of a bespoke probabilistic program the architecture will synthesize for reasoning about this problem. That synthesis proceeds next, by marshaling relevant general knowledge from memory and assembling it into informal (natural-language) descriptions of hypothetical dependencies between variables, along with informal probabilities. For example, the system might posit that *"Intrinsic strength varies widely from athlete to athlete, but let's imagine that players tend to be weak, average or strong, and uniformly around that . . . ,"* or *"The amount of effort that an athlete puts into any given race is a continuous parameter. . . . This effort is somewhat contingent on an athlete's intrinsic strength, because stronger athletes probably tend to be more likely to put in extra effort."* Finally, these informal model fragments are used to synthesize a formal code-based world model in a PPL; this grounds the predicates expressing the questions posed and supports coherent, rational inferences to answer them.

While still preliminary, this approach does appear capable of making judgments quantitatively similar to humans on these tasks – and outperforming simple language models on their own. In particular, the MSA prototype appears to better capture how people reason coherently in novel situations, when suddenly called upon to consider a variable or factor that just wasn't in their mental model before. We asked a second set of participants to propose new explanatory variables or factors that might affect match outcomes or how one would judge the strength of a player. Some human responses we got included: *"Avery seems to have come down with a stomach virus between rounds, but has decided to compete anyway." "Kay didn't get enough sleep last night and can barely stay awake during the race." "Ness took an energy drink and it just kicked in to give them extra energy for their matches."*[21] We then gave these same facts back to the first set of participants, in addition to match outcomes, and asked them to judge players' strengths conditioned on this new information. The MSA handles these scenarios too, following the same process as above. It first proposes, informally in natural language, causal mechanisms that link the new information to what is already known, and then translates those proposals into formal PPL representations of how its world model must be expanded. Probabilistic inference in the expanded model yields results that, although messy like the real world, are much closer to human judgments than reasoning in natural language (the pure LLM) alone.

This is just a first demonstration, but it points to the larger promise of the MSA approach. Because the models synthesized in this architecture are small ones – instantiating just the relevant variables needed for a given problem – inference and decision-making are tractable and efficient in a way they could never be if we had one big symbolic model for the whole world in our heads. Because the models are constructed on the fly, we can in principle reason about any situation, so long as we can retrieve at least some relevant background information that can connect the question to the observations we've made.

The ability to assemble the relevant variables from background experience and check them against the specific, present context also illustrates more broadly the functions natural language plays in human thinking beyond merely communicating ideas. On the one hand, language provides the substrate for almost everything we come to know beyond our direct experience – everything learned explicitly or implicitly, abstractly or concretely, from people telling us things, or reading, or other cultural means. At the same time, language is the basis for a general content-addressable associative memory that indexes this world knowledge together with every memory of the symbolic worlds or fragments of models we've built before, and allows us to efficiently retrieve candidate variables, relations between variables, and symbolic model components useful for the current task at hand.

Implementing all these functions in the weights of code-trained LLMs, as our prototype does, is not the only way to build an MSA. But it does mirror something of how neuroscientists think knowledge is represented in the weights of the brain's long-term memory: not always or even usually in an explicit and structured form, but in a distributed and implicit manner that can be marshaled into structured world models to support coherent rational reasoning as and when needed for a task.

In sum, this architecture may offer an approach to building a serious and scalable machine-executable theory of how human language and rational thought interact, consistent with our understanding of their evolutionary, developmental, and neural bases. The MSA uses the foundation of probabilistic programming to capture how people reason coherently about the world, whether in long-standing core domains such as intuitive physics and intuitive psychology or flexibly beyond those to all the situations we can think about in small bespoke models. And it incorporates breakthroughs in language modeling from artificial intelligence, not as an end-to-end paradigm for all of intelligence, but as a tool for all the ways that language extends, transforms, and scales the human capacity for rationality, mental model building, and coherent reasoning.

I want to close this essay by looking forward to the biggest open scientific questions at the intersection of human and artificial intelligence and the societal impacts that may come from answering them. There are many open questions about the model synthesis architecture, chief among them are whether and in what sense it actually resembles how the brain works. Even if it is right in broad strokes, the actual mechanics of this architecture will surely be more complex – in how and when the brain synthesizes explicit world models across different domains of cognition, what form its languages of thought takes and how they interact with natural language, and how model synthesis and inference build on and interleave in online thinking.

But to the extent that these ideas do capture something fundamental about how the brain works – how intelligence grows from a foundation of rationality, and how language both emerges from and expands these core capabilities – they could have

significant and wide-ranging implications for fields from computing and psychology to education and economics.

If we can build AI grounded in a functional understanding of human intelligence, we may develop far more efficient and effective systems, able to reason coherently and act rationally using much less data, computation, and energy. By understanding how the brain works as an organ of rationality, and how that function can break down, we may develop therapies to address mental illness and improve mental health by targeting specific broken cerebral processes. By better understanding how our minds grow, we may better diagnose and treat developmental disabilities. And by better understanding how people communicate and learn, and interpret information at scale, we may develop far more predictive economic models and effective educational techniques. In the process of making these discoveries, we may even answer some of the biggest questions our field has long puzzled over: What kind of computation is cognition? How do these computations work mechanistically in the brain? And how did our minds arise through evolution, culture, and development?

In my view, these questions and many important ones like them are unlikely to be answered purely by continuing along the LLM scaling route or by publishing bespoke cognitive models alone. To a cognitive scientist, the way large language models are used today in AI, where language is the source of thinking and everything else is a tool that these models use to become more rational, is backward. Instead, we should see language as a tool – a superpower, but still a tool – that emerges out of already rational minds and enhances and extends them. Likewise, we need to move past thinking of symbols in terms of the brittle, intractable, hard-coded formalisms of classical AI (or earlier Bayesian cognitive science). Symbolic systems need not be fixed, massive, monolithic codebases. Instead, they can be small, bespoke, and transient models, fast to build, quick to do inference in, and cheap enough to cast aside. Code doesn't need to be "written by hand" but can be constructed evolutionarily: learned from experience (including language), and rewritten, reused, and recomposed by the mind itself.

By combining the tools of modern language models, probabilistic programming, and Bayesian cognitive science, we have been able to build prototypes of this way of thinking, much in the same way we view the mind as marshaling intelligence out of multiple different systems that build on our core rationality. We are only at the beginning, but I hope this work could be the basis for rebuilding the bridge between the fields of artificial intelligence and cognitive science and their working together toward a foundational architecture that achieves human-like language and thought. That may just lead to AI systems that really are the thought partners we've always wanted, that make us truly smarter and better off, and that start to answer some of the fundamental questions about how our minds work and where our intelligence comes from.

AUTHOR'S NOTE

This essay is based on a keynote talk, "Scaling Intelligence the Human Way: Building a Shared Future for AI and Cog Sci," I presented on August 1, 2025, at the 47th Annual Meeting of the Cognitive Science Society. It represents joint work with many collaborators, but especially Lio Wong and Tyler Brooke-Wilson, who originated the ideas of rational meaning construction and model synthesis architectures (respectively) in their PhD work and who continue to lead the development of this research program. Katherine Collins provided essential contributions as a colead on much of this work, and key contributions also came from Alexander Lew, Gabriel Grand, Jacob Andreas, Vikash Mansinghka, Cedegao Zhang, Lance Ying, Tan Zhi-Xuan, Kaya Stechly, Tobi Gerstenberg, Noah Goodman, and Timothy O'Donnell. I am grateful for the support of an AI2050 Senior Fellowship from Schmidt Sciences, a MacArthur Fellowship, and grants from the Office of Naval Research, the Air Force Office of Scientific Research, the U.S. National Science Foundation, and the MIT Siegel Family Quest for Intelligence. Thanks also to Gabe Stein and Jennifer Sunoo for writing assistance, and to Mira Bernstein, Laura Schulz, and James Manyika for comments on this essay.

ABOUT THE AUTHOR

Joshua B. Tenenbaum, a Member of the American Academy since 2020, is Professor of Computational Cognitive Science in the Department of Brain and Cognitive Sciences (BCS) at MIT, a Principal Investigator at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), and the Director of Science at the MIT Siegel Family Quest for Intelligence. He was named a MacArthur Fellow in 2019 in recognition of his contributions to mathematical psychology and Bayesian cognitive science, and is the recipient of the Troland Award from the National Academy of Sciences and the Howard Crosby Warren Medal from the Society of Experimental Psychologists.

ENDNOTES

[1] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 49 (236) (1950): 433–460, https://doi.org/10.1093/mind/LIX.236.433.

[2] Andrej Karpathy (@karpathy), "Jagged Intelligence The word I came up with to describe the (strange, unintuitive) fact that state of the art LLMs can both perform extremely impressive tasks (e.g. solve complex math problems) while simultaneously struggle with some very dumb problems. E.g. example from two days ago - which number is bigger, 9.11 or 9.9? Wrong. https://t.co/3C7pCdBShQ," X (formerly Twitter), July 25, 2024, 1:50 p.m., https://x.com/karpathy/status/1816531576228053133.

[3] Kyle Mahowald and Anna A. Ivanova, "Google's Powerful AI Spotlights a Human Cognitive Glitch: Mistaking Fluent Speech for Fluent Thought," The Conversation, June 24, 2022, https://doi.org/10.64628/AAI.wvs9mp5dy.

[4] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, et al., "The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing," *Proceedings of the National Academy of Sciences* 118 (45) (2021): e2105646118, https://doi.org/10.1073

/pnas.2105646118; Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, et al., "Dissociating Language and Thought in Large Language Models," *Trends in Cognitive Sciences* 28 (6) (2024): 517–540, https://doi.org/10.1016/j.tics.2024.01.011; and Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, et al., "On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior," arXiv (2020), https://doi.org/10.48550/arXiv.2006.01912.

[5] Alex Warstadt, Aaron Mueller, Leshem Choshen, et al., "Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora," *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, ed. Alex Warstadt, Aaron Mueller, Leshem Choshen, et al. (Association for Computational Linguistics, 2023), 1–6, https://doi.org/10.18653/v1/2023.conll-babylm.1; and Colton Casto, Anna Ivanova, Evelina Fedorenko, and Nancy Kanwisher, "What Does It Mean to Understand Language?" arXiv (2025), https://doi.org/10.48550/arXiv.2511.19757.

[6] Michael Tomasello, *The Evolution of Agency: Behavioral Organization from Lizards to Humans* (MIT Press, 2022).

[7] Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum, "Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines," *Science* 349 (6245) (2015): 273–278, https://doi.org/10.1126/science.aac6076; and 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 (2019), https://doi.org/10.1017/S0140525X1900061X.

[8] Thomas L. Griffiths, Nick Chater, and Joshua B. Tenenbaum, *Bayesian Models of Cognition: Reverse Engineering the Mind* (MIT Press, 2024), https://mitpress.mit.edu/9780262049412/bayesian-models-of-cognition. This work in turn builds on pioneering rational analyses of generalization, categorization, and memory by cognitive scientist Roger Shepard and psychologist John Anderson in the late 1980s and early 1990s, and a computational framework for the study of vision pioneered by neuroscientist David Marr in his book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Company, 1982).

[9] For an introduction to probabilistic programming in cognitive science and artificial intelligence, see Noah D. Goodman, Joshua B. Tenenbaum, and the ProbMods Contributors, *Probabilistic Models of Cognition*, 2nd ed. (2016), https://www.probmods.org; and The Gen.jl Team, "Gen Tutorials," https://www.gen.dev/tutorials (accessed January 15, 2026).

[10] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, et al., "Pyro: Deep Universal Probabilistic Programming," *The Journal of Machine Learning Research* 20 (28) (2019): 1–6, https://jmlr.org/papers/v20/18-403.html; 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 *PLDI 2019: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Association for Computing Machinery, 2019), 221–236, https://doi.org/10.1145/3314221.3314642; Kartik Chandra, Tony Chen, Joshua B. Tenenbaum, and Jonathan Ragan-Kelley, "A Domain-Specific Probabilistic Programming Language for Reasoning About Reasoning (or: A Memo on Memo)," *Proceedings of the ACM on Programming Languages* 9 (OOPSLA2) (2025): 784–814, https://doi.org/10.1145/3763078; and McCoy R. Becker, Alexander K. Lew, Xiaoyan Wang, et al., "Probabilistic Programming with Programmable Variational Inference," *Proceedings of the ACM on Programming Languages* 8 (PLDI) (2024): 2123–2147, https://doi.org/10.1145/3656463.

[11] Lionel Wong, Gabriel Grand, Alexander K. Lew, et al., "From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought," arXiv (2023), https://doi.org/10.48550/arXiv.2306.12672.

[12] Technically this translation process itself is a probabilistic inference. The meaning function parameterized by a large language model encodes a distribution on likely meanings of a sentence in the form of a distribution over expressions in the probabilistic language of thought. See ibid., note 14.

[13] Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz, "Visual Cognition in Multimodal Large Language Models," *Nature Machine Intelligence* 7 (1) (2025): 96–106, https://doi.org/10.1038/s42256-024-00963-y.

[14] 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) (2013): 18327–18332. See also Kevin A. Smith, Jessica B. Hamrick, Adam N. Sanborn, et al., "Intuitive Physics as Probabilistic Inference," in *Bayesian Models of Cognition*, ed. Griffiths, Chater, and Tenenbaum, ref. 10. It might seem strange to posit that the brain contains a physics simulation engine, but Nancy Kanwisher's lab at MIT has used functional neuroimaging to show that indeed it might, in a candidate "physics network" spanning frontal and parietal regions that intriguingly are also involved in action planning and perhaps how we understand others' actions and even novel tool use. Importantly, neither people's judgments nor those of the Bayesian simulation-based models are completely accurate. Because they only approximately and probabilistically infer the underlying physical scene, they show systematic "physical illusions," such as seeing stacks as unstable that in fact are not. The model and people make these errors in the same ways, seeing the same physical illusions, providing further evidence that this is how intuitive physics works (and doesn't always work) in our brains.

[15] Cedegao Zhang, Lionel Wong, Gabriel Grand, and Josh Tenenbaum, "Grounded Physical Language Understanding with Probabilistic Programs and Simulated Worlds," *Proceedings of the Annual Meeting of the Cognitive Science Society* 45 (2023), https://escholarship.org/uc/item/7018f2ss.

[16] Schulze Buschoff, Akata, Bethge, and Schulz, "Visual Cognition in Multimodal Large Language Models."

[17] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum, "Action Understanding as Inverse Planning," *Cognition* 113 (3) (2009): 329–349, https://doi.org/10.1016/j.cognition.2009.07.005; 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) (2016): 589–604, https://doi.org/10.1016/j.tics.2016.05.011; Tan Zhi-Xuan, Jordyn Mann, Tom Silver, et al., "Online Bayesian Goal Inference for Boundedly Rational Planning Agents," in *Advances in Neural Information Processing Systems* 33, ed. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, et al. (Curran Associates, Inc., 2020), 19238–19250, https://proceedings.neurips.cc/paper_files/paper/2020/hash/df3aebc649f9e3b674eeb790a4da224e-Abstract.html; Lance Ying, Katherine M. Collins, Megan Wei, et al., "The Neuro-Symbolic Inverse Planning Engine (NIPE): Modeling Probabilistic Social Inferences from Linguistic Inputs," arXiv (2023), https://doi.org/10.48550/arXiv.2306.14325; Lance Ying, Ryan Truong, Katherine M. Collins, et al., "Language-Informed Synthesis of Rational Agent Models for Grounded Theory-of-Mind Reasoning On-The-Fly," arXiv (2025), https://doi.org/10.48550/arXiv.2506.16755; Pedro A. Tsividis, Joao Loula, Jake Burga, et al., "Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning," arXiv (2021),

https://doi.org/10.48550/arXiv.2107.12544; and Cédric Colas, Tracey Mills, Ben Prystawski, et al., "Language and Experience: A Computational Model of Social Learning in Complex Tasks," arXiv (2025), https://doi.org/10.48550/arXiv.2509.00074.

[18] Tyler Brooke-Wilson, "Bounded Rationality as a Strategy for Cognitive Science" (PhD diss., Massachusetts Institute of Technology, 2023).

[19] Lionel Wong, Katherine M. Collins, Lance Ying, et al., "Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models," arXiv (2025), https://doi.org/10.48550/arXiv.2507.12547.

[20] Based on the Bayesian Tug of War from Noah D. Goodman, Joshua B. Tenenbaum, and Tobias Gerstenberg, "22: Concepts in a Probabilistic Language of Thought," in *The Conceptual Mind: New Directions in the Study of Concepts*, ed. Eric Margolis and Stephen Laurence (MIT Press, 2015), 623–654, https://doi.org/10.7551/mitpress/9383.003.0035.

[21] Wong, Collins, Ying, et al., "Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models."