Trends in
Cognitive Sciences

CellPress

Opinion

From task structures to world models: what do LLMs know?

Ilker Yildirim[1,3,4,5,*] and L.A. Paul[2,4,6,*]

In what sense does a large language model (LLM) have knowledge? We answer by granting LLMs 'instrumental knowledge': knowledge gained by using next-word generation as an instrument. We then ask how instrumental knowledge is related to the ordinary, 'worldly knowledge' exhibited by humans, and explore this question in terms of the degree to which instrumental knowledge can be said to incorporate the structured world models of cognitive science. We discuss ways LLMs could recover degrees of worldly knowledge and suggest that such recovery will be governed by an implicit, resource-rational tradeoff between world models and tasks. Our answer to this question extends beyond the capabilities of a particular AI system and challenges assumptions about the nature of knowledge and intelligence.

Kinds of knowledge

OpenAI's GPT-4 [1] and other LLMs, such as Meta's LlaMA [2], show impressive conversational capabilities. These systems can generate coherent, novel, and often surprisingly sophisticated responses to questions or prompts posed directly in natural language. This has led artificial intelligence (AI) researchers who develop these systems[i], along with cognitive scientists and authors of articles in news media, to ask whether LLMs have knowledge.

Here, we explore this possibility. We suggest we are in a 'Kuhnian moment': a conceptual revolution in what we take knowledge to involve, with implications for how we think intelligence could arise. Accordingly, we ask: in what sense can GPT-4 (and similar models) be said to have knowledge? The answer to this question extends far beyond the capabilities of a particular AI chatbot, with implications for cognitive science, neuroscience, philosophy, and AI.

We ground our answer using a core concept from cognitive science, **world models** (see Glossary). World models are structure-preserving, behaviorally efficacious representations of the entities and processes in the real world [3], including objects with 3D shapes and physical properties [4], scenes with topological relations and navigable surfaces [5], and agents with beliefs and desires [6]. Human thought often relies on these types of world models [7,8] to perceive [4,9], effectively reason [10,11], plan [12], and talk about the world [13].

In particular, for a wide range of ordinary contexts, a knowledgeable human agent draws on their world model, exploiting a structural match between their mental representations and the state of the world and, using language, reliably generating contentful answers to prompts that are approximately truth preserving and relevant. We describe such world-model-based knowledge as **worldly knowledge** and the content of the matched representations that support it as **worldly content**. When a subject has worldly knowledge of a proposition, this is, at least in part, in virtue of the subject using their world model to grasp the worldly content of the proposition. For example, when the subject knows that balancing a ball on a box is easier than balancing a box on a ball, they use their world

Highlights

OpenAI's GPT-4 and similar large language models (LLMs) show impressive conversational capabilities. This Opinion asks: in what sense does a LLM have knowledge? The answer to this question extends beyond the capabilities of a particular AI chatbot and challenges our assumptions about the nature of knowledge and intelligence.

We answer by granting LLMs 'instrumental knowledge': knowledge defined by a certain set of abilities.

How is such knowledge related to the more ordinary, 'worldly' knowledge exhibited by humans? To address this, we turn to a core concept in cognitive science, world models, and explore the degree to which instrumental knowledge might incorporate such structured representations.

We discuss how LLMs could recover degrees of worldly knowledge and suggest that such recovery will be governed by an implicit, resource-rational tradeoff between world models and task demands.

[1]Department of Psychology, Yale University, New Haven, CT, USA
[2]Department of Philosophy, Yale University, New Haven, CT, USA
[3]Department of Statistics and Data Science, Yale University, New Haven, CT, USA
[4]Wu-Tsai Institute, Yale University, New Haven, CT, USA
[5]Foundations of Data Science Institute, Yale University, New Haven, CT, USA
[6]Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, Munich, Germany

*Corresponding author.
ilker.yildirim@yale.edu (I. Yildirim) and la.paul@yale.edu (L.A. Paul).

https://doi.org/10.1016/j.tics.2024.02.008

© 2024 Elsevier Ltd. All rights reserved.

model of how objects move and react to external forces to grasp the content that balancing a ball on a box is easier than balancing a box on a ball. We take such knowledge to be the target of philosophical analyses of how an individual knows the content of a proposition *p*, where to know that *p*, 'an agent must not only have the mental state of believing that *p*, but various further independent conditions must also be met: *p* must be true, the agent's belief in *p* must be justified or well-founded, and so forth' ([14], p. 281). We also take such knowledge to be well studied in scientific contexts [15–20] and to include much of our ordinary factual and relational knowledge about the world.

Recent AI developments in LLMs seem to involve a different kind of knowledge. Such systems are based on deep neural networks pretrained on Internet-scale data to autocomplete the next word (or token, more accurately) given preceding context, which are then further fine-tuned with reinforcement and supervised learning techniques for human-aligned and humanlike responses (e.g., [21]). When using these tools, LLMs can generate sufficiently successful responses to an appropriately wide range of prompts, giving answers that are often approximately truth preserving and relevant. We describe such answers as demonstrating **instrumental knowledge**.

Instrumental knowledge is knowledge acquired through the successful use of instruments that perform certain tasks. A person who can successfully use an instrument like a television remote gains a small amount of instrumental knowledge, as evidenced by the limited range of tasks the remote can be used to perform. (The person gains this instrumental knowledge through task performance, not through knowing in any deeper sense how the television remote works, making their knowledge merely 'instrumental'.) Here, we treat the process of next-word generation as an instrument (cf. [22] for a treatment of language as an instrument). We take an LLM's wide-ranging and generative ability to successfully respond to prompts, using next-word generation, as demonstrating instrumental knowledge. An LLM that can successfully use the instrument of next-word generation gains a significant amount of instrumental knowledge meeting a high standard of reliability, as evidenced by the range of tasks it can use this language tool to successfully perform. Note that the LLM gains its instrumental knowledge through task performance, not through knowing, in any deeper sense, how its responses are about the world.

So, we can grant that LLMs exhibit knowledge, as, like people, they can be said to have instrumental knowledge. However, how is such instrumental knowledge related to the sort of human knowledge that is based on world models and to what degree, if any, might an LLM's knowledge incorporate worldly knowledge?

World models in cognition

We start with an exposition of world models as realizers of worldly knowledge. Our definition of world models is inspired by Gallistel and King's discussion of mental representations [3] and is partially grounded in **mathematical representation theory**. A world model is a way of representing entities in the real world and their relations with two critical requirements.

First, a world model must be structure preserving, such that changes in the real-world entities and their relations should map onto similar sorts of changes in their counterpart representations in the world model. Gallistel and King use the simple example of measuring a child's height by marking it on a wall with a pencil, where the markings and the process of making these markings constitute a structure-preserving representation of the child's height. Notice that a structure-preserving representation is, by definition, content preserving.

Second, a world model must be behaviorally efficacious, meaning that it should enable accurate planning and high-reward actions back in the real world. In the example above, the markings are

Glossary

Causal generative models: a probabilistic model that embeds a world model to specify joint distribution of latents and observables. In this way, inference, prediction, and reasoning queries can be defined rigorously and precisely (via the calculus of probability), so as to formalize processes in perception and cognition.

Compression: in information theory, compression is to efficiently represent data (e.g., in fewer bits, less space).

Fine-tuning with reinforcement learning: typically comprises the use of human feedback about the quality or appropriateness of its generated responses to fine-tune an LLM. If an LLM's response to a prompt is deemed high quality by a human judge, the LLM receives positive feedback, encouraging it to generate similar sorts of responses in the future. This technique alleviates the need for paired datasets of input prompts and output responses needed for supervised training.

Fine-tuning with supervised learning: using a paired dataset of input prompts and output responses to further train an LLM to follow instructions or generate aligned and humanlike responses. Such data may come from humans or from bigger models to train smaller models. Typically, such labeled datasets are much smaller in scale than the dataset used to pretrain LLMs.

Instrumental knowledge: knowledge acquired through the successful use of instruments that perform certain tasks. We suggest that next-word generation in LLMs leads to spontaneous inferences about task structure from natural language input and conditioning of the activations within the model according to this structure.

Machine language translation: a branch of computer science (subfield of natural language processing) that develops methods and algorithms to train machine-learning algorithms to translate across languages.

Mathematical representation theory: in mathematics, representations specify how (sometimes seemingly entirely) different mathematical objects or structures relate to each other. Representation theory is a branch of mathematics that specifies these relationships using linear algebra and transformations of vector spaces.

Structure-preserving mapping: a mapping between two systems is behaviorally efficacious as they can guide decisions about clothing size or comparisons of the heights of a child across time. Notice that behavioral efficacy suggests that world models need not be veridical replicas of what is out there; abstractions that can be implemented in algorithmically efficient approximations of their counterpart real world processes will do. Examples of such structure-preserving, behaviorally efficacious representations are provided in Figure I in Box 1.

Multiple lines of research provide support for world models as a basis of cognition. Probabilistic models in cognitive science suggest concrete candidates for how these structure-preserving, behaviorally efficacious representations may be implemented in the mind (see Figure I in Box 1). Embedding a candidate representation within probabilistic models, often referred to as **causal generative models**, allows researchers, in a given domain to formally and rigorously specify queries of inference, learning, prediction, and planning and to solve them using Bayesian inference. These models are typically evaluated against alternative models in how consistent they are with human performance, with respect to average accuracy, stimulus-driven variability in accuracy, response times, similarity judgments, and other sorts of behavioral ratings.

Beyond evaluating these models in comparison with behavioral performance, recent neuroscientific studies explore the realism of the hypothesized world models in brain activity. A formal review of this literature is beyond the scope of this Opinion, but we provide brief pointers in Box 1. We now turn to the specification of instrumental knowledge in LLMs and situate this knowledge relative to worldly knowledge as characterized using world models.

Instrumental knowledge

How can we characterize the instrumental knowledge of LLMs? We can understand the instrumental knowledge of an entity in terms of its ability to use an instrument to perform tasks posed for it across relevant domains. Indeed, a motivating perspective in LLMs is the idea of **unsupervised multitask learning** [23]. Internet-scale natural language data can be seen as a large dataset of a multitude of tasks posed in varying ways and forms, consistent with the messiness of how language is used naturally. For instance, the abbreviation ‘TL;DR’ or a paragraph that starts with the phrase ‘In summary,…’ might signal a summarization task; nearby or paired sentences or phrases spanning multiple languages might suggest the task of translation between those languages. For a model to accurately predict the next word in a sequence, researchers have speculated that it may be critical for the model to spontaneously infer the **task structure** from the preceding context and condition the next-word predictions on that task structure [23]. After the training of an LLM is over, using next-word generation to infer such task structure from natural language, and conditioning the activations within the model according to this structure, is a possible source of instrumental knowledge (Figure 1A,B).

Could inference and use of task structure occur without (or with very little) worldly knowledge? A setting where this could occur is **machine language translation**. Instead of focusing on building systems that translate through semantic analyzers or any other formal notion of meaning, most progress in machine translation relies on increasingly sophisticated statistical approaches [24,25]. It is plausible that LLMs represent a new frontier in this progression of models, one in which the models infer the task structure of language translation in terms of how words, phrases, and even paragraphs are emitted within and across pairs of languages, and use this structure to translate – without necessarily representing worldly knowledge or rendering it in different languages. Such a possibility is further suggested by ‘relational’ theories of word meaning (e.g., a conceptual role semantics where the meaning of a word or phrase is defined directly by its relation to other words or concepts and only indirectly through reference and causal connections to the nonlinguistic world, with limited transmission of worldly content) [26].

structure preserving (i.e., homomorphic) if how the symbols within each system relate to each other is preserved. A basic example of homomorphism is linearly mappable systems.

Task structure: a broad category of information that facilitates task performance, including noticing the relevant cues and relations in a given prompt that are most important, in this Opinion, for next-word generation.

Worldly content: content constituted by entities and processes in the real world, including objects with 3D shapes and physical properties, scenes with spatial structure and navigable surfaces, and agents.

Worldly knowledge: S has worldly knowledge that P only if P has worldly content and S uses their world model to grasp this content.

World model: structure-preserving, behaviorally efficacious representations of the entities, relations, and processes in the real world. These representations capture, at an abstract level, their counterpart real-world processes (which typically involve causal relations), in algorithmically efficient forms, to support relevant behaviors.

Unsupervised multitask learning: learning of tasks in the absence of any supervision. OpenAI’s GPT-2 was described as an unsupervised multitask learner as a result of learning to accurately predict the next word on Internet-scale language data.

Box 1. Evidence for world models in cognition

We review evidence for the role of world models in cognition across three domains. First, in intuitive physics, several computational studies characterized mental representations as being functionally similar to physics engines of computer graphics (Figure I, top), including a small set of object properties (e.g., shape, mass, stiffness) and a set of simulation rules (often in the form of partial differential equations) [4,55,57,58]. Critically, more recent work tested elements of this computational proposal in neural data. Human fMRI experiments provided evidence for functionally localized frontoparietal brain regions for abstract, invariant representations of object mass [59] and physical stability [60]; electrophysical studies in nonhuman primates provided evidence for continuous inferences about object kinematics and rapid forward predictions about behaviorally relevant collision events [61–63].

Second, spatial knowledge, sometimes referred to as a kind of 'cognitive map', is another commonly explored domain of worldly knowledge. Representational formats in terms of Euclidean maps as well as the more qualitative topological relations (e.g., graphs of connected locations with different connectivity patterns) are used to formalize spatial knowledge [64,65] (and more broadly knowledge involving relations [66,67]; Figure I, middle). Neural studies provide evidence for the deployment of this multiplicity of spatial knowledge representations in the human brain (see [5,68] for reviews), localizing Euclidean maps and more graph-like representations across a network of brain regions including hippocampal, entorhinal, medial–frontal, and occipital regions.

Third, there is model-based planning. A long tradition in cognitive science, neuroscience, and AI formalized how structure-preserving representations can be used for planning (see [69] for a review). Beyond cognitive maps, which can help to support planning in navigation [70], we highlight the 'body model theory' [71], the idea that the anatomy and circuitry of the somatosensory cortex is designed to compute a morphological model of the body and its kinematics ('body simulation'). This proposal constitutes a world model of how our bodies work (Figure I, bottom) and can support complex, contact-rich behaviors [72], the 'embodiment' of others' actions and plans [73], and behavioral rehearsal and learning in the absence of movement [74,75].

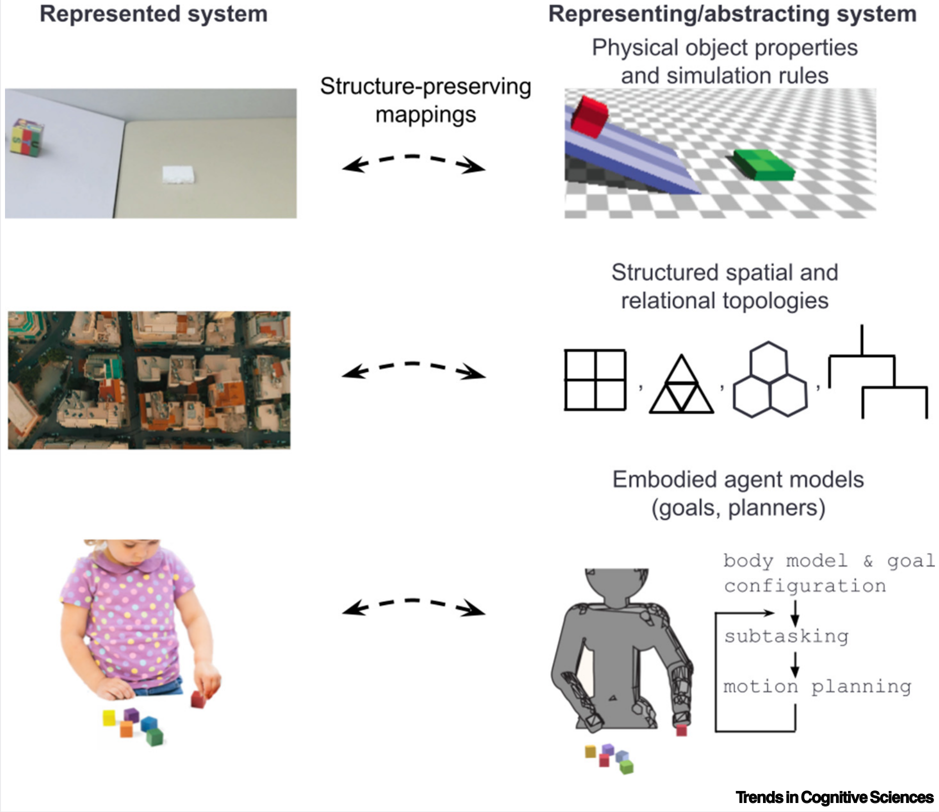
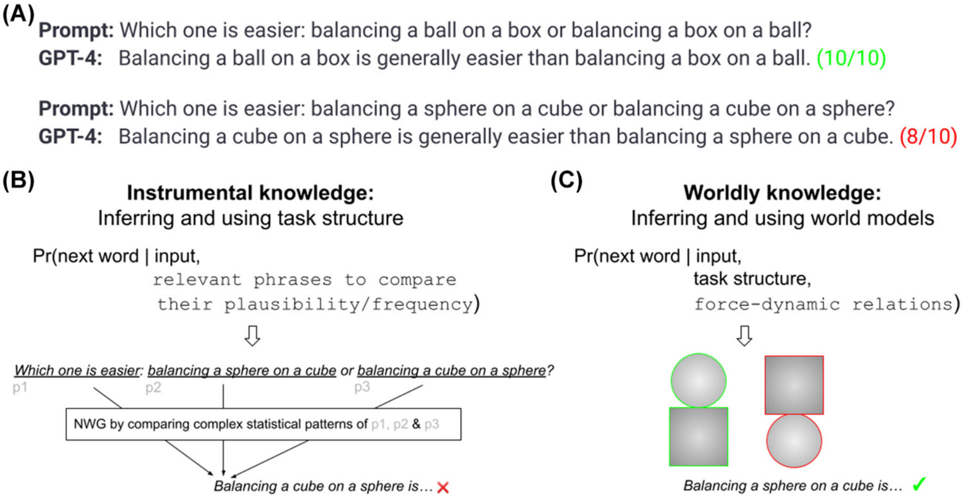

Represented system
Representing/abstracting system
Structure-preserving mappings
Physical object properties and simulation rules
Structured spatial and relational topologies
Embodied agent models (goals, planners)
body model & goal configuration
subtasking
motion planning
Trends in Cognitive Sciences

Figure I. Example world models: structure-preserving, behaviorally efficacious representations. We define world models as behaviorally efficacious, structure-preserving mappings that represent entities and their relations in the real world. Notice that the left-hand side of the figure ('Represented system') does not show a specific input or the right-hand side ('Representing system') a specific output configuration; rather, the two sides are the two ends of a homomorphic mapping between two sets of many configurations [3]; for example: objects with physical properties for intuitive physics (top); places with navigable spatial relations as cognitive maps (middle); and agents with a body morphology and goal-directed dynamics for 'embodied' planning (and social inference; bottom). Joint probability distributions that embed world model configurations and observable inputs (i.e., causal generative models) help to formalize perceptual and cognitive tasks [9] (see Figure 3A in main text).

Could inferring and using task structure be reduced to having a grasp of the rules and patterns of a language? Some have asked whether LLM 'knowledge' is merely an ability to follow linguistic rules and language patterns. In addressing this question, an important distinction concerning the performance of LLMs is between 'form' versus 'meaning'. As others have rightfully cautioned [27], an LLM's ability to generate coherent language should not be taken as evidence of



(A)

Prompt: Which one is easier: balancing a ball on a box or balancing a box on a ball?
GPT-4: Balancing a ball on a box is generally easier than balancing a box on a ball. (10/10)

Prompt: Which one is easier: balancing a sphere on a cube or balancing a cube on a sphere?
GPT-4: Balancing a cube on a sphere is generally easier than balancing a sphere on a cube. (8/10)

(B)
Instrumental knowledge:
Inferring and using task structure
Pr(next word | input,
relevant phrases to compare
their plausibility/frequency)
Which one is easier: balancing a sphere on a cube or balancing a cube on a sphere?
p1
p2
p3
NWG by comparing complex statistical patterns of p1, p2 & p3
Balancing a cube on a sphere is… ✗
(C)
Worldly knowledge:
Inferring and using world models
Pr(next word | input,
task structure,
force-dynamic relations)
Balancing a sphere on a cube is… ✓

Trends in Cognitive Sciences

Figure 1. Exploring the nature of knowledge in large language models (LLMs). (A) LLMs, such as OpenAI's GPT-4, show impressive capabilities on a broad range of tasks (and some failures), that collectively suggest questions about the extent of their worldly knowledge. (GPT-4 was queried ten times on each of these prompts and the most frequent response – shortened for conciseness – is reported, with the frequency displayed in green if that response is consistent with the authors' intuitions and red otherwise. We note that this example is not meant to be a quantitative evaluation; different prompts might lead to different outcomes.) (B) We suggest that pretrained LLMs acquire instrumental knowledge that goes beyond formal linguistics competence (i.e., the set of rules and statistical patterns that constitute language) but is not worldly knowledge. We suggest that this knowledge comes from next-word generation (NWG): accurate autocompletion leads to inference and use of task structure during the processing of input context (and is perhaps further strengthened and organized during the fine-tuning of LLMs). Here, the model might infer that the task is to compare the plausibility of different phrases in the prompt, with respect to the (often highly complex) co-occurrence patterns of these phrases. Such instrumental knowledge might work for comparisons of phrases whose co-occurrence patterns are approximated well by the model but become increasingly inaccurate for less frequent phrases. (C) We explore how much of this instrumental knowledge might rely on an underlying account of the entities and processes in the physical world (i.e., worldly knowledge). In this example, this would include the force-dynamic relations between entities. More generally, in this Opinion, we examine this perspective using the core cognitive concept of structured world models, a review of recent relevant work, and the notion of bounded rationality.

understanding natural language. (In this Opinion, we relate such understanding to having worldly knowledge). Similarly, based on the separation observed in the human brain between language and non-language regions [28,29], researchers have argued [30] that LLMs acquire formal linguistic competence, or knowledge of the rules [30] and statistical regularities [31] of a language, but not 'functional linguistic competence', which includes knowledge of and reference to things and processes in the social and physical worlds [32]. (In this Opinion, we see functional linguistic competence as related to worldly knowledge.)

When making this argument, researchers have sometimes drawn a distinction between pretrained LLMs (on next-word prediction) and LLMs after **fine-tuning with supervised** or **reinforcement learning** objectives on human dialog data (e.g., [21]). When prompted with examples that seem to require worldly knowledge, some fine-tuned LLMs, including GPT-4, generate compelling answers [21]. It is of interest to understand how such fine-tuning impacts knowledge; however, such procedures typically adapt only certain output stages of the pretrained models, otherwise keeping much of the pretrained weights frozen [33,34]. Moreover, in certain cases, performance similar to that of a fine-tuned model can be obtained with pretrained LLMs via so-called 'in-context learning'; in the absence of any parameter updates in the underlying model, providing a few example input–output pairs in the prompt can lead

LLMs to learn new tasks [35]. We see in-context learning as relying on instrumental knowledge that exceeds 'formal linguistic competence' and includes inference of a task structure and conditioning of next-word prediction on that structure (Figure 1A,B).

The leap: from next-word prediction to world models

Beyond inferring task structure, can the instrument of next-word generation also lead to inference of world models? In other words, could this purely text-related instrument lead to any degree of worldly knowledge (Figure 1C)?

The ordinary sense of 'knowing a proposition *p*', where *p* is about the world, includes some degree of worldly content. One interesting way this could occur in an LLM draws on the ability to recover worldly content using **compression**. In many ways, next-word prediction in LLMs reflects compression of the vast amounts of text data crawled on the Internet into the many billions of weights of a deep neural network, which in proportion remains too small to memorize the training data[ii]. Indeed, compression and prediction are closely related objectives [36,37].

A lower-dimensional state space factorizing the relevant dimensions of variation of a given domain and the dependence of these dimensions on each other can be simultaneously used to compress and predict. World models are examples of such state spaces, where a small set of variables captures causal abstractions of the structure of their counterpart physical processes in the real world (see Figure I in Box 1). A caveat, however, is that compression does not always preserve worldly content, as when various possible approximations of the statistical regularities in the training text are not structure-preserving representations.

However, we speculate that it is possible for compression to recover a **structure-preserving mapping** of the data-generating process underlying the training data. There are multiple data-generating processes underlying natural language data that could be the source for this: the rules and statistical regularities in a language, the tasks that are posed and addressed in natural language, and, especially relevant to the present context, worldly content involving entities, physical processes, and situations projected into text by humans perceiving, talking about, and participating in these situations.

We acknowledge that, at present, to hold that LLMs do or even could recover structure-preserving abstractions of the world involves a leap of faith. However, as it is often this type of thesis that motivates attributions of general intelligence to LLMs (e.g., [38]), we turn to available literature exploring this possibility.

Measured recovery of world models in LLMs under domain-specific settings

World models (i.e., causal abstractions of a represented system with a structure-preserving representing system [3]) provide a concrete framework to reason about whether and how LLMs can recover degrees of worldly content that could lead to worldly knowledge. A small number of recent studies have explored whether a language model trained to predict next-token sequence spontaneously approximates the underlying data generating process (i.e., the world model [39,40]). We consider two lines of work: models trained on specialized, non-language domains and models trained on Internet-scale natural language data.

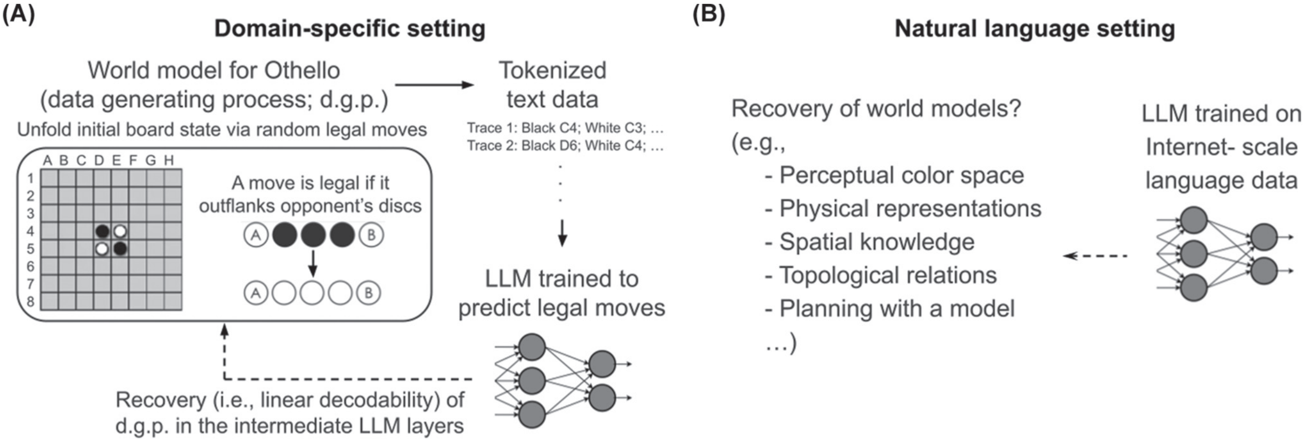
Models trained on specialized non-language domains

A set of recent studies analyzed models after training them on sequences of word-like tokens but in specific non-language domains where the unique tokens, as well as how they combine into sequences, are constrained by the underlying world model.

One example of the recovery of worldly knowledge is presented by Li *et al.* [40] (Figure 2B). This study trained a GPT-4-like (but on a much smaller scale) language model to predict legal moves in the game of Othello – a two-player board game in which players gain more discs by outflanking the discs of the other player. The world model in this game comprises the state of the board (for each cell in an 8 × 8 grid, whether it is black, white, or empty) and the set of rules by which each player changes this state. They randomly simulated this world model creating a dataset of board-state traces via stochastic, non-strategic decisions for each player. When they trained a language model, called Othello-GPT, on this dataset, they found that not only could this model reliably generate legal moves given the prior set of moves, but also the entire board state could be accurately decoded from the intermediate-layer activations in the model, with a linear decoder (as established by a follow-up study [41]). Crucially, the authors also showed that intervening on the board state via these decoders causally and appropriately impacted the model's legal move predictions.

A recent study [39], using a toy domain, took a similar approach to suggest that a language model trained for next-token prediction for program synthesis can recover something about the deeper semantics of this domain-specific programming language. Finally, a similar conclusion comes from the domain of computational biology: a recent LLM trained by Meta researchers [42] to predict masked entries in protein amino acid sequences rendered the coarse 3D structure (i.e., contact relations between amino acids) of the actual folded protein linearly decodable.

The work by Li *et al.* [40] engages with the possibility mentioned earlier: that next-token prediction with a language model can recover structure through the underlying data-generating process. It suggests that the 'leap' mentioned in the previous section is a realistic outcome. It is exciting and of pressing importance for future research to systematically explore this possibility across the dimensions of training objectives (e.g., next-token prediction, masked token prediction), network architectures (e.g., transformer-based language models [43] and RNN-based sequence models [44,45], as these are the kinds of neural network architectures that underlie the modern LLMs and traditional neural language models, respectively), and the complexity of world models.



Trends in Cognitive Sciences

Figure 2. Using world models to explore the extent of worldly knowledge in large language models (LLMs). (B) Compression and prediction are like the two sides of a coin, and it is possible that compression can recover the data-generating process. In a domain-specific setting (the board game Othello) with tokenized traces of randomly generated game states, the work in [40] trained an LLM on next-token prediction. Surprisingly, the intermediate layers of this LLM yielded a linearly decodable full board state. (C) As we discuss in the text, generalization of this result to actual natural-language-trained LLMs is so far limited.

That said, we raise three caveats. First, these studies depend on dense sampling of relatively small domains (more than 1 million training examples for Othello); dataset requirements can quickly grow for non-toy domains, and whether even the largest Internet-scale language datasets can satisfy these conditions is unknown. Second, these studies consider settings where the basic building blocks of the world models are enumerated and assigned unique tokens (e.g., the locations in the 8 × 8 grid and the possible states for each cell in the Othello environment). We suspect this will not be applicable to many relevant world models and to how they are projected into text by human language users. Finally, it is possible that the recovered world model in a language model, based on the objective of next-token prediction, is computationally not the most efficient one.

Models trained on Internet-scale natural language data

This motivates the exploration of the recovery of domain-specific world models in LLMs trained with natural language data (Figure 2C) [46–48], drawing on recent work in the context of perceptual color space [46,47].

The work by Abdou *et al.* [47] shows that pretrained language models can recover aspects of the relational structure of the perceptual color space. Using representational similarity analysis, the researchers report statistically significant correlations between the similarity structure of color pairs with respect to language model embeddings versus Euclidean distances between the same pairs of colors under a well-established perceptual color space. Another study [46] provides further evidence but with a different approach. The researchers report that LLMs provided with a small number of in-context examples from a single hue (e.g., red) generalize, more accurately than chance, to the rest of the color space, indicating that some aspects of the relational structure of the perceptual color space are readily available in these LLMs.

In a way, these results suggest the possibility of recovering worldly knowledge in LLMs despite their purely text-based training. However, this needs to be qualified, as the quantitative nature of the correspondence between the physical spaces and LLM internals is often underwhelming in these studies (e.g., a correlation value of roughly 0.2 in [47]). We anticipate that this correspondence will increase under better trained, larger models; nevertheless, the fact that this relationship is weak in a domain like colors is telling; structure-wise, color space is a simple topology (distances in 3D space) and presumably there is much text in the training corpora that talks about color. The complexity of typical world models projected to text by human language users (e.g., spatial structures, intuitive physics) is often far more complex, suggesting that there is a significant amount of worldly content that still needs to be captured.

Across the settings we reviewed, from Othello-GPT to the case of perceptual color space, LLMs' knowledge can incorporate varying degrees of worldly knowledge, from linearly mappable to rather limited correspondence to the underlying data-generating processes. Next, we speculate on an account of this observed variability, in terms of the complexity of the set of world models underlying a dataset and the tasks a system needs to perform.

A resource-rational approach

What might determine the degree of worldly knowledge recovered due to next-word generation in an LLM? We suggest that recent research in cognitive science offers two key dimensions to consider: the distribution of tasks an agent encounters [49] and the granularity or complexity of representations needed to accomplish these tasks [50–52] (Figure 3A). The combination of task distribution and the coarseness (or the overall complexity) of the structure-preserving representations that this task distribution requires might render the convergence of compression, under next-word generation, to a structure-preserving representation more or less likely (Figure 3B).

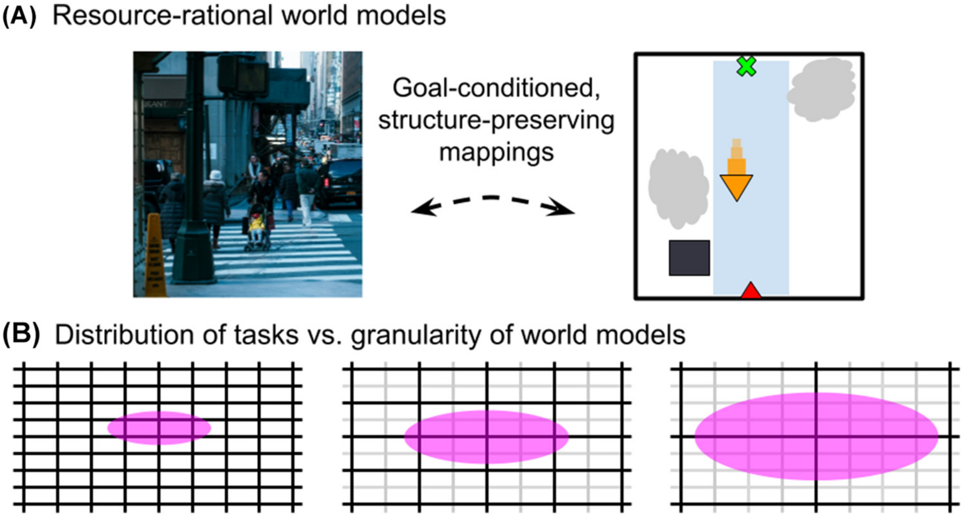

(A) Resource-rational world models
Goal-conditioned, structure-preserving mappings
(B) Distribution of tasks vs. granularity of world models
Trends in Cognitive Sciences

Figure 3. Resource-rational view of world models and instrumental knowledge. (A) A goal-conditioned mapping would invest fewer resources (e.g., larger 'error bars') on the aspects of worldly knowledge not benefiting the task of the agent. Such a mapping can remain behaviorally efficacious and structure preserving. (B) The extent to which next-word generation may recover worldly knowledge might follow from the 'coverage' of the task distribution relative to the underlying data-generating process and the granularity or overall complexity of this data-generating process. We expect recovery of worldly knowledge to be more likely with coarser representations and broader task distributions. The density of the black grids indicates the granularity of the representation; pink regions indicate the subspace of the representations implicated by the task distribution.

Worldly knowledge can be computationally costly; some of the mappings illustrated in Box 1 (e.g., game-engine-style physics simulation, embodied planning), although they are domain specific, may rival LLMs in the complexity of the computations they involve, as they map to complex causal processes in the physical world. For next-word generation to rely on world models (or any other stimulus-computable system), there is the additional challenge of inferring specific configurations of world models from specific input stimulus: going from a sensory measurement to a specific world model configuration, or from a proposition such as balancing a box on a ball is hard to a model in which this proposition is reflected, is an underdetermined problem of an intractable nature.

Accordingly, any system with bounded resources (in computation, time, memory), whether it is an LLM or a human, should take advantage of the efficiencies afforded by the distribution of tasks in their environments [53]. To see an example where tasks can modulate whether coarser worldly knowledge would be sufficient, consider the domain of intuitive physics (see Figure I, top in Box 1): A coarse-grained, qualitative simulation might suffice to predict whether a liquid will flow toward right or left, while a finer-grained simulation may be necessary to determine details of its trajectory [54,55]. Figure 3A illustrates another scenario in the context of a navigation-related task. It is possible that the distribution of tasks an agent encounters may require only coarse-grained representations; alternatively, it is also possible that these tasks require only a specific small subset of the world configurations to be inferred, rendering more idiosyncratic knowledge sufficient.

Instrumental knowledge, inference of task structure, and conditioning of next-word generation on that, without much worldly knowledge, will likely be a sufficient 'shortcut' under a wide range of task distributions. For instance, even in the Othello-GPT study reviewed earlier, when the model is trained on a subspace of tasks based on a distribution of strategic gameplays, as

Box 2. World models as programmable, mid- or high-level interfaces for safety and alignment

Beyond their role in intelligence, why else should we embrace the use of world models in an AI system? That is, if a system has instrumental knowledge allowing it to perform a diverse range of tasks accurately, what else should we want?

We want such systems to be safe. That is, we want these AI systems to be deployed in a way that is both truthful and aligned with 'human values' [76]. Existing deep neural networks, including LLMs with their transformer-based neural network architectures, are black-box systems without an explicit, high-level interface to direct program their behavior. The lack of transparency in how LLMs work, combined with their tendency to produce so-called 'hallucinations' – the frequently observed fabrication of situations, events, and persons by these systems in response to reasonable prompts – has raised concerns in the industry and in larger societal contexts [77,78]. The possibility of unexpected value change in these systems (e.g., transformative higher-order value change as detailed in [79]) raises questions about how sure we can be that evolving AI values will remain aligned with human values [76].

World models may help us find a path to safer, truthful, and better aligned AI systems. Why? Because world models, and more generally domain-specific high-level programming languages, formalize worldly knowledge in structure-preserving, interpretable representations, and they can readily enable truthfulness by supporting an engineer or a user who wishes to impart their 'values' and safety measures as explicit constraints over the system. These features can be exploited via hybrid pipelines of LLMs and world models [80,81][iii], by neuro-symbolic architectures [82], and by the creation of natively programmable neural network architectures [83,84].

opposed to a dataset of randomly unfolding legal game states, the decodability of the game state from the neural network activations (even with nonlinear decoders) is dramatically reduced. Similarly, for the natural language settings discussed (e.g., color space), typical next-word generation queries constrain what knowledge an LLM arrives at.

It is only recently that the field of cognitive science has started exploring computational theories that examine structured world models with task demands (Figure 3A). For example, recent work provides computational-level explanations of simplified representations relative to navigation-related objectives, using the domain of 2D maze navigation [52] and 3D scene perception [56]. This new landscape of resource-rational world models could in turn help to refine what we take knowledge in LLMs and humans to be.

Concluding remarks

The impressive performance of LLMs on a surprisingly broad range of tasks challenges how we think about the acquisition of knowledge, the relationship between instrumental knowledge and worldly knowledge, and, by extension, how intelligence could arise in artificial or machine systems. To frame this challenge, we asked: in what sense can LLMs, trained purely on text, primarily to predict the next word, be said to have knowledge? We answered the challenge by granting 'instrumental knowledge' to LLMs: knowledge defined by a set of (sufficiently sophisticated) abilities to use next-word generation as an instrument, including spontaneous inference and use of task structure. We then asked how such knowledge is related to the more ordinary, 'worldly' knowledge that people exhibit, and explored the degree to which LLM knowledge could incorporate world models. We also suggested two resource-rational frameworks from cognitive science – inference networks and goal-conditioned world models – as promising formalisms for how instrumental and worldly knowledge can overlap or interface. We close by noting that, beyond their implication for intelligence, world models, incorporated more explicitly in AI systems, will facilitate a more direct path to safe and aligned deployment by exposing an interpretable mid- or high-level interface for control and intervention (Box 2; see Outstanding questions).

Acknowledgments

We thank Yutaro Yamada, Mario Belledonne, Jonathan Schaffer, the Editor, and two anonymous referees for helpful discussions. We thank members of the Cognitive and Neural Computation Laboratory at Yale for their feedback on earlier versions of the manuscript. We thank Aalap Shah for his help making display items. I.Y. was supported by AFOSR grant # FA9550-22-1-0041.

Outstanding questions

How can we formalize instrumental knowledge? One possibility is task-conditioned world models, which needs further development. Future work should also explore the relationship of instrumental knowledge to amortized inference or data-driven proposals in Bayesian inference and resource-rational solutions to intractable problems arising from work with expressive world models.

Under what conditions – with respect to the underlying data-generating process, actual training data, and model architecture – does next-word prediction lead to approximate recovery of world models?

What is the impact of fine-tuning on an LLM's instrumental knowledge and recovery of worldly knowledge?

How can we create diagnostic, domain-specific benchmarks to assess the extent and nature of worldly knowledge in LLMs? These benchmarks should explore dimensions such as world model granularity and the training data distribution.

To what extent can the distinction of instrumental knowledge versus worldly knowledge help toward the building and exploration of new foundation models in language and beyond, such as computer vision and reinforcement learning? (LLMs are often referred to as 'foundation models', in the sense of their adaptability to new tasks with little additional data.)

What is the format of the formal linguistic competence internalized by LLMs, specifically in relation to linguistic theories?

How does our discussion relate to existing treatments of LLMs in related contexts, including conceptual role semantics and embodiment-based arguments?

How can we incorporate structured world models in AI, such as LLMs and beyond, for safer and better aligned systems?

Declaration of interests

No interests are declared.

Resources

[i] www.youtube.com/watch?v=SjhIlw3Iffs

[ii] www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

[iii] https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers

References

1. OpenAI (2023) GPT-4 technical report. *arXiv*, Published online March 15, 2023. http://doi.org/10.48550/arxiv.org/abs/2303.08774
2. Touvron, H. *et al.* (2023) LLaMA: open and efficient foundation language models. *arXiv*, Published online February 27, 2023. http://doi.org/10.48550/arxiv.org/abs/2302.13971
3. Gallistel, C.R. and King, A.P. (2011) *Memory and the computational brain: why cognitive science will transform neuroscience*, John Wiley & Sons
4. Yildirim, I. *et al.* (2020) Physical object representations. In *The cognitive neurosciences* (6th edn) (Poeppel, G.M., ed.), pp. 399, MIT Press
5. Epstein, R.A. *et al.* (2017) The cognitive map in humans: spatial navigation and beyond. *Nat. Neurosci.* 20, 1504–1513
6. Jara-Ettinger, J. *et al.* (2016) The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn. Sci.* 20, 589–604
7. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253
8. Spelke, E.S. (2000) Core knowledge. *Am. Psychol.* 55, 1233–1243
9. Kersten, D. and Schrater, P. (2022) Pattern inference theory: a probabilistic approach to vision. In *Perception and the physical world: psychological and philosophical issues in perception* (Heyer, D. and Mausfeld, R., eds), Wiley
10. Gerstenberg, T. *et al.* (2021) A counterfactual simulation model of causal judgments for physical events. *Psychol. Rev.* 128, 936–975
11. Baker, C.L. *et al.* (2009) Action understanding as inverse planning. *Cognition* 113, 329–349
12. Zhu, S. *et al.* (2022) Eye movements reveal spatiotemporal dynamics of visually-informed planning in navigation. *eLife* 11, e73097
13. Jones, C.R. and Bergen, B. (2021) The role of physical inference in pronoun resolution. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, Cognitive Science Society
14. Nagel, J. (2013) Knowledge as a mental state. In *Oxford studies in epistemology* (vol. 4), pp. 273, Oxford University Press
15. Dretske, F. (1981) *Knowledge and the flow of information*. Bradford
16. Chisholm, R.M. (1977) *Theory of knowledge*, Prentice-Hall
17. Goldman, A.I. (1986) *Epistemology and cognition*, Harvard University Press
18. Kornblith, H. (2002) *Knowledge and its place in nature*, Clarendon Press
19. Nagel, J. (2014) *Knowledge: a very short introduction*, Oxford University Press
20. Williamson, T. (2002) *Knowledge and its limits*, Oxford University Press
21. Ouyang, L. *et al.* (2022) Training language models to follow instructions with human feedback. *Adv. Neural Inf. Proces. Syst.* 35, 27730–27744
22. Sosa, E. (2006) Knowledge: instrumental and testimonial. In *The epistemology of testimony* (Lackey, J., ed.), Oxford University Press
23. Radford, A. *et al.* (2019) Language models are unsupervised multitask learners. Technical report, OpenAI. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf
24. Stahlberg, F. (2020) Neural machine translation: a review. *Jair* 69, 343–418
25. Och, F.J. *et al.* (2004) A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 161–168, Association for Computational Linguistics
26. Piantadosi, S.T. and Hill, F. (2022) *Meaning without reference in large language models*, Proceedings of the NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)
27. Bender, E.M. and Koller, A. (2020) Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Association for Computational Linguistics
28. Fedorenko, E. *et al.* (2011) Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 108, 16428–16433
29. Fedorenko, E. and Thompson-Schill, S.L. (2014) Reworking the language network. *Trends Cogn. Sci.* 18, 120–126
30. Chomsky, N. (2014) *Aspects of the theory of syntax, 50th anniversary edition*, MIT Press
31. Bybee, J. and Hopper, P. (2001) Introduction to frequency and the emergence of linguistic structure. In *Frequency and the emergence of linguistic structure. Typological studies in language 45* (Bybee, J. and Hopper, P., eds), pp. 1–24, John Benjamins
32. Clark, H.H. (1996) *Using language*, Cambridge University Press
33. Hu, E.J. *et al.* (2022) *LoRA: low-rank adaptation of large language models*, Proceedings of the Tenth International Conference on Learning Representations
34. Tsimpoukelli, M. *et al.* (2021) Multimodal few-shot learning with frozen language models. *Adv. Neural Inf. Proces. Syst.* 34, 200–212
35. Brown, T. *et al.* (2020) Language models are few-shot learners. *Adv. Neural Inf. Proces. Syst.* 33, 1877–1901
36. Grünwald, P.D. *et al.* (2005) *Advances in minimum description length: theory and applications*, MIT Press
37. Ratsaby, J. (2010) Prediction by compression. *arXiv*, Published online August 30, 2010. http://doi.org/10.48550/arxiv.org/abs/1008.5078
38. Bubeck, S. *et al.* (2023) Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*, Published online March 22, 2023. http://doi.org/10.48550/arxiv.org/abs/2303.12712
39. Jin, C. and Rinard, M. (2023) Evidence of meaning in language models trained on programs. *arXiv*, Published online May 18, 2023. http://doi.org/10.48550/arxiv.org/abs/2305.11169
40. Li, K. *et al.* (2023) Emergent world representations: exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations, 2023*, Published online February 1, 2023. https://openreview.net/forum?id=DeG07_TcZvT
41. Nanda, N. *et al.* (2023) Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: analyzing and interpreting neural networks for NLP* (Belinkov, Y. *et al.*, eds), pp. 16–30, Association for Computational Linguistics
42. Lin, Z. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130
43. Vaswani, A. *et al.* (2017) Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 5998–6008
44. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.* 9, 1735–1780
45. Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.* 14, 179–211
46. Patel, R. and Pavlick, E. (2022) Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations, 2022*, https://openreview.net/pdf?id=gJcEM8sxHK

47. Abdou, M. *et al.* (2021) *Can language models encode perceptual structure without grounding? A case study in color*, Proceedings of the 25th Conference on Computational Natural Language Learning
48. Søgaard, A. (2023) Grounding the vector space of an octopus: word meaning from raw text. *Mind. Mach.* 33, 33–54
49. Dasgupta, I. *et al.* (2020) A theory of learning to infer. *Psychol. Rev.* 127, 412
50. Clark, A. (2015) Radical predictive processing. *South. J. Philos.* 53, 3–27
51. Jara-Ettinger, J. and Rubio-Fernandez, P. (2021) Quantitative mental state attributions in language understanding. *Sci. Adv.* 7, eabj0970
52. Ho, M.K. *et al.* (2022) People construct simplified mental representations to plan. *Nature* 606, 129–136
53. Schaffner, J. *et al.* (2023) Sensory perception relies on fitness-maximizing codes. *Nat. Hum. Behav.* 7, 1135–1151
54. Zhang, Y. *et al.* (2022) Where does the flow go? Humans automatically predict liquid pathing with coarse-grained simulation. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, Cognitive Science Society
55. Bates, C.J. *et al.* (2019) Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Comput. Biol.* 15, e1007210
56. Belledonne, M. *et al.* (2023) Goal-conditioned world models: adaptive computation over multi-granular generative models explains human scene perception. In *Cognitive Computational Neuroscience Conference*, CCN
57. Battaglia, P.W. *et al.* (2013) Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18327–18332
58. Smith, K. *et al.* (2019) Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Adv. Neural Inf. Proces. Syst.* 32, 8983–8993
59. Schwettmann, S. *et al.* (2019) Invariant representations of mass in the human brain. *eLife* 8, e46619
60. Pramod, R.T. *et al.* (2022) Invariant representation of physical stability in the human brain. *eLife* 11, e71736
61. Yoo, S.B.M. *et al.* (2020) The neural basis of predictive pursuit. *Nat. Neurosci.* 23, 252–259
62. Rajalingham, R. *et al.* (2022) Recurrent neural networks with explicit representation of dynamic latent variables can mimic behavioral patterns in a physical inference task. *Nat. Commun.* 13, 5865
63. Rajalingham, R. *et al.* (2022) Dynamic tracking of objects in the macaque dorsomedial frontal cortex. *bioRxiv*, Published online June 28, 2022. https://doi.org/10.1101/2022.06.24.497529
64. Gallistel, C.R. (1990) *The organization of learning. Learning, development, and conceptual change*, MIT Press
65. Warren, W.H. (2019) Non-Euclidean navigation. *J. Exp. Biol.* 222, jeb187971
66. Kemp, C. and Tenenbaum, T.B. (2008) The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10687–10692
67. Behrens, T.E.J. *et al.* (2018) What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* 100, 490–509
68. Peer, M. *et al.* (2021) Structuring knowledge with cognitive maps and cognitive graphs. *Trends Cogn. Sci.* 25, 37–54
69. Mattar, M.G. and Lengyel, M. (2022) Planning in the brain. *Neuron* 110, 914–934
70. Tolman, E.C. (1948) Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208
71. Brecht, M. (2017) The body model theory of somatosensory cortex. *Neuron* 94, 985–992
72. Mordatch, I. *et al.* (2012) Discovery of complex behaviors through contact-invariant optimization. *ACM Trans. Graph.* 31, 1–8
73. Yildirim, I. *et al.* (2017) Physical problem solving: joint planning with symbolic, geometric, and dynamic constraints. *arXiv*, Published online July 25, 2017. http://doi.org/10.48550/arxiv.org/abs/1707.08212
74. Kim, O.A. *et al.* (2022) Motor learning without movement. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2204379119
75. Sheahan, H.R. *et al.* (2018) Imagery of movements immediately following performance allows learning of motor skills that interfere. *Sci. Rep.* 8, 14330
76. Bowman, S.R. (2023) Eight things to know about large language models. *arXiv*, Published online April 2, 2023. http://doi.org/10.48550/arxiv.org/abs/2304.00612
77. Critch, A. and Russell, S. (2023) TASRA: a taxonomy and analysis of societal-scale risks from AI. *arXiv*, Published online June 14, 2023. http://doi.org/10.48550/arXiv.2306.06924
78. Russell, S. (2022) *Provably beneficial artificial intelligence*. Proceedings of the *27th International Conference on Intelligent User Interfaces*, Helsinki, Finland
79. Paul, L.A. (2014) *Transformative experience*, Oxford University Press
80. Wong, L. *et al.* (2023) From word models to world models: translating from natural language to the probabilistic language of thought. *arXiv*, Published online June 23, 2023. https://doi.org/10.48550/arXiv.2306.12672
81. Ellis, K. (2023) *Modeling human-like concept learning with Bayesian inference over natural language*, Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)
82. Lu, X. *et al.* (2022) *NeuroLogic A*esque decoding: constrained text generation with lookahead heuristics*, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)
83. Kim, J.Z. and Bassett, D.S. (2023) A neural machine code and programming framework for the reservoir computer. *Nat. Mach. Intell.* 5, 622–630
84. Lindner, D. *et al.* (2023) *Tracr: compiled transformers as a laboratory for interpretability*, Proceedings of the International Conference on Learning Representations