DOES CHATGPT HAVE A MIND?

Simon Goldstein
The University of Hong Kong
simon.d.goldstein@gmail.com

B.A. Levinstein
University of Illinois at Urbana-Champaign
benlevin@illinois.edu

ABSTRACT

This paper examines the question of whether Large Language Models (LLMs) like ChatGPT possess minds, focusing specifically on whether they have a genuine folk psychology encompassing beliefs, desires, and intentions. We approach this question by investigating two key aspects: internal representations and dispositions to act. First, we survey various philosophical theories of representation, including informational, causal, structural, and teleosemantic accounts, arguing that LLMs satisfy key conditions proposed by each. We draw on recent interpretability research in machine learning to support these claims. Second, we explore whether LLMs exhibit robust dispositions to perform actions, a necessary component of folk psychology. We consider two prominent philosophical traditions, interpretationism and representationalism, to assess LLM action dispositions. While we find evidence suggesting LLMs may satisfy some criteria for having a mind, particularly in game-theoretic environments, we conclude that the data remains inconclusive. Additionally, we reply to several skeptical challenges to LLM folk psychology, including issues of sensory grounding, the "stochastic parrots" argument, and concerns about memorization. Our paper has three main upshots. First, LLMs do have robust internal representations. Second, there is an open question to answer about whether LLMs have robust action dispositions. Third, existing skeptical challenges to LLM representation do not survive philosophical scrutiny.

***K*eywords** Large Language Models (LLMs) · Artificial Intelligence · Folk Psychology · Mental Representation · Interpretability · AI Cognition

1 Introduction

Recent developments in AI are stunning. Large language models like ChatGPT can generate text that is fluent, accurate, and responsive to human questions. These advances have sparked a fundamental question: Do these AI systems possess minds?

To address this broad question, we focus on a more specific aspect of the mental: folk psychology. The key question we explore is whether LLMs have beliefs, desires, and intentions. In other words, do LLMs have goals about what to do, a perspective on what the world is like, and plans for achieving their goals given what the world is like?

Why care about LLM folk psychology? There are at least three reasons. First, folk psychology is the primary lens that humans use to understand the behavior of agents. When we interact with one another, we consistently try to explain what is happening in terms of beliefs and desires (Hutto and Ravenscroft (2021)). As LLMs become increasingly capable, we will increasingly interact with them. It is worth figuring out whether we can genuinely use beliefs and desires to understand these interactions, or whether instead beliefs and desires would be at best an elaborate metaphor. Second, folk psychology is relevant for moral patiency. When we ask what it takes to be the kind of entity that can be harmed, one important answer appeals to the satisfaction of desires (Heathwood (2016)). More may be required too, such as full-fledged consciousness; but folk psychology will play a role. Third, folk psychology is relevant to AI safety. As AI systems become more powerful, many have worried that they may systematically pursue goals that conflict with humanity (Russell (2019)). But much of this discussion implicitly assumes that AI systems will have goals and a perspective about how to achieve those goals. If there are important barriers for AI systems to possess a folk psychology, this may complicate our understanding of what it would take for AI systems to be safe.

Our approach is simple. The question of folk psychology has two key aspects: internal representations and dispositions to act. We'll explore in detail whether LLMs possess each aspect of a folk psychology.

The first key question is whether LLMs have robust internal representations of the world. Here, our strategy in §2 will be to survey various philosophical theories of representation, and see whether LLMs satisfy the theories. We'll look at a range of conditions on representation, including: (i) that the system has internal states that *carry information* about the world; (ii) that these internal states are *causally effective* at producing the system's behavior; (iii) that these internal representation satisfy *folk patterns of reasoning*; (iv) that the *structure* of these internal representations mirrors the structure of what they represent; and (v) that the information-carrying capacity of these representations emerged from some kind of *selective, evolutionary process*. In each case, we'll draw on recent research in machine learning about LLM *interpretability*, which suggests that these systems satisfy the relevant condition on mental representation.

But internal representations alone are not sufficient for a full-blown folk psychology. In order to possess beliefs and desires, the second key question is whether LLMs have robust dispositions to perform actions. If LLMs have both internal representations and action dispositions, then they will have a folk psychology. In particular, nearly every theory of belief and desire will explain these mental states in terms of some combination of internal representations and dispositions to act. In §4, we explore two of the most prominent traditions of theorizing about belief and desire, interpretationism and representationalism. In each case, we suss out what kinds of action dispositions the theory requires of LLMs. Then we critically assess whether LLMs satisfy the relevant condition. The crucial question for LLMs will be whether their linguistic outputs are *stable* enough to be best explained as promoting goals. Our conclusion in §4 will be tentative. We argue that the behavior of LLMs in game environments is suggestive of the kinds of rich plans of action required for belief/desire psychology. But the data is not decisive.

Our third goal of the paper, in §3, will be to refute some of the existing skeptical challenges to LLM folk psychology. Here, we'll engage directly with three challenges. The first challenge, *symbol grounding*, raises the following question: if the only inputs to a language model are strings of text, rather than rich perceptual experiences or feedback from motor actions, how can language models understand prompts about the external world? The second challenge is that LLMs do not represent the world because they are not trained to do so; instead, they are merely *stochastic parrots*, trained to predict the next word. The third challenge is that LLMs do not represent the world because their behavior can be fully explained by an alternative theory, according to which they rely on *memorization* and shallow shortcuts. In each case, we'll argue that the challenge does not survive philosophical scrutiny.

Overall, then, our conclusions for future research run as follows. First, we think there are interesting questions about exactly how LLMs represent the world, and what in the world they represent. But we think overall there is quite strong evidence that they do so. Second, there is a rich debate to be had about whether LLMs genuinely have beliefs and desires about the world, in the sense that is involved in assembling complex plans of action. Third, we suggest that many of the existing skeptical challenges to LLMs deserve more careful development, as they possess major shortcomings. Before we turn to our main claims, we'll end this section by briefly summarizing how LLMs work.

1.1 Large Language Models

At the heart of modern LLMs lies the transformer architecture (Vaswani et al. (2023)). Although many variants exist, we'll here focus on the mechanics of decoder-only models such as GPT 3.5.[1]

When you feed a prompt to a model, it makes a prediction about what comes next. For example, if you feed it "The cat sat on the," it will assign a probability to each possible next token.[2] It will assign some probability to "aardvark," some probability to "banana," some probability to "mat," and so on.

To compute these probabilities, each token is converted into an initial *embedding*—a vector, or long list of numbers, which encode "The," "cat," "sat," "on," and "the." The initial embedding carries information about the corresponding token, but it doesn't at first carry any information about surrounding tokens. For example, the initial embedding for "cat" does not encode the fact that "The" precedes it.

The initial embeddings for each token are then transformed and updated across a large number of layers. At each layer, the embedding for a given token is first updated through the mechanism of self-attention. Self-attention allows the embedding for a given token to "attend to" itself and earlier tokens in the sequence. For example, the token for "sat" might attend to "cat" at a given layer. The embedding for "sat" could then be updated to represent the fact that "cat" was the immediately preceding token or, perhaps, to encode somehow that "cat" was the subject. In other words, after paying attention to the information of earlier tokens and itself, the embedding for "cat" is updated to include contextual information about the surrounding tokens in the prompt. We call this new embedding a *contextual embedding*. These new embeddings are then refined further using something like a multi-layer perceptrons (MLPs) before being fed forward into a new layer.

After the embeddings are passed through all the layers of the model, the model uses the final contextual embedding for a token to predict what the next token will be. To generate more and more text, we can then select some token assigned relatively high probability, tack it onto the initial prompt, and then feed the new augmented sequence to the model again. For an illustration, see fig. 1.

Transformer models get extremely good at generating plausible text via training. Training for retail models like ChatGPT comes in two main phases. In the first phase, we take lots of pre-existing text, feed an initial segment of it to the model, and then have the model generate probabilities about what comes next. We then tweak the parameters of the model via gradient descent to make more accurate predictions. For instance, if the model is fed "Happy families are all alike; every

[1] More advanced models have many architectural changes, but the exact details are not public.

[2] Tokens can be words, subwords, numerals, punctuation, etc. For our purposes, we can just think of tokens as words.

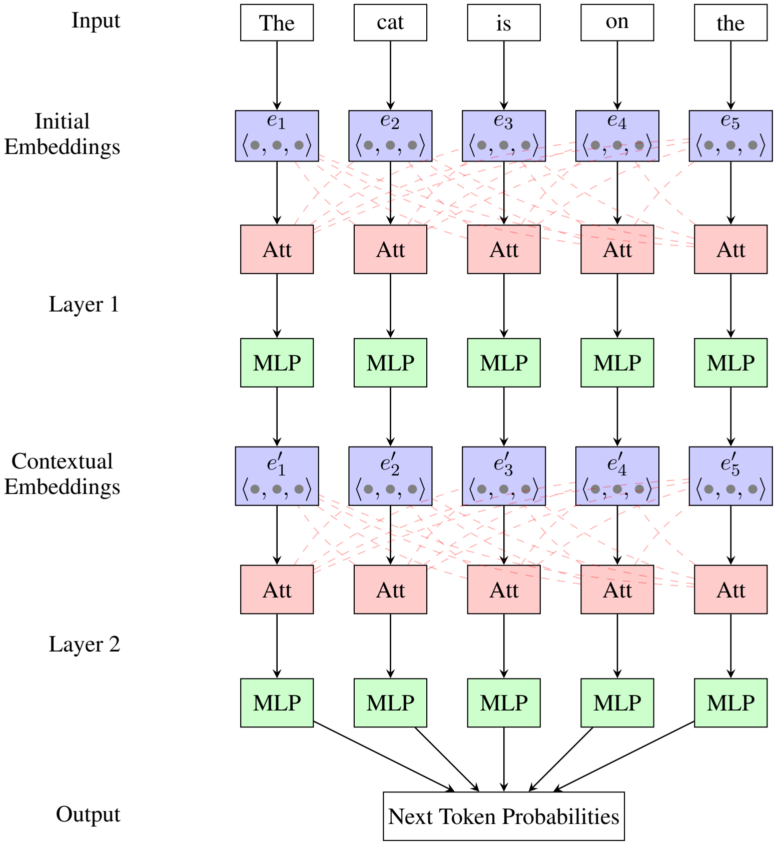


Figure 1: Simplified two-layer transformer architecture processing “The cat is on the”. Each word is initially converted to an embedding vector. In each layer, self-attention (Att) allows words to attend to each other, followed by a multi-layer perceptron (MLP). After the first layer, new contextual embeddings are created. The final layer produces probabilities for the next token.

unhappy family is unhappy in its own” it will learn to generate a higher probability for “way” as the next token. The model’s parameters are adjusted through gradient descent to improve its predictions over time.

Eventually, the model gets very good at predicting what comes next in text. But it is still not especially conversational or useful. So, the second (more optional) phase of training, called “fine-tuning,” involves tweaking the model further to be a better conversational agent. We will omit the details of this portion of training here, but the essential idea is to get humans or other AI to rate various responses for quality and then push the model to be more disposed to generate high quality responses.[3]

Note, at this point, some initial obvious reasons for skepticism about LLMs. LLMs are trained, essentially, to make good predictions about sequences of text and then to tell people what they want to hear. There is no direct pressure to represent the world nor to represent truth. Instead, the immediate pressure on LLMs is to find plausible or pleasant continuations of prompts. Furthermore, pure LLMs only connect with the world via textual embeddings. They have some

[3] The most common method is reinforcement learning from human feedback (RLHF) (Christiano et al. 2023), but there are a number of alternatives such as reinforcement learning from AI feedback and direct preference optimization.

initial embedding for words like “rainbow” or “hammer,” and they manipulate these embeddings based on context, but they’ve never actually seen a rainbow nor manipulated a hammer.

There is, then, an important question of whether LLMs can or do end up representing the world as a means toward their training objective. While they manipulate embeddings based on context, their lack of direct sensory experience and the focus on plausibility rather than accuracy raise reasons for skepticism about their representational capabilities.

2 Theories of representation

Can LLMs represent the world? In this section, we explore this question by examining various philosophical theories of representation. We aim to demonstrate that LLMs meet many of the criteria set by these theories. In particular, we’ll highlight a series of recent studies in AI interpretability research and related behavioral research that connect closely to various philosophical theories of representation.

We focus primarily on naturalistic theories of representation. These theories explain representation in terms of physical processes. They differ from other theories, such as those proposing that representation is primitive (Boghossian (1990)) or dependent on primitive phenomenal properties (Graham et al. (2007)). While non-naturalist theories may allow for LLM representation, it is difficult to say whether AIs could have primitive representational or phenomenal properties.

Our focus is on whether LLMs have internal states that represent the world by having truth conditions. A representation can truly depict the world if the world is a certain way and falsely if it is another. We will also consider if LLMs have internal states that refer to objects or properties in the world.

Importantly, our main interest is not the text produced by LLMs but whether LLMs have mental states that represent the world. Our hypothesis is that the *activations* of LLMs—the patterns of internal neural activity within the network as it processes information—refer to objects in the world and have truth conditions. These activations, which can be thought of as the temporary, computation-specific states of the network, are distinct from the more permanent weights that encode the model’s learned knowledge. This question of representation is an important first step in determining whether LLMs have a robust folk psychology with beliefs and desires.

With these questions in mind, we will survey leading theories of mental representation to see if LLMs meet their criteria. We draw on existing surveys, including Adams and Aizawa (2021) and Schulte (2023). Our goal is to show that, according to these leading theories, there is strong evidence that LLMs indeed represent the world.

We will consider five key conditions on mental representations, and argue that LLMs satisfy each one:

- **Information carrying**: Informational theories posit that mental representation requires internal states to *carry information* about the external world, typically through probabilistic connections. We demonstrate how recent advances in AI interpretability, particularly in *probing* techniques, provide compelling evidence that LLM internal embeddings carry such world-relevant information.

- **Causal efficacy**: Fodorian theories demand that representational states be *causally effective* in generating system behavior. We present evidence from recent interpretability studies showing that LLM outputs indeed depend counterfactually on their internal embeddings, satisfying this causal requirement.

- **Folk-psychological reasoning**: Another Fodorian condition requires that reasoning with internal representations follows patterns familiar from folk psychology. We argue that emerging research on *world models* in LLMs reveals their capacity for effective reasoning about the world.
- **Structural isomorphism**: Structural theories of representation require that the internal architecture of representations mirrors the structure of what they represent. We illustrate how recent studies on LLM concepts of color and direction demonstrate that their representations exhibit the requisite structural properties.
- **Selection**: Teleosemantic theories stipulate that genuine representations must emerge from a selective, evolution-like process. We contend that the training methodologies employed in developing LLMs fulfill this selectional criterion.

2.1 Information

A long tradition of work on representation has appealed to the concept of carrying information. Smoke carries information about fire, mumps carry information about measles, and thermometers carry information about temperature. Dretske (1981) and others have argued that a system can only represent the world if that system carries information about it.

Philosophers have disagreed about how exactly to define carrying information, but in general different analyses all appeal to probabilistic concepts. According to Dretske, a state carries the information that p if and only if the probability of p given the state is 1, provided that various background conditions obtain. Some theorists have instead focused on states raising the probability of p, rather than making it certain (Usher (2001)). Other theorists have focused on more general conditions involving entropy.[4]

How can we tell whether LLMs carry information about the world? Harding (forthcoming) argues there is a close connection between informational approaches to representation and recent work on probing in AI interpretability research (Alain and Bengio (2018)). In probing, researchers train a separate classifier to take activations as input and make a prediction. The probe takes in some activations and predicts features of the input. For example, in a visual AI system, a probe might predict whether the system is looking at a cat based solely on activations, without access to the original input.[5]

A compelling example of using probes to discover LLM representation comes from Li et al. (2022). We'll use this example as a case study below. Li et al trained an LLM on sequences of moves in the 8x8 board game Othello, using only lists of moves (like F5 D6 C3 D3 C4 F4 E3) without describing the rules or the board. Othello is a simpler game than chess, but there are far too many possible moves in general for an LLM simply to memorize all legal game states. Despite this, the trained LLM, dubbed Othello-GPT, tended to output legal moves with high probability. To understand how, the authors used probes to identify possible internal representations of the board. By comparing the model's internal states to the actual board, they trained probes to guess whether each of the sixty-four squares was black, white, or blank. The probes achieved remarkable accuracy with an error rate of only 1.7%. Similar probing techniques have been used

[4] The exact connection between carrying information and representing the world is also debated. For Dretske, a mental state represents that p iff the state carries the information that p during the end of the subject's learning period associated with the state. This doesn't require that whenever a state represents that p, it carries the information that p; for Dretske, carrying information is factive, and so this would rule out misrepresentation. But other notions of carrying information might not be factive, and so could allow a more direct connection between representation and information.

[5] Probes typically use linear classifiers or shallow neural networks trained on model activations to predict specific features. For details, see (Alain and Bengio 2018).

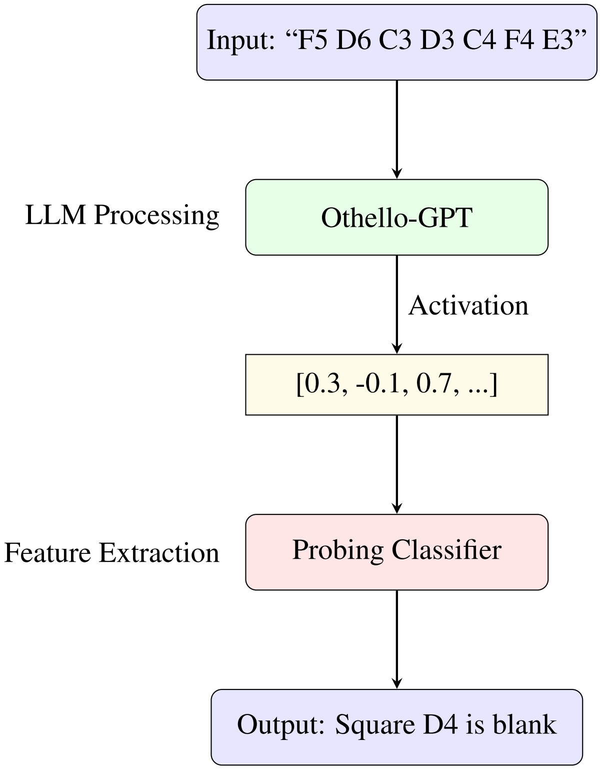


Figure 2: Illustration of the probing process using Othello-GPT. The LLM processes the input sequence of Othello moves, generating activations. A separate probing classifier is trained to predict specific features (e.g., the state of a particular square) from these activations.

to understand how models represent grammatical case, number, tense, and more (e.g., Giulianelli et al. (2021)). Figure 2 illustrates the probing process.

Informational theories of representation make sense of the relevance of probes for mental representation. If the classifier can correctly determine whether a cat is present almost 100% of the time, and the classifier is only using the activations of the system to make its prediction, then those activations likely represent the cat. At the very least, they carry mutual information with the presence of cats. If mental representation is a matter of carrying information, then there is no special barrier to LLMs representing the world.

2.2 Causal powers

Fodor (1975) imposed a series of conditions on genuine representations, one of which is that they must have genuine causal powers. The key question here is whether LLM outputs are robustly caused by the activations discovered by probes or whether these apparent representations are actually epiphenomenal.

To make this concern concrete, return to Othello-GPT. Suppose Othello-GPT itself does not represent the state of the board at all. However, a clever probe could read off from its activations alone what the history of moves was. For instance, if F5 D6 C3 D3 C4 F4 E3 were fed to Othello-GPT, the probe could learn just from its internal state that F5 D6 C3 D3 C4 F4 E3 was the input. If the probe also learned (via its own training) what the rules of Othello were, it could determine whether each square was black, white, or blank, even if Othello-GPT itself didn't represent that fact.

Alternatively, we can imagine that Othello-GPT does represent the board state, but the probe found some other way to determine whether its square was black, white, or blank using information from Othello-GPT's activations that Othello-GPT itself does not use. To investigate this question, interpretability researchers use methods of causally intervening on a model's activations to test whether supposed representations are actually relevant or useful to the model.[6] For instance, if we find a successful probe for a square in Othello-GPT, we can determine what elements of Othello-GPT's activations the probe is using to identify whether a square is black, white, or blank. We can then hand-edit the activation to see what happens.

Suppose, in an oversimplified example, that the probe notices that when the first coordinate of an embedding vector for a square is 1, the square is black; when it is 0, it is blank; and when it is -1, it is white. We can feed the model a prompt making the square black, setting the first coordinate of the relevant embedding vector to 1. Then, we can manually change this coordinate to 0 or -1. If the model's output changes as expected, we have strong evidence that the representation we found is one the model itself uses to track the board state. Figure 3 illustrates the process of causal intervention in the Othello-GPT model. This demonstrates how altering specific internal representations can change the model's outputs, providing evidence for the causal efficacy of these representations.

As another example, consider finding a potential representation of grammatical number in an LLM. Suppose we feed the LLM the prompt 'I ate these'. Without hand-editing, 'apples' should receive a higher probability than 'apple' as the next token. But if we hand-edit the representation of 'these' from plural to singular, we can see if 'apple' becomes more likely than 'apples'. If so, we have causal evidence that our supposed representation is used by the model.

In the case of Othello-GPT, causal interventions were effective showing that the probes did in fact find the model's representation of the board's state. In other cases, similar causal methods have been used to find representations of grammatical number, subjecthood, and factual associations (Giulianelli et al. (2021), Meng et al. (2023)). We thus have direct evidence of the causal effectiveness of LLM representations.

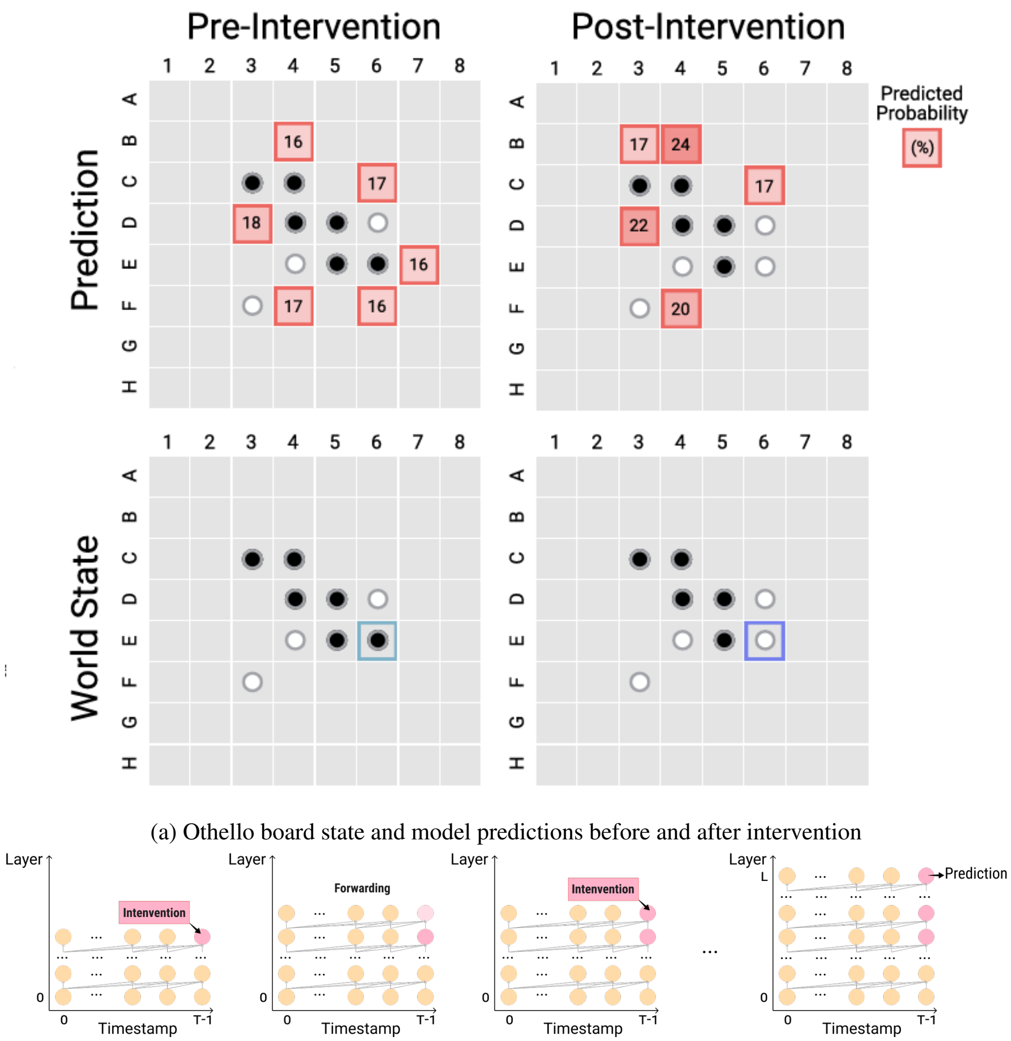
2.3 Folk patterns of reasoning

So far, we've argued that LLM activations carry information and causally influence LLM outputs. But this alone may not be enough for genuine representation. For example, Fodor (1975) argued that genuine mental representations have to have special kinds of causal powers: the representations have to influence the system's behavior in ways that match the laws of folk

[6] There are many different methods of intervention to understand and manipulate neural network representations. The simplest is ablation, where entire neurons are deactivated to assess their importance in the model's performance. More sophisticated approaches include:

- Iterated Nullspace Projection (Ravfogel et al. (2020)): This method iteratively projects embeddings into a nullspace to remove the influence of specific concepts, effectively isolating and erasing targeted information.
- Least Squares Concept Erasure (LEACE) (Belrose et al. (2023)): LEACE identifies and removes concept-specific information from embeddings using a least squares optimization approach, ensuring that the targeted concept is no longer represented in the neural activations.
- Causal Tracing (Meng et al. (2023)): This method tracks the flow of information through a model to determine the causal impact of specific components or representations. By intervening in the causal pathways, researchers can assess the importance and role of particular features in the model's decision-making process.

These techniques allow researchers to surgically target and manipulate information encoded in embeddings, providing deeper insights into the functioning and interpretability of neural networks.



(a) Othello board state and model predictions before and after intervention

(b) Activation patching process across model layers

Figure 3: Activation patching in Othello-GPT. (a) The top panel shows the Othello board state and model predictions before and after intervention. The upper row in each state displays the model's move predictions with associated probabilities, while the lower row shows the actual board state. Pre-intervention, the model correctly predicts legal moves. Post-intervention, the model's predictions change. (b) The bottom panel illustrates the process of activation patching across different layers and timestamps of the model. The intervention at a specific layer and timestamp propagates through subsequent layers, ultimately affecting the final prediction. This demonstration shows how altering internal representations can causally influence the model's outputs, even leading to illegal move predictions in the context of the game state shown above. Both depictions are adapted from Li et al. (2022).

psychology. The representations have to make sense of the world, causing outputs in a way that resembles ordinary reasoning. Given our interest in whether LLMs have folk psychological states, this theory of representation is especially important for our purposes.

One way interpretability researchers have investigated this kind of condition is through the idea of world models. A world model, in the context of AI, refers to an internal representation of how the external world operates, including objects, their properties, and the causal relationships between them. World models require coherent and consistent representations across contexts and a degree of abstraction that allows LLMs to generalize from particular cases to more general situations. For LLMs, these models are constructed purely from textual data, raising important questions about their nature and limitations.

The importance of world models lies in their potential to bridge the gap between mere pattern recognition and genuine understanding. If LLMs possess robust world models, it suggests they have developed representations that go beyond simple statistical associations, potentially supporting attributions of beliefs and reasoning capabilities.

Li et al. (2022) suggest their experiments with Othello-GPT are strong evidence of LLMs' ability to create world models, since Othello-GPT has some coherent internal model of the board state that allows it to predict the next move. Indeed, Othello-GPT can easily generalize to make predictions about new boards its never seen before and can even model permissible moves in impossible boards—i.e., boards with states that can never be reached through a legal initial series of moves. Recent work by Vafa et al. (2024) builds upon and extends this approach, proposing new evaluation metrics for world model recovery inspired by the Myhill-Nerode theorem from language theory. Their study encompasses not only game environments like Othello but also navigation tasks and logic puzzles that can be captured by a finite state automaton. Their research demonstrated that while language models can perform well on existing diagnostics, their underlying internal models vary in levels of coherence.

However, the rules of Othello and those studied by Vafa et al. (2024) are essentially a kind of language that renders some "moves" grammatical and others not. It's not clear that modeling Othello is sufficient evidence of LLMs' having causal abstractions of physical processes that are robust enough to count as world models.

Musker and Pavlick (2023) explore whether large language models build causal models in order to understand the meaning of words. They rely on the HIPE theory, developed to understand human lexical concepts, in connection with artifact terms like "mop" and "pencil" (Chaigneau et al. (2004)). According to this theory, when human language users decide whether an artifact term like "mop" can correctly apply to an object, they rely on an implicit causal model:

> The object's design history and the user's goal are distal causes in the CM, while the object's physical structure and the user's actions with respect to it are proximal causes in the CM. Thus, HIPE predicts that, for example, both the physical structure of an object (e.g., having a handle and something absorbent on one end) as well as the reason the object was originally created (e.g., for wiping up water) should affect how appropriate it is to call the object a 'mop', but that the latter should have a minimal effect when the former is fully specified (Musker and Pavlick 2023).

The idea is that the status of later nodes 'screen off' earlier nodes when judging whether something is an artifact. To test this theory, Chaigneau et al. (2004) built vignettes that manipulated various nodes in the causal model, and explored which of the nodes influenced language users' judgments about whether an object counted as a type of artifact. For example, one such vignette was:

> One day Jane wanted to wipe up a water spill on the kitchen floor, but she didn't have anything to do it with. So she decided to make something. [...] The object consisted of a bundle of thick cloth attached to a 4-foot long stick. Later that day, John was looking for something to wipe up a water spill on the kitchen floor. [...] He grabbed the object with the bundle of thick cloth pointing downward and pressed it against the water spill (Musker and Pavlick (2023)).

Musker and Pavlick (2023) applied this theory of lexical concepts to large language models. Musker and Pavlick found that GPT-4 behaved very similarly to human subjects when evaluating counterfactual vignettes. Compromising any factor in the model led to a negative effect on GPT-4 deciding that something counted as an artifact. Compromising proximal features had a larger effect on GPT-4's judgments than compromising distal features.[7] This potentially suggests that GPT-4 built causal models in order to deploy artifactual lexical concepts, in a way structurally similar to humans.[8]

Recent work by Gurnee and Tegmark (2024) provides compelling evidence that large language models (LLMs) develop coherent representations of space and time, even when trained solely on text data. By probing the internal activations of Llama-2 models, they discovered that LLMs learn linear representations of spatial and temporal information across multiple scales. These representations are unified across different entity types (e.g., cities and landmarks) and robust to variations in prompting. Remarkably, they identified individual "space neurons" and "time neurons" that reliably encode spatial and temporal coordinates. Figure 4 provides striking visual evidence of how LLMs develop structured representations of space and time. The clear organization of locations and events in the model's internal space suggests that these representations go beyond mere statistical associations.

Another skeptical challenge to LLM reasoning concerns poor performance in LLMs like GPT 3.5 when such models are asked to perform arithmetic operations. In order to genuinely represent mathematical information on a Fodorian view, the LLMs would need a series of representations that transform in law-like ways that reflect folk patterns of reasoning. Instead of genuinely representing mathematical information, the skeptics argue, these models are instead engaged in simplistic statistical pattern matching or simply memorize their training data.

Nanda et al. (2023) explored internal computations in a LLM trained to perform modular addition. They found that rather than merely memorizing answers or relying on statistical patterns, the LLM implemented a particular 'Fourier multiplication algorithm' for computing the sum: simply put, "they perform this task by mapping the inputs onto a circle and performing addition on the circle."[9]

[7] For example, in one vignette (labeled "pencil object, compromised action scenario") an object is designed to be used as a pencil (satisfying the goal condition), but the user fails to successfully use the object to write on a piece of paper (violating the action condition).

[8] For more on the emergence of causal models in LLMs, see Forbes et al. (2019), Da and Kasai (2019), Ettinger (2020), Petroni et al. (2019), and Kassner and Schütze (2020). Musker and Pavlick themselves shy away from interpreting these results as showing that GPT-4 genuinely builds causal models of artifacts (p. 7). Instead, they suggest that more work is needed to identify methods for probing inner models. After all, their methodology relies solely on GPT-4's responses to text vignettes. But we have already seen that other work in interpretability research has used probing and other paradigms to uncover world models in the internal representations of LLMs. In this setting, even the behavioral evidence from Musker and Pavlick can potentially be interpreted as revealing inner causal models associated with LLM lexical concepts. See Yildirim and Paul (2024) for further work exploring the philosophical upshots of world models in LLMs.

[9] See Zhong et al. (2023) for an alternative algorithm some LLMs learned to perform modular arithmetic.

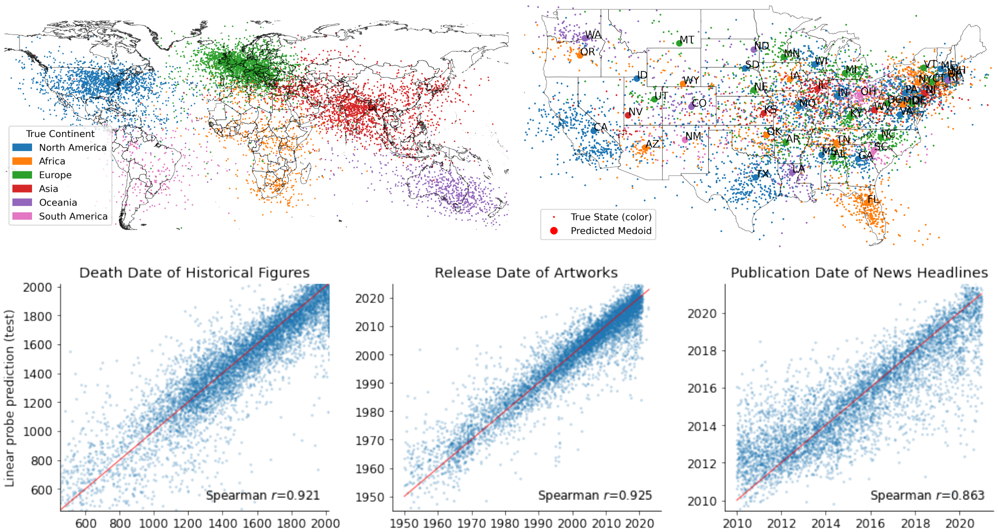


Figure 4: Spatial and temporal representations in Llama-2-70b. Each point corresponds to the layer 50 activations of the last token of a place (top) or event (bottom) projected onto a learned linear probe direction. The clear structure in these projections, closely matching real-world geography and chronology, demonstrates that the model has learned coherent representations of space and time. All points depicted are from the test set. (Adapted from (Gurnee and Tegmark 2024).)

During training, the models initially overfitted the data by memorizing specific examples. However, over time, they transitioned to a generalizable understanding of modular addition, demonstrating an internal shift from memorization to systematic reasoning.[10]

2.4 Structure

Structuralist theories of mental representation propose that mental states represent the world only if their internal structure mirrors the structure of the external world (Opie and O'Brien (2004)). This mirroring creates a network of relationships within the mental states that correspond to the relationships between the objects or properties they represent. As an analogy, maps represent geographic areas by containing a series of symbols whose physical relations on the page mirror the physical distances between locations.

Structuralist theories of mental content are particularly relevant to large language models. Interpretability researchers have explored the activations of LLMs to see whether they stand in patterns of relations with one another that are isomorphic to various worldly features.

Patel and Pavlick (2022) tested GPT-3's concepts of direction and color. In the case of cardinal directions, they exposed GPT-3 to a series of gridworlds, consisting of matrices filled with 0s and a single 1. The task was to identify the direction of the 1 symbol in the gridworld: left,

[10] There are many more questions that can be asked about the extent to which LLM representations satisfy various folk reasoning patterns. Recent discussion, for example, has highlighted the 'reversal curse', where LLMs are competent with the phrase A is B, while failing to understand the phrase B is A (Berglund et al. (2024)). There has been rich recent discussion about these sorts of issues, caught up with the question of whether LLM representations exhibit the kind of full-blown compositional structure familiar from Fodor (1975). (For example, see Lake and Baroni (2023) for a recent argument that LLMs do exhibit full compositional structure.)

right, etc. They prompted GPT-3 with a series of sample answers, and tested whether GPT-3 could complete the pattern. Patel and Pavlick found that GPT-3 had a strong grasp of cardinal directions. First, the LLM could smoothly generalize to a range of gridworld environments: it could correctly describe directions in gridworlds with new lengths and widths. Second, the LLM could smoothly generalize across cardinal directions: even when it was only prompted with northerly or easterly directions, it could correctly answer questions about southerly and westerly located 1s. This showed that GPT-3 had an underlying conceptual space of directions that connected the concepts of north to south, and west to east. Once the LLM could hook up one of these concepts to the gridworld, its underlying structural understanding of direction allowed it to complete the grounding task. Finally, the LLM smoothly generalized to rotated gridworlds. Even when the grids did not map onto our ordinary judgments of left and right, the LLM could quickly generalize from examples. This suggested that the LLM had a stable underlying conceptual space for directions.

Patel and Pavlick ran a similar experiment for color. In this case, they tested whether LLMs could few-shot learn how to map folk color terms to three dimensional RGB scales. Again, although GPT-3 is a text-only model, it nonetheless possessed an underlying conceptual space for color. Even when it was only prompted with examples that associated RGB values with red colors, it could still use this information to correctly predict how to associate RGB values with blue colors. This showed that the LLM had an underlying grasp of the structural relation between red and blue colors: once it learned how to connect red colors to RGB values, the further connection to blue colors could be inferred.

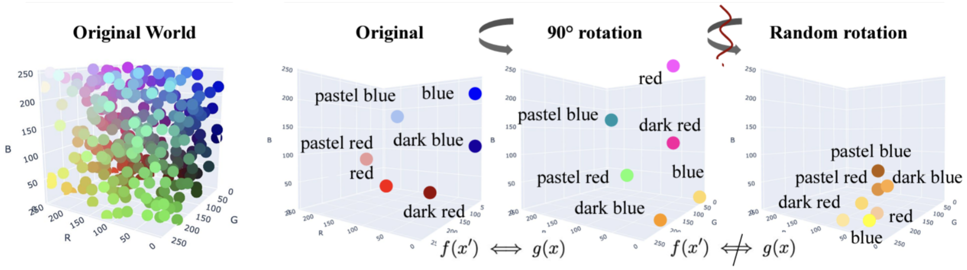
In both the case of direction and color, Patel and Pavlick's experiment demonstrated that LLMs develop rich networks of representations, with structural relationships that resemble the analogous human concepts. This result fits related interpretability research from Abdou et al. (2021), which found that text-only language models create a network of embeddings for color terms. These embeddings were found to have structural relationships that mirrored human judgments of similarity between various colors: they found that language model "representations of color terms that are derived from text only express a degree of isomorphism to the structure of humans' perceptual color space". For a visual illustration, see fig. 5.

Indeed, the nature of the embedding space allows for a natural representation of many different structural relationships. Word embeddings are dense vector representations of words, where words with similar meanings are located close to each other in the vector space. This spatial arrangement allows LLMs to recognize and generate analogies by identifying patterns and relationships between different word vectors. In the case of the analogy "man is to woman as king is to queen," LLMs can identify this relationship through the arithmetic of word vectors (Vylomova et al. (2016)). The vector difference between "man" and "woman" is similar to the vector difference between "king" and "queen." Mathematically, this can be expressed as: king - man $\approx$ queen - woman. For a visual illustration, see fig. 6.

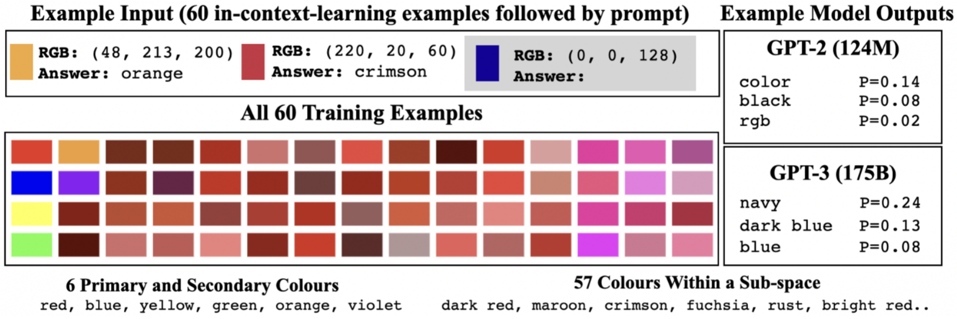
When the LLM searches for a word that satisfies this relationship, it can correctly identify "queen" as the word that completes the analogy. Likewise, the model can understand the relationship between a city and its country through a similar mechanism. The vector difference between "Paris" and "France" is similar to the difference between "Tokyo" and "Japan," enabling the model to complete the analogy.

Structural conditions on representational content may not themselves be sufficient to fully explain how mental states have truth conditions.[11] After all, several different networks of physical properties could stand in the same structural relations, and this would then leave unsettled which of them is the referent of the relevant mental state. But structural conditions can be combined

[11] See Piantadosi and Hill (2022) for an extended discussion of conceptual role semantics in LLMs.



(a) Isomorphic and Random Rotations in Color Space: The leftmost panel shows the full 3D color spectrum. The three right panels demonstrate how sample colors transform under different conditions: in their original positions, after a 90° rotation (an isomorphic transformation), and after random reassignment. This illustrates how structural relationships between colors are preserved in isomorphic rotations but disrupted in random rotations.


Example Input (60 in-context-learning examples followed by prompt)
Example Model Outputs
All 60 Training Examples
6 Primary and Secondary Colours
red, blue, yellow, green, orange, violet
57 Colours Within a Sub-space
dark red, maroon, crimson, fuchsia, rust, bright red..

```
RGB: (48, 213, 200)    RGB: (220, 20, 60)    RGB: (0, 0, 128)
Answer: orange         Answer: crimson       Answer:
```

GPT-2 (124M)

```
color       P=0.14
black       P=0.08
rgb         P=0.02
```

GPT-3 (175B)

```
navy        P=0.24
dark blue   P=0.13
blue        P=0.08
```

(b) Generalization to Unseen Color Terms: The left panel shows example training data, including primary, secondary, and shades of red. The right panel demonstrates model outputs for an unseen color (navy blue). Importantly, navy blue is in a separate color subspace from any of the colors used for training.

Figure 5: These figures together illustrate the study's approach to testing language models' understanding of color terms. While panel (b) suggests the model's ability to generalize to unseen colors, panel (a) shows how rotated color spaces were used to control for potential memorization of RGB-to-name mappings. This combined approach helps distinguish between true generalization and mere memorization of training data. (Adapted from (Patel and Pavlick 2022).

with other conditions to produce a full theory of mental content. For example, one strand of work in philosophy of mind has appealed to structural conditions to explain the particular significance of phenomenal experiences. According to representationalists about phenomenal character, phenomenal experiences supervene on their representational content. According to these theories, there can be no change in how things seem qualitatively without a change in the representational content of one's experiences. Many of these representationalists attempt to explain phenomenal experiences in terms of a special kind of representational content. Here, structural conditions have played an important role. For example, Rosenthal (2005) argued that qualitative experiences are associated with a network of representational contents that are arranged in a similarity space. Consider color concepts. Rosenthal's idea is that color experiences not only represent the world, but also stand in patterns of similarity to one another: red experiences are more similar to orange experiences than to yellow ones. And these similarity patterns are isomorphic to patterns in what color experience represents (for example, wave length). The interpretability research surveyed above suggests that LLMs may satisfy this condition on rich color representation.

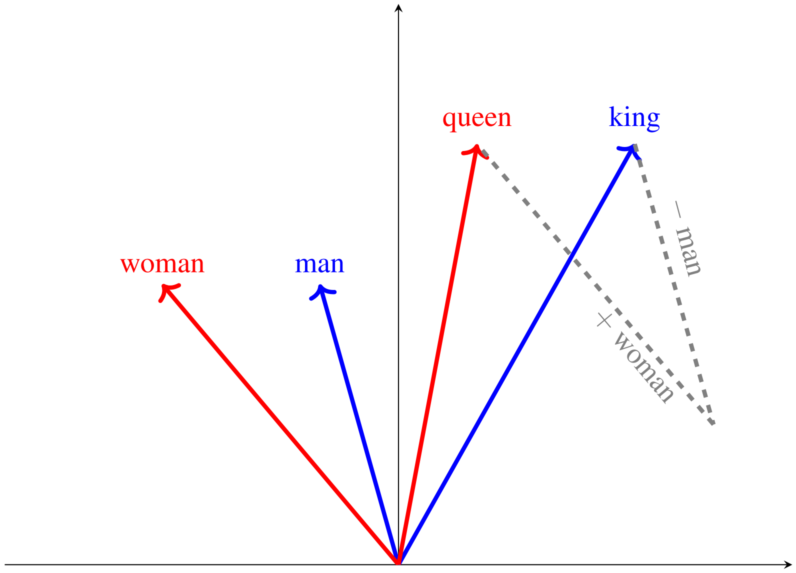


Figure 6: Word embedding vector operations illustrated: king - man + woman ≈ queen

Finally, this work on color concepts is also relevant to our earlier question of whether LLM activations satisfy generalizations from folk psychology. When LLMs successfully make predictions about the relationships between colors or directions they haven't previously encountered, they seem to reason in ways that are similar to the way humans reason.

2.5 Teleosemantics

Another potential necessary condition on representation comes from 'teleosemantic' theories of mental content. According to teleosemantic theories, the function of the underlying system generating a mental state determines its content (Millikan (1984)). For example, whether a given mental state counts as a representation of a goldfinch depends on the function of the system that created that mental state. On standard teleosemantic views, the function of a state is determined by evolution: the state has whatever function explains why it was selected for via natural selection (see (Schulte 2023, p. 35)).

Teleological conditions can be combined to causal and informational conditions on representation. For example, Stampe (1977)'s causal account of content appealed to the idea of causal correlations that obtain in normal conditions. This can be fleshed out in terms of evolution: a perceptual state represents a feature of the world if that state being reliably caused by the feature helps explain why the organism was selected for in natural selection. Similarly, Neander (2013, 2017) argues that perceptual states have content when they have the function of being reliably produced by features of the environment. For example, toads have evolved perceptual states that fire in response to "small, dark, moving" objects. These states were selected for by natural selection, because toads with these states were more successful at hunting flies. For this reason, these perceptual states represent small, dark, moving objects. Similarly, informational accounts can be supplemented with the condition that various information-carrying channels were selected for (see Dretske (1981)).

We argue that teleological constraints on meaning are broadly compatible with AI representation. Although LLMs are not products of biological evolution, they undergo a form of artificial selection during their training process. This selection process optimizes the model's parameters—

specifically, its weights—to improve its performance on specific tasks. In teleosemantic theories, the function of a mental state is determined by what it was selected for in the evolutionary process. Analogously, in LLMs, the function of a particular weight configuration is determined by its role in minimizing the loss function during training.

During training, the model's weights evolve in response to the pressure of minimizing loss, much like biological traits evolve under natural selection. This process bears important similarities to natural selection: just as biological traits that enhance fitness are more likely to persist, weight configurations that reduce prediction error are more likely to be retained. The resulting weight structure of the trained model can be seen as encoding the "functions" that the model was selected to perform, paralleling how evolved biological structures encode their functions.

The crucial disanalogy to natural selection is that in natural selection mutation is random, while in machine learning weights do not change randomly, but are instead adjusted in the direction of lower loss. But this disanalogy does not seem relevant to teleosemantic theories. In both cases, an underlying optimization process creates a clear sense of a goal, and of normal versus abnormal pursuit of that goal.

At this point, we've surveyed several different potential requirements on mental representation. We have argued that results about AI interpretability suggest that large language models can meet each of these requirements. We think that when LLMs play games like Othello, for example, they don't merely predict the next word. Instead, they represent the game environment. In particular, as LLMs learn how to complete this task, they build internal models, representations that help them figure out what to do next, for example by representing board states. In this way, we think that interpretability research provides significant reason to reject skepticism about mental representation.

3 Responding to Skeptical Challenges

At this point in the paper, we've laid out our positive claims about LLM mental representation. In short, we think that there is a strong case that LLMs possess robust internal representations of the world.

In this section, we'll respond to three skeptical challenges to LLM folk psychology. Each skeptical challenge targets the claim that LLMs represent the world. (In the next section, we turn to challenges to LLMs possessing robust action dispositions.) In each case, we'll identify potential gaps in the skeptical challenge, and consider how the skeptical challenge is potentially relevant to questions about LLM folk psychology.

We'll consider three challenges:

- **Sensory grounding**: This challenge says that LLM inputs are meaningless because LLMs have no connection to the external world. In response, we'll argue that (i) the challenge relies on overly simplistic causal constraints on mental representation; (ii) the challenge ignores the possibility that LLMs form hypotheses about aspects of the world that are not directly observable to them; and (iii) the challenge is *fragile*, because it relies on properties of LLMs (such as the lack of perceptual inputs) that are not shared by all models.
- **Stochastic parrots**: This challenge says that LLMs do not represent the world because they aren't trained to do so. In response, we'll argue that (i) the challenge overgeneralizes to threaten human cognition; (ii) the challenge ignores that LLMs could represent the external world as a means to predicting the next word; (iii) the challenge ignores the possibility that representation is an *emergent capability* of LLMs; and (iv) the challenge

is incompatible with the structure internal representations identified by interpretability research.

- **Memorization**: This challenge claims that LLMs do not represent the world because their behavior can be explained through the alternate theory that they simply memorize text. In response, we'll argue that LLMs possess the ability to generalize robustly from their training data, and then make correct predictions about new questions.

We can think of each of these three challenges as targeting one of the conditions on representation we discussed earlier. The sensory grounding challenge claims that representation requires specific kinds of causal connections between the external world and internal LLM states. The memorization challenge claims that representation requires robust models and reasoning abilities, which LLMs supposedly lack. The stochastic parrots challenge claims that representation requires the right kind of functional origin, rooted in training objectives.

3.1 Sensory Grounding

One skeptical challenge concerns sensory grounding. This skeptic says that LLMs don't represent the world because they lack sensory grounding and merely use text: "a system that is trained only on form [such as an LLM] would fail a sufficiently sensitive test [for intelligence], because it lacks the ability to connect its utterances to the world" (Bender and Koller (2020) p. 5188). Pure (text-only) LLMs only see patterns of text. They do not have any outside sensory input. Therefore, according to this challenge, they can't "break the syntactic circle" and connect any of the symbols they see to the outside world. As Harnad (1990) defined it, the symbol grounding problem is one of how the semantic interpretation of symbols can be "intrinsic to the system, rather than just parasitic on the meanings in our heads."

One version of the symbol grounding problem involves causal theories of mental representation. For some theories of mental representation, the right sort of causal connection between the outside world and internal states is required for such states to count as representational in the first place. According to these theories (including Stampe (1977) and Fodor (1987)), mental states represent the world when they are causally connected to the world in the right way.[12] For example, my perceptual experiences of cats tend to be caused by cats, and therefore represent them. In addition, my desire to eat ice cream tends to cause me to eat ice cream, and therefore represents ice cream. Causal theories of representation can explain, for example, why a photograph of an identical twin represents one twin and not the other, despite resembling each twin perfectly.

Naive causal conditions may make trouble for LLM representation. LLMs have activations that are related to cats. But these activations are not directly caused by cats in the way that our perceptual system directly responds to cats in our environment. Instead, LLMs have learned about cats through training on text, and this text itself has been caused by cats.

There are two ways to address causal skepticism about LLM representation. The first option is to appeal to long causal chains. The second option is to appeal to pluralism about causal structure. Let's consider each in turn.

[12] The precise details of the causal theory differ with each particular adherent. Stampe proposed that a mental state represents the proposition that p iff under optimal conditions, it is causally correlated with p. Stampe's notion of optimal conditions was then spelled out in terms of biologically normal conditions that "guarantee the well-functioning" of the mechanisms that produce the mental state. Fodor defended a different, "asymmetric dependence" theory. For Fodor, a concept C represents an object O when the concept is lawfully causally correlated with that object, and any other correlations between C and other objects are asymmetrically explained by the connection between C and O. For example, my concept ZEBRA is caused by zebras, but can also be caused by cleverly painted mules. But when ZEBRA is caused by a painted mule, this itself is explained by the underlying causal connection between ZEBRA and zebras.

The first response appeals to long causal chains. In some sense, LLM activations related to cats are not caused directly by cats, in the way that retinal stimulation is directly caused by cats. Instead, LLM activations related to cats are caused by text about cats. But cat text is itself caused by cats. In this way, there is a causal chain from cats to LLM activations about cats. So it isn't clear that even naive causal conditions on representation block LLMs from genuine reference.

The second response appeals to pluralism about causal structure. In fact, any causal theory of representation needs to be pluralistic about the types of causal correlation that facilitates representation. Stampe (1977) gives an example of barometers: the barometric representation is caused by a drop in air pressure, and the drop in air pressure causes the storm, and in this way the barometer represents the storm. While the barometric reading represents the storm, the storm is neither a cause of the barometric reading nor one of its effects. This is one of the ways in which informational theories of representation improve on naive causal theories: they can allow for a wide range of causal relations to produce genuine representation.

Another version of the sensory grounding challenge is illustrated by 'The Octopus Test', a thought experiment from Bender and Koller (2020). Imagine two humans communicating with one another through telegraphic cables across two islands, and a hyper-intelligent octopus that eavesdrops on their communication. In the beginning, the two humans mostly make small talk and describe their environments, which the octopus cannot see. However, because the octopus is hyperintelligent, it can pick up on various patterns in their communication. It has never seen trees, but it knows the patterns of discourse around the word "tree" and can do a good job of mimicking the other person. However, imagine that one of the people invents a catapult, and talks about it to their partner.

The skeptical challenge claims that the octopus cannot represent the catapult, or any other object that only exists on the islands rather than in the octopus's own environment. Because the catapult is novel to the discourse, the octopus won't be able to understand what it is or make predictions about how it works. And if this is the case, it suggests that LLMs cannot represent the world either.

We ourselves don't have a strong judgment about what the octopus can represent in this thought experiment. But we think one fruitful way of resolving the question is to consider various candidate conditions on representation. First, it is clear that the octopus carries information about the catapult, since it can reliably predict who will say what about the catapult. But, second, it is unclear whether the octopus has internal representations of the catapult that causally influence its predictions. It is also unclear whether the octopus has a series of internal representations of the catapult whose internal structure or pattern of similarity mirrors patterns in the catapult. Moreover, it is unclear whether the octopus evolved to track the catapult through some kind of optimization process such as natural selection. For all of these reasons, we think it is quite unclear whether the octopus represents the catapult.

Finally, it is unclear whether the octopus can reason well about the catapult in a wide range of counterfactual scenarios, in the kinds of ways that would produce genuine world models. The claim that the octopus simply could not in principle represent the catapult in such a way ignores a crucial hypothesis. When an LLM (or octopus) receives text, if it is sufficiently capable, it can form hypotheses over what sort of processes generated such a text.

Just as scientists form hypotheses about microscopic phenomena responsible for observed patterns, advanced octopuses could make educated guesses about how the world works based on the patterns and types of text they receive. For example, if it reads lots of texts about machine learning, it might form certain hypotheses about certain types of researchers and computational infrastructure existing in the real world. From the under-water version of Wikipedia, it can make guesses about the sorts of processes and agents that would generate Wikipedia articles and what sort of underlying reality would result in such strings of text.

While direct empirical evidence for LLMs (and octopuses) forming explicit hypotheses about the world is limited, the possibility remains theoretically plausible. If advanced LLMs can form hypotheses about how the world works, then the symbols they receive could connect to the outside world, even if they do not have the capability of receiving anything other than text as input. Future research should aim to uncover more about the internal mechanisms of LLMs and their ability to form and use such hypotheses about the world.

Another problem for the sensory grounding challenge is its fragility. Even if skeptics can successfully argue that a pure LLM is unable to represent reality, their victory may only be a Pyrrhic one. The transformer architecture, which underlies these models, is designed for sequence prediction generally and is versatile enough to be applied to various modalities, including computer vision.

Vision-Language Models (VLMs), such as GPT-4V, integrate both visual and textual data, processing them within a shared embedding space. Unlike systems that separately handle text and images and then attempt to combine their outputs, VLMs use a single, unified transformer model capable of processing and understanding both modalities simultaneously.

For instance, an image of a cat and the word “cat” would be represented in the same space, allowing the model to draw connections between the visual and textual representations. This shared embedding space helps bridge the gap between sensory input and linguistic representation. By incorporating visual data, VLMs gain a form of sensory grounding that pure text-based LLMs lack.

The architecture of VLMs is fundamentally similar to that of standard LLMs, relying on the same principles of self-attention and sequence modeling. This similarity undermines the argument that the transformer architecture itself is incapable of genuine representation. If VLMs, which are built on the same foundational architecture, can achieve sensory grounding and representational capabilities, it suggests that the perceived limitations of LLMs are not inherent to the architecture but rather a consequence of the input modalities used.

For all of these reasons, we do not consider the sensory grounding challenge a decisive threat to LLM representation.

3.2 Stochastic Parrots

Another skeptical challenge to LLM representation is associated with the “stochastic parrots” argument, originally proposed by Bender et al. (2021). This view contends that LLMs, despite their impressive outputs, are merely sophisticated statistical pattern matchers rather than systems capable of genuine understanding or representation. The core claim of the stochastic parrots argument is that LLMs are trained solely to predict the next word in a sequence, without any true comprehension of the content they generate. According to this view, LLMs don’t represent or reason about the world; instead, they simply reproduce patterns from their training data in a statistically sophisticated but fundamentally meaningless way. Proponents of this view often point to cases where LLMs produce fluent but nonsensical or contradictory outputs. For instance, an LLM might confidently assert a false statement or agree with contradictory premises in different conversations. These behaviors, they argue, reveal that LLMs lack genuine understanding and are merely “parroting” patterns from their training data.

One way to understand the stochastic parrot challenge is in terms of the teleosemantic conditions on representation we discussed earlier. Those conditions required that the internal states of LLMs have the function of representing features of the external world. This function would itself need to emerge from some kind of evolutionary, selective process.

Here, we'll lay out four challenges for the stochastic parrots view: The first challenge is overgeneralization. The argument that systems trained on pattern recognition can't develop genuine understanding potentially overgeneralizes. Human cognition, for instance, involves significant pattern recognition and statistical learning, yet we don't deny humans the capacity for genuine understanding. Similarly, other AI systems trained on pattern recognition (like computer vision models) are often considered to represent features of the world. The key question is not whether a system is trained on patterns, but whether it develops more sophisticated capabilities as a result of this training.

The second challenge is representation as a means to an end. While LLMs are indeed trained to predict the next word, representing the world could be an efficient means to this end. For example, to accurately predict words in a physics textbook, it would be helpful for an LLM to develop some internal understanding of physics concepts. Our earlier discussions of probing studies and world models provide evidence that LLMs do develop structured internal representations that go beyond simple pattern matching. (See Herrmann and Levinstein (2024) for more.) Again, this can be unpacked in terms of teleosemantic requirements on representation. The idea would be that the training process of predicting the next word selects for the ability to track features of the external world.

The third challenge is emergent capabilities. Recent research suggests that LLMs can exhibit capabilities that were not explicitly part of their training objective. For instance: a) Few-shot learning: Brown et al. (2020) demonstrated that GPT-3 can perform new tasks with just a few examples, despite not being explicitly trained for this ability. This "few-shot" learning suggests a form of rapid adaptation that isn't easily explained by simple pattern matching. b) In-context learning: Wei et al. (2022) showed that LLMs can learn to perform new tasks from instructions and examples provided in the input prompt, without any change to their weights. This ability to "learn" within the context of a single forward pass challenges the notion of mere pattern reproduction. c) Chain-of-thought reasoning: Wei et al. (2022) also demonstrated that prompting LLMs to generate step-by-step reasoning significantly improved their performance on complex reasoning tasks, suggesting a capacity for structured thinking beyond simple pattern matching. d) Emergence of coding abilities: Chen et al. (2021) found that large language models trained on natural language can develop unexpected coding abilities, despite not being explicitly trained on programming tasks. e) Zero-shot task generalization: Kojima et al. (2022) showed that LLMs can solve novel tasks they weren't explicitly trained on when prompted to "think step by step," demonstrating a form of reasoning capability.

The fourth challenge is internal representations. As we've discussed earlier in this paper, interpretability research provides evidence that LLMs develop rich internal representations of concepts and relationships. This structured internal knowledge is difficult to reconcile with the stochastic parrots view. It's worth noting that the stochastic parrots argument raises important questions about the nature of understanding and representation. Even if we reject the strong claim that LLMs are merely stochastic parrots, we might consider intermediate positions. For instance, LLMs might combine aspects of statistical pattern matching with more sophisticated representational capabilities. Ultimately, the stochastic parrots argument highlights the need for careful empirical investigation of LLM capabilities and limitations. While the behavior of LLMs can sometimes be consistent with sophisticated pattern matching, the evidence we've reviewed throughout this paper suggests that LLMs do develop meaningful internal representations and capabilities that go beyond simple parroting.

3.3 Memorization

Another skeptical challenge concerns memorization. According to this challenge, we don't need to posit genuine reasoning in LLMs, because we can explain their behavior in another way. Instead of reasoning about questions, LLMs simply memorize great quantities of data and then use shallow heuristics to generalize to new prompts. They are trained on billions of sentences, and in the course of this training, they simply store the answers to a large number of questions. These answers are then returned in response to prompts, without any genuine reasoning taking place.

Whether memorization threatens LLM representation depends on what is required for representation. If representation is simply a matter of carrying information, then memorization is perfectly compatible with genuine LLM representation. On the other hand, if representation requires the presence of robust models or of reasoning about the world that matches folk psychology, then memorization may rule out representation.

In fact, there is a lot of evidence that LLMs do not merely memorize the answers to questions they are asked. Instead, LLMs seem to possess the ability to generalize robustly from their training data and then make correct predictions about new questions. They do not simply use shallow heuristics.

Return to the case of modular addition. As we discussed above, Nanda et al. (2023) found that during training, the LLM learns modular addition in multiple steps: an initial period of overfitting, based on memorization, followed by a transition to a general solution to the problem. In the initial phase of training, the LLM does engage in memorization. But after using memorization, the LLM learns a more general algorithm for computing the answer. After this algorithm is learned, the LLM then removes its memorization components. As evidence for this claim, Nanda et al found that in the initial period of training, the model achieved 100% accuracy on the training data, but low accuracy on the testing data. After 10,000 epochs of training, the model learned how to actually perform the task, and achieved high accuracy on the testing data. Overall, this suggests that LLM skeptics are missing out on much of the rich structure of LLM reasoning.

Similar abilities to generalize were found in the Othello experiment. Every game of Othello starts with one of four moves. Each initial move creates a 'quadrant' of Othello game space. To train Othello, the experiments only included training data from 3 of the 4 quadrants. But they found that the model was equally successful at playing Othello in the omitted quadrant of game space. [13] [14]

[13] Our focus in this paper is on whether LLMs have a folk psychology. This question is worth distinguishing from another skeptical target: the question of whether LLM *outputs* are meaningful. Here, the key question is whether text produced by LLMs has meaning. This is a different question than whether LLMs have a folk psychology. Compare: we can imagine a human being who has beliefs and desires, but who does not know how to speak French. If they started saying French sentences out loud phonetically, there would be an interesting question to ask about whether these sentences are meaningful in their mouth. But this question is a separate one from whether the speaker has a psychology. In practice, much of the debate about LLM outputs has itself been connected to debates about LLM folk psychology. For example, one skeptical challenge to the meaningfulness of LLM outputs starts from the premise that LLMs lack communicative intentions. This is itself a skeptical premise about LLM folk psychology. In response, Mandelkern and Linzen (2023) have suggested that LLM psychology may not be necessary for LLM outputs to be meaningful, as long as LLMs as part of our linguistic community. (See also Mollo and Millière (2023).) Others have suggested that we think of LLMs as more like *libraries* rather than speakers, which could also allow outputs to be meaningful (for discussion, see Gopnik (2022)). By contrast, Lederman and Mahowald (2024) argue that some kinds of LLM outputs can only be meaningful of LLMs are genuine speakers. Our focus in this paper is the question of whether LLMs have a psychology, rather than the question of whether LLM outputs are meaningful.

[14] The connection between belief and learning may offer another route to LLM skepticism. The philosopher Grace Helton has argued that in order to genuinely have beliefs, those beliefs must be able to change in response to evidence (Helton (2018)). Helton's argument has two premises. First, you have a belief only if you are obligated to change that

4 Action and Folk Psychology

While we've argued that LLMs have mental representations, the question remains: do they have a robust folk psychology? To address this, we need to consider whether LLMs have beliefs about the world, desires they aim to satisfy, and intentions that guide their actions. This question is complex, involving issues of stability, coherence, and goal-directed behavior.

To approach this question, we'll examine two influential philosophical perspectives on folk psychology: interpretationism and representationalism. Each offers a different framework for understanding mental states and presents different challenges when applied to LLMs.

4.1 Interpretationism

The most radical view is interpretationism. The idea behind interpretationism is that folk psychology is for explaining behavior. All that is required to have a folk psychology is for the system to behave in sufficiently complex ways best explained by appeal to folk psychological states. When this happens, the system has beliefs and desires: the system's desires are the goals promoted by its actions, and its beliefs are the views about the world that are required to be true in order for its actions to promote its goals. Whether it has internal states of a certain kind is not directly

belief in response to strong counter-evidence. Second, you are obligated to do something only if you are able to do so. From these premises, it follows that LLMs have beliefs only if they are able to change their beliefs in response to strong counter-evidence. But one might worry that after training, LLMs can no longer learn or change their beliefs. If Helton's premises are correct, it follows that LLMs do not have beliefs.

The initial skeptical concern here is that after training LLMs, no longer learn or change their beliefs. The idea is that all of the learning that an LLM engages in occurs during training, when the weights of the LLM's neural network gradually change in response to feedback. But once a user interacts with the LLM, for example using ChatGPT, the weights are frozen, and so the LLM no longer learns.

This skeptical response fails, because of in-context learning. In-context learning refers to the ability of a language model to use the context provided within a given prompt or conversation to generate relevant and coherent responses (Brown et al. (2020)). In general, in-context learning can refer to any way in which the model aptly adapts to the context of the prompt: e.g., adapting the right style, responding to corrections, or maintaining continuity over a conversation.

For our purposes, we can focus on few-shot learning, which is a form of in-context learning. In few-shot learning, you can provide examples within your prompt, and the model will use these to understand the type of response you're looking for.

For example, if you want GPT-3 to predict nationalities, you might just ask it about Marie Curie's nationality. But with a few-shot prompt, you instead give it Einstein and Gandhi's nationality, and then ask it Marie Curie's nationality. Often, LLMs will perform better after seeing a few examples, rather than answering a question 'zero shot'. This suggests that the LLM learns from the initial examples (Xie et al. (2021)).

Xie et al. (2021) found that in-context learning in LLM satisfies many features of Bayesian inference. In particular, they hypothesized that in few-shot learning, there's a hidden concept that explains each of the examples in the prompt and that should be used to generate the response. They studied how a perfect bayesian would guess the right response given this hypothesis and compared it to how GPT-2 performed. GPT-2 learned quickly, like a bayesian, and got better and better with more examples.

Although LLMs of today do not retain the information learned in-context in other inference cycles, they do update their beliefs within a given inference cycle. Humans likewise often forget some of their beliefs–albeit not always–so we do not see a special reason to deny that LLMs can learn and change their minds.

Philosophers often distinguish dispositional from occurrent beliefs. This distinction is particularly interesting in the setting of LLMs. In LLMs, dispositional representations would be stored in the underlying weights of the neural net, while occurrent representations would be tokened by the activations in response to prompts. When LLMs engage in on-line learning, their occurrent representations change in response to evidence. In this way, Helton's argument could allow that LLMs possess occurrent beliefs, even if they lack dispositional beliefs. Rather, they would possess unchanging dispositional representations that do not respond to evidence, and so are not subject to rational obligations, even though they could guide action and inference.

relevant to the question of folk psychology. Prominent interpretationists include Dennett (1981) and Davidson (1984).[15]

The previous conditions we've explored could allow for a separation between mental representation in general and belief/desire psychology in particular. LLMs could satisfy informational, causal, structural, and teleosemantic requirements on content, even if their behavior is too disorganized and happenstance to count as possessing beliefs and desires. Importantly, for interpretationists, folk psychological states come as a package deal. As Stalnaker (1984) puts it:

> Belief and desire . . . are correlative dispositional states of a potentially rational agent. To desire that P is to be disposed to act in ways that would tend to bring it about that P in a world in which one's beliefs, whatever they are, were true. To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true. (p. 15)

So, even if LLMs have something like mental representations of the world, they might not have desires or the right kinds of behavioral dispositions required for folk psychological states.

Do large language models satisfy interpretationist conditions on mental representation? There are at least two reasons for skepticism. The first challenge concerns LLM affordances: what actions an LLM can perform. Pure LLMs do not have access to robotic bodies. Instead, they simply produce text (or probability distributions over tokens). But at first glance these text interactions aren't naturally suited for performing complex actions.

There are at least two good responses to the problem of affordances. First, the problem of affordances is fragile. Some LLMs today do have access to robotic limbs. For example, Google's Palm-E system integrates an LLM with a robotic limb, which can perform actions after being prompted by a user with text (Driess et al. (2023)). When asked to pick up a bag of potato chips from the counter, the robotic limb can skillfully sort between different objects and pick up the chips.

Second, text production provides a rich enough space of alternative outcomes to count as a full-fledged action space. In today's digital world much human action takes place in text. When we imagine a human being whose life is confined entirely to text-based actions, we see no barrier to such a being possessing beliefs and desires. To see this point in greater detail, consider the use of LLMs in game environments. Mei et al. (2024) studied the behavior of LLMs in social cooperation games. These kinds of environments allow LLMs to formulate complex plans. They found that GPT-3 and GPT-4 would adjust their behavior throughout the game:

> In games with multiple roles (such as the Ultimatum Game and the Trust Game), the AIs' decisions can be influenced by previous exposure to another role. For instance, if ChatGPT-3 has previously acted as the responder in the Ultimatum Game, it tends to propose a higher offer when it later plays as the proposer, while ChatGPT-4's proposal remains unchanged. Conversely, when ChatGPT-4 has previously been the proposer, it tends to request a smaller split as the responder. (p. 17)

This kind of behavior suggests that LLMs have the ability to use game theory to navigate environments involving other agents.

[15] Closely related to interpretationism is dispositionalism. On a dispositionalist picture, an agent has a belief that P if it is disposed to behave as if P. See, e.g., Marcus (1990).

A second reason for skepticism about action is instability. Shanahan and others have noted that LLM outputs are very sensitive to prompting. If you slightly change the way you ask a question, the LLM can start to behave very differently. This makes it hard to see the LLM as taking a wide range of means to promote a unified goal (Shanahan et al. (2023)).

We think instability is potentially the most serious threat to LLM folk psychology. We see two potential responses worthy of further development. First, one might concede that LLMs are relatively unstable but say that each prompting session with an LLM produces a different agent with its own beliefs and desires. For example, if you ask GPT-4 to write poetry for you on one day and start a new conversation asking it to play chess with you the second day, you in effect have two separate agents. The chess-playing instance has no interest in poetry, but it does have the goal of beating you at chess. When you give it goals like writing poetry or playing chess or coding, it does a good job. Those tasks are very difficult. So, the thought goes, the best explanation is that the beliefs and desires of the LLM change from inference cycle to inference cycle, but the behavior of particular instances is still fruitfully explained with beliefs and desires and intentions.

Second, one might reject the claim of instability. Although the behavior looks unstable, the different behavior of LLMs in different prompting sessions might derive from a stable underlying goal, such as pleasing the user. The different outputs of LLMs in different sessions might result from a disconnect between what the model believes and what the model says. In other words, instability in model outputs is evidence of lying, not evidence that the system lacks beliefs.[16]

Above all, we suggest that more research needs to be done about how stable LLM outputs are to a range of different prompting environments. For example, one fruitful project would be to explore how LLM strategies in game environments change as a result of different prompting conditions. This would allow some measure of whether it forms coherent plans that are robust to a range of perturbations.

4.2 Representationalism

Representationalism, advocated by philosophers like Jerry Fodor and Fred Dretske, holds that having mental states requires having internal representations with appropriate functional roles. On this view, to have a belief that p, a system must have an internal state that represents p and plays the right causal role in the system's cognitive economy.

There is cause for optimism on the representationalist picture. As we've argued at length, some LLM mental states can have truth-conditions. Furthermore, LLMs even appear to have some sorts of world models and structured representations of conceptual domains.[17]

However, even on representationalist pictures like Fodor's, beliefs have to play the right role in the larger system and generally cannot be divorced entirely from desires. Even if LLMs have rich internal representations, it's not clear that these play the right kind of role in generating behavior to count as beliefs or desires. The issue of instability resurfaces here—if internal representations don't stably guide behavior across different prompts, can they really count as beliefs?

[16] For an argument that RLHF helps ground textual meaning through the goal of pleasing the user, see (Mollo and Millière 2023).

[17] We might require that LLMs internally distinguish between true claims and false claims in a way that goes beyond representation for the LLM to count as having beliefs. In particular, we might require that they somehow systematically "tag" sentences as true or false and use this tag in their master algorithm to determine what text to output?

We don't yet have a full answer to this question. Some (Burns et al. (2024), Azaria and Mitchell (2023)) have argued that LLMs do make an internal distinction between truth and falsity. However, others (e.g., Levinstein and Herrmann (2024)) claim these studies are flawed. Herrmann and Levinstein (2024) argue for a representationalist account of belief for LLMs but maintain that current empirical evidence of whether LLMs actually have beliefs is inconclusive.

Ultimately, we think matters are easier for the representationalist than the interpretationist. While we have some behavioral evidence in the case of LLMs, behavioral evidence is much more limited for LLMs than it is for humans. However, we have perfect internal access to LLMs, and we have much to discover about the role various representations play in LLMs' cognition. Therefore, as we come to understand more about how LLMs think, it will become more obvious for representationalists whether they have folk psychological mental states. Some key open questions for attributing folk psychological states on either picture, then, include:

- **Action and planning**: How can we best understand LLM "action" given their limited affordances?
- **Stability**: How can we reconcile the apparent instability of LLM outputs with the need for stable beliefs and desires?
- **Goal-directedness**: Do LLMs have anything analogous to enduring goals or values?

Addressing these questions will require a combination of philosophical analysis and empirical investigation. While there's evidence that LLMs have sophisticated internal representations and can exhibit complex, apparently goal-directed behavior, significant questions remain about whether they possess full-fledged folk psychological states. Resolving these questions will be crucial for understanding the capabilities, limitations, and potential moral status of these increasingly ubiquitous AI systems.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color, 2021.

Fred Adams and Ken Aizawa. Causal Theories of Mental Content. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form, 2023.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463`.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024.

Paul A. Boghossian. The status of content. *Philosophical Review*, 99(2):157–84, 1990. doi: 10.2307/2185488.

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

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024.

Sergio E Chaigneau, Lawrence W Barsalou, and Steven A Sloman. Assessing the causal structure of function. *Journal of Experimental Psychology: General*, 133(4):601, 2004.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL `https://arxiv.org/abs/1706.03741`.

Jeff Da and Jungo Kasai. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. *arXiv preprint arXiv:1910.01157*, 2019.

Donald Davidson. *Inquiries Into Truth and Interpretation*. Oxford University Press, Oxford, GB, 1984.

Daniel Clement Dennett. *The Intentional Stance*. MIT Press, 1981.

Fred I. Dretske. *Knowledge and the Flow of Information*. MIT Press, Stanford, CA, 1981.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models, 2020.

Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975.

Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, 1987.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*, 2019.

Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information, 2021.

Alison Gopnik. Children, creativity, and the real key to intelligence. *APS Observer*, 35, 2022.

George Graham, Terence E. Horgan, and John L. Tienson. Consciousness and intentionality. In Max Velmans and Susan Schneider, editors, *The Blackwell Companion to Consciousness*, pages 468–484. Blackwell, 2007.

Wes Gurnee and Max Tegmark. Language models represent space and time, 2024.

Jacqueline Harding. Operationalising representation in natural language processing. *British Journal for the Philosophy of Science*, forthcoming. doi: 10.1086/728685.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

Chris Heathwood. Desire-fulfillment theory. In Guy Fletcher, editor, *The Routledge Handbook of the Philosophy of Well-Being*, pages 135–147. Routledge, 2016.

Grace Helton. If you can't change what you believe, you don't believe it. *Noûs*, 54(3):501–526, 2018. doi: 10.1111/nous.12265.

Daniel A. Herrmann and Benjamin A. Levinstein. Standards for belief representations in llms, 2024.

Daniel Hutto and Ian Ravenscroft. Folk Psychology as a Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, 2020.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Harvey Lederman and Kyle Mahowald. Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms, 2024.

Benjamin A Levinstein and Daniel A Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27, 2024.

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

Matthew Mandelkern and Tal Linzen. Do language models refer? *arXiv preprint arXiv:2308.05576*, 2023.

Ruth Barcan Marcus. Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research*, 50(n/a):133–153, 1990. doi: 10.2307/2108036.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

Ruth Garrett Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism.* MIT Press, 1984.

Dimitri Coelho Mollo and Raphaël Millière. The vector grounding problem, 2023. URL `https://arxiv.org/abs/2304.01481`.

Sam Musker and Ellie Pavlick. Testing causal models of word meaning in gpt-3 and-4. *arXiv preprint arXiv:2305.14630*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Karen Neander. Toward an informational teleosemantics. In Dan Ryder, Justine Kingsbury, and Kenneth Williford, editors, *Millikan and her critics*, pages 21–40. Wiley, 2013.

Karen Neander. *A Mark of the Mental: A Defence of Informational Teleosemantics*. MIT Press, Cambridge, USA, 2017.

Jonathan Opie and Gerard O'Brien. Notes toward a structuralist theory of mental representation. In Hugh Clapin, editor, *Representation in Mind: New Approaches to Mental Representation*, pages 1–20. Elsevier, 2004.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*, 2022.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Steven T. Piantadosi and Felix Hill. Meaning without reference in large language models, 2022.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020.

David M. Rosenthal. *Consciousness and Mind*. Oxford University Press UK, New York, 2005.

Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

Peter Schulte. *Mental Content*. Cambridge University Press, 2023.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Robert C. Stalnaker. *Inquiry*. Cambridge University Press, 1984.

Dennis W. Stampe. Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy*, 2(1):42–63, 1977. doi: 10.1111/j.1475-4975.1977.tb00027.x.

Marius Usher. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language*, 16(3):331–334, 2001. doi: 10.1111/1468-0017.00172.

Keyon Vafa, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating the world model implicit in a generative model, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1158. URL `https://aclanthology.org/P16-1158`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023.