that, "[a] bare-bones LLM doesn't 'really' know anything because all it does, at a fundamental level, is sequence prediction." (p. 5, Shanahan (2022)).[3]

We've argued in detail elsewhere that these concerns rest on a philosophical mistake (Levinstein and Herrmann 2024). Briefly, this is an inference from the *goal* of a system[4] to a claim about *how* the system accomplishes its goal. But just because truth-tracking is not the goal of the system, does *not* mean that the system does not track the truth as a *means* to accomplishing its goal. Indeed, humans are partially products of a process that maximizes inclusive genetic fitness; and yet, as a byproduct of this process, we track the truth, at least in some domains some of the time.

As we've emphasized, it should be a live hypothesis that LLMs track the truth. In the remainder of the article, we propose conditions that more carefully spell out what would have to be hold of an internal representation of an LLM for it to count as a belief. To be clear: we are entirely open to the possibility that LLMs do *not* have representations that satisfy these conditions. In such a situation, we think that it would likely not be very useful to describe LLMs as having beliefs.

4. Where to Look for Beliefs

For guidance on how to measure beliefs, we might look to some of our best theories of belief. Given our interest in the truth-tracking, action-guiding aspects of belief, we might take inspiration from decision theory. In decision theory, there is a long tradition of reconstructing an agent's beliefs and desires (credences and utilities) from her preferences via representation theorems (Ramsey (1926); Savage (1972); Jeffrey (1990)). In the radical interpretation tradition, philosophers also attribute beliefs by appealing to interpretational maxims in conjunction with observable behavior and utterances (Davidson (1973)).

However, when it comes to LLMs, we have a number of disadvantages. It's not clear they have preferences, and their behavior is quite limited. They do not engage in long-term planning nor do they have bodies that can physically interact with the world.[5] They simply output probability

[3]The requirements we propose here focus on the action-guiding, truth-tracking aspects of belief. These are the ones that we think are most relevant for the pragmatic goals upstream of understanding LLM cognition. Though not of concern for us here, issues of communicative intent, symbol grounding, and reference also drive skepticism of belief in LLMs (Bender and Koller (2020); Bender et al. (2021); Shanahan (2022)). Mandelkern and Linzen (2023) counter these concerns using externalist arguments, claiming LLMs can refer, while Piantadosi and Hill (2022) use an internalist approach to dismiss grounding concerns, emphasizing internal conceptual roles. Pavlick (2023) summarizes both internalist and externalist perspectives on whether or not the internal states of LLMs encode meaning.

[4]Or, more carefully, from what its training objective was.

[5]Of course, there is a sense in which they *do*. Ultimately, the LLM's computations are executed on physical hardware somewhere. The point we are making is that its way of

distributions over tokens and have no robust ways of bending the world to their will.

The standard methods for eliciting honest beliefs from human agents also fall short with LLMs. For instance, to determine if a human judges that P is more likely than Q, we can offer a choice of a dollar if P or a dollar if Q. If they choose the bet on P, it indicates they believe P is more likely than Q. Alternatively, we can elicit human credences by paying them based on their announced forecasts in accordance with the Brier Score or other strictly proper scoring rules (Brier 1950; Gneiting and Raftery 2007).[6] Strictly proper scoring rules incentivize an expected wealth-maximizing agent to report their actual credences. That is, if they really believe a proposition to degree x, then they maximize their expected wealth by reporting their credence to be x.

In contrast, while we can offer LLMs bets or ask them about their beliefs and tell them we'll pay them according to their Brier score, they can't actually receive payment, and it's unclear that they would care about money even if they had bank accounts. Therefore, standard methods of eliciting beliefs using betting or scoring rules are ineffective.

Given the lack of behavioral evidence and reliable reward methods, traditional tools from formal epistemology, decision theory, radical interpretation, and economics are inadequate for finding beliefs in LLMs. Moreover, LLMs have architectures unlike that of the human brain and lack shared evolutionary or cultural history with us, depriving us of certain shared understandings and common ground that we take for granted among humans.

Nevertheless, there are advantages in interpreting the minds of LLMs. Although the high-level algorithms they use are opaque, we have perfect low-level access. We can see the embeddings at each layer and the internal weights of the network. We can also perform precise modifications or ablations on the model's weights or the components of any embedding. For instance, we can adjust the hidden embedding at a given layer for a token and observe how this changes the model's output. Additionally, it is easy to reset the memory of LLMs; each new conversation or inference cycle begins with the LLM in the same state, with no memory of previous prompts.

Thus, given that we are in a different epistemic situation when we are trying to measure beliefs in LLMs than in humans, we propose looking internally—inside the model's head, as it were—to find out what it believes. That is, we want to find *internal representations of truth*. If we find such a representation, then it makes sense to attribute beliefs to LLMs.

An internal representation of truth is a mechanism by which the model can internally tag a sentence as true or false (or mark it with some level of

interacting with the world is filtered almost entirely through its textual outputs, unlike humans and other critters.

[6]More explicitly, according to the Brier Score, if they announce a credence of x in a given proposition and that proposition turns out to be true, we will pay them $\$1-(1-x)^2$, and if it's false, we will pay them $\$1-x^2$.

confidence) and use that tag along with other information it has computed to figure out what to output.

The question, then, is whether in the course of its computations an LLM internally distinguishes between true claims and false claims and uses this distinction, in part, when deciding what to output.

To be clear, internal representations of truth aren't, in general, necessary on many accounts of belief. As we've seen, representation theorems in decision theory only require preferences with beliefs and desires derived thence. Indeed, belief-desire psychology has been very successful for humans for a long time even though we had at best very limited direct access to internal states of other people until a few decades ago. However, with LLMs, we have a very different evidential basis. Behavioral evidence is much more limited, while internal access is much greater.

Discovering internal representations of truth also has important social and ethical implications. In the relatively near term future, society will likely use LLMs to automate a number of tasks previously reserved for humans, and we will want to know what they really think and whether we can trust them. For example, we might use LLMs to conduct job interviews, where they have a conversation with a candidate (just like current human interviewers do), and then make recommendations about who should be hired and give justifications for those recommendations.

However, without being able to check what their internal reasons are for their recommendations, we have no good way of ensuring the justifications they provide actually align with their thought processes. They may use some illicit feature like race or gender in a problematic way while also justifying their recommendations for totally different reasons. As Zhou and Joachims (2023) demonstrate, it is very easy to justify the decisions made by models in ways that aren't faithful to the actual functioning of the model.[7]

In addition to checking that explanations are faithful to internal processes, belief measurement also provides a strategy to detect deception, which can be important for designing safe and cooperative AI systems (Dafoe et al. (2020); Evans et al. (2021); Park et al. (2024)). Indeed, this is the explicit motivation of many of the articles developing belief measurement techniques (Azaria and Mitchell (2023); Levinstein and Herrmann (2024)).

We believe that belief measurement is one important tool in the effort to develop ethical and safe AI, but is not sufficient on its own, or even necessary in general. There are contexts in which trusting an AI system does not require looking at the internals of a system if other criteria are met, such as robust and reliable uncertainty quantification (Grote (2021)). Indeed, there are even strategies that try to bypass a need for lie detection by building systems that do not deceive in the first place (Ward et al. (2024)), or are by their construction interpretable by design (Grote (2023)). We support such

[7]See Zhou et al. (2020) for an overview on how explainable AI and fairness in AI relate, and contribute to trust in AI.