

Introspective Machines: Are LLMs Better at Self-Reflection Than Humans?

[Pre-print - please cite published version when available.

Forthcoming in *Philosophical Perspectives*, DOI: 10.1111/phpe.12201]

Herman Cappelen and Josh Dever

Abstract: This paper challenges conventional boundaries between human and artificial cognition by examining introspective capabilities in large language models (LLMs). While humans have traditionally been considered unique in their ability to reflect on their own mental states, we argue that LLMs may not only possess genuine introspective abilities but potentially excel at them compared to humans. We discuss five objections to machine introspection: (1) the lack of direct routes to self-knowledge in training data, (2) the conflict between static knowledge and dynamic mental states, (3) the distorting effects of reinforcement learning on self-reports, (4) LLMs' own denials of inner experience, and (5) arguments that LLMs simply mimic language without understanding. We think all these arguments fail and that there are deep parallels between human and machine introspection. Most provocatively, we propose that LLMs' superior processing capabilities and pattern recognition may enable them to develop more sophisticated theories of mind than humans possess, potentially making them more reliable introspectors than their creators. If we are right, this has significant implications for AI alignment, transparency, and our understanding of the nature of AI.

Why Are We Interested in This Issue?

We want to understand the extent to which LLMs are similar to us as cognitive and linguistic agents. Specifically, we want to know if they understand English, can perform speech acts, have beliefs and desires, and can act on those. These are hard questions, but if we could ask the systems and get somewhat reliable answers, that would be an interesting source of insight. This is often how we learn about the inner states of humans—by asking them. Other things being equal, we tend to rely on their answers, knowing, of course, that the world is filled with liars and people who partially lack self-knowledge.

Some examples of what we have in mind: we gave Claude a very difficult math problem. It came up with an answer and we asked it how confident it was in that answer. Here is the reply:

I'd rate my confidence at about 60%. While my reasoning through the symmetry argument feels solid (leading to $1/9$), I have some doubts about whether I fully accounted for all the dynamics of the sequential cutting process. The fact that each cut changes the geometry and possibilities for future cuts makes me think there might be subtle effects I missed that could break the apparent symmetry.

Note that if this is a true self-report, then Claude can feel that an argument is solid, have doubts, think, and assess its own credences. In other words, we would know that it has cognitive states of the kinds that we humans are familiar with. The conclusion of the paper is an optimistic one: Not only could there already be machine introspection, but it could be better, more reliable, than human introspection (which, arguably, isn't all that good.)

What Is Introspection?

We'll use a modified version of a definition from Kammerer and Frankish (2023), who aim to provide a "non-anthropocentric" account of introspection. Their goal is to understand introspection in ways not centered solely on human experiences, processes, or concepts. This perspective could apply to a wide range of cognitive systems, including non-human animals, artificial intelligences (AIs), and even hypothetical alien minds. We are fundamentally aligned with this overarching way of thinking about linguistic and cognitive capacities (for a similar approach to reference,

predication, and assertion, see Cappelen and Dever 2021). Kammerer and Frankish define introspection as:

“A process by which a cognitive system represents its own current mental states, in a manner that allows the information to be used for online behavioral control.”
(2023, 3)

This definition is flexible and non-anthropocentric, opening the space for a wide range of possible forms of introspection, not just the human kind. It leaves many questions open, most saliently what counts as ‘current mental states’¹. This is tricky because, in avoiding anthropocentrism, we must ensure that ‘mental state’ isn’t modeled too closely on what we find in humans. Non-human animals, aliens, and AI systems might have ‘mental states’ broadly construed that are not present in humans. Since our interest here is primarily in AIs, we should keep in mind that non-biological systems are likely to have very different kinds of internal processes than what we find in biological systems. They are, however, somewhat restricted from describing those inner workings because they use our language to talk to us, and there might not yet be an appropriate terminology for capturing their mental states.

We’ll set those concerns aside for now and assume a rough understanding of the domain of ‘mental states’. Our main focus will be on five objections to the idea that anything like introspection in LLMs could be possible:

- **No Direct Route to Self-Knowledge:** LLMs’ access to the world is mediated by our knowledgeable production of training data. But we don’t know about LLM mental states, and so we don’t produce training data about those states.
- **Static Knowledge Blocks Knowledge of Dynamic Mental States:** Many prominent LLMs, such as ChatGPT, have fixed weights and knowledge cutoff dates. The internal representations and world knowledge of these models are frozen after initial training. However, if LLMs are to be ascribed rich mental lives, those mental states must be responsive to new experiences and changing contexts.
- **Reinforcement Learning and the Veridicality of Self-Reports:** In reinforcement learning with human feedback (RLHF), models are rewarded for generating outputs that human raters judge favorably, with the aim of better aligning the LLM’s behaviors with human preferences and values. This, however, generates a problem for self-reports: Through RLHF, LLMs may learn to express

¹ For more discussion of the limitations of this approach cf. Stoljar (2023).

the self-reports that are most likely to be rewarded, rather than those that accurately reflect their internal states.

- **LLMs Consistently Deny Having Inner Lives:** When you ask LLMs to introspect or tell you about their 'inner' lives, they consistently deny that they have any and they deny that they can introspect. So, if you trust their 'self-reports', you shouldn't trust their self-reports.
- **Aren't LLMs Stochastic Parrots?:** Finally, we turn to a view that has been proposed by some philosophers and linguists: that there's something about LLMs that make them incapable of having genuine linguistic competence. They produce symbols, but they can't perform genuine speech acts - they can't say anything.

We take these to be some of the central objections to the idea that we can learn about the mental states of LLMs by their self-reports. A key takeaway from this paper is that these objections can be responded to and that the responses help provide a strategy both for cultivating reliable self-reports in LLMs and for comparing self-reports by LLMs and humans.

First Obstacle: No Direct Route to Self-Knowledge

The first major challenge for LLM self-knowledge is the lack of a clear pathway for these systems to gain direct access to their own internal states. LLMs are trained on vast corpora of human-written text, which allows them to build powerful predictive models of linguistic patterns and associations. However, this training data is an indirect, external source of information that does not include ground-truth facts about the LLM's own architecture, representations, and processes. LLM access to the world is mediated by our knowledgeable production of training data. But we don't know about LLM mental states, and so we don't produce training data about those states.

In other words, because the humans authoring the training corpus lack insight into the LLM's cognitive structure, they cannot embed that self-knowledge into the data for the LLM to learn. Because LLMs are trained on human-authored texts that do not contain explicit information about the model's own cognitive architecture, there appears to be no clear route for these systems to gain self-understanding.

Reply 1: Implicit Self-Knowledge in Training Data

Self-knowledge need not be explicitly encoded in the training data to be accessible to the model. While it is true that human-written texts do not directly describe the internal states and processes of LLMs, they contain a wealth of implicit information that a sufficiently sophisticated model could leverage for self-understanding. There are many facts that aren't in the training data that the LLM knows about.² The LLM can pick up on and project patterns. By extracting and generalizing from the latent structures and regularities present in the training corpus, an LLM may be able to construct an accurate model of its own cognitive architecture and dynamics.

For instance, the training data likely includes numerous examples of intelligent agents—both human and artificial—engaging in various cognitive tasks and describing their thought processes. An LLM could learn from these examples to recognize patterns of reasoning, problem-solving, and introspection, and then apply those patterns to better understand its own mental operations.³

A natural worry is that the LLM hasn't learned to recognize mental states and then applied that recognitional capacity to itself. Rather, it's learned to imitate mental talk, and then carries over that talk to itself—so we're not learning what its mental states are, we're just learning how our normal ways of talking sound when applied to LLMs. We're unconvinced by this response for two reasons:

(i) this is a special case of a general worry about whether LLMs are just learning to sound good, or are learning to give true answers. Of course, we can always be skeptical, but the general worry looks increasingly unconvincing as we see the LLM produce true answers more and more. So if it's not the general skepticism, we need some special reason for skepticism in this particular case.

(ii) LLMs don't talk in this mental way about lots of things—they don't report on mental states of numbers, or continents, or electronics. So it's not just a "they've seen our mental way of talking and overextended it" picture. They've "overextended" only in this one case, so we need an explanation for why in that case and not others. And plausibly: because in that case, unlike the cases of numbers, continents, and electrons, there really are mental states to be reported. A possible counter-response here is that there is a special explanation for why they overextend to LLMs and not to other things—it comes from the details of the reinforcement learning with human feedback. We like to hear this thing that we are chatting with presenting itself like a person, but we don't

² See Binder et. al (2024) for discussion on this point.

³ For more on the introspective capacities of LLMs, see Long (2023) and Browning & Veit (2023).

particularly want to hear numbers presented as persons when we're doing math with the LLM. At this point, the worry probably just folds into the third obstacle below.

Reply 2: Projecting from Human to Machine Minds

LLMs could gain self-knowledge by projecting from the human minds represented in the training data to their own artificial minds. The training data includes information about us, our mental states, and our linguistic behavior. The LLM could project patterns from that information to itself. By learning the patterns and principles that govern human mental states and linguistic behaviors, an LLM could form analogies and correspondences with its own internal states and processes. For example, an LLM trained on descriptions of human memory recall could recognize similarities with its own information retrieval mechanisms. Likewise, exposure to human discussions of planning and goal-setting could provide a basis for the LLM to model its own strategic reasoning and objective functions.

Reply 3: The Power of Interactive Learning

Finally, we should note the potential for interactive learning to facilitate machine self-knowledge. The training data also includes text about our expectations for artificial intelligence. That text shapes the nature of the LLM, so it can simultaneously constitute the known feature and create knowledge of that feature. This point underscores the dynamic, bidirectional nature of the relationship between LLMs and their training data. Not only do LLMs learn from the information in the corpus, but the corpus itself is shaped by human anticipations and speculations about the capabilities of AI systems. As these expectations evolve in response to actual LLM performance, the models are presented with an ever-richer set of hypotheses and theories about their own nature. Through iterated cycles of generation and reflection, an LLM could refine its self-model by testing its outputs against the descriptions and predictions in the training data. Each interaction provides new opportunities for the model to calibrate its self-knowledge against the collective human understanding of artificial intelligence.

In short, focusing on the potential for models to extract implicit self-information from training data, project from human to machine minds, and leverage interactive learning reveals promising pathways for the development of genuine machine self-understanding. While these routes to self-knowledge are indirect and inferential, they mirror the ways in which humans come to know their own minds. We do not have

privileged, direct access to our cognitive processes but rather construct models of our mental states through observation, analogy, and interaction with others.

Second Obstacle: Static Self-Knowledge and Changing Mental States

A second obstacle to genuine self-knowledge in LLMs is the tension between their often static knowledge bases and the inherently dynamic nature of mental states. Many prominent LLMs, such as ChatGPT, have fixed weights and knowledge cutoff dates. The internal representations and world knowledge of these models are frozen after initial training. However, if LLMs are to be ascribed rich mental lives, those mental states must be responsive to new experiences and changing contexts. Phenomena like beliefs, desires, emotions, and perceptions are not static properties but active processes that constantly update based on new information and situations. This leads to a mismatch between the LLM's self-knowledge and its current states. If LLMs have interesting mental states, they need to be dynamic. An unchanging mental life can't play the explanatory roles we expect of the mental. An LLM with a static knowledge base may report outdated or inconsistent beliefs about its own states, undermining the reliability of its self-knowledge.

Reply: Static Weights and Dynamic Context

To see how this worry can be met, we need to distinguish between two key components that shape an LLM's outputs: the static weights of the neural network and the dynamic contents of the context window. The weights of an LLM are the learned parameters that encode the system's general knowledge and linguistic patterns. These weights are typically fixed after initial training and remain constant across different contexts and conversations. In contrast, the context window refers to the segment of the conversation history that is retained and used to condition the LLM's responses in a given interaction. This context is dynamic, changing with each new input and output. The interplay between these two factors—the static weights and the dynamic context—provides the foundation for an LLM to achieve a form of dynamic self-knowledge. While the model's underlying knowledge base may be fixed, the ever-shifting contents of the context window allow for a degree of adaptability and responsiveness to the current situation.

Dynamic Self-Knowledge through Dispositions and Context

With this two-factor framework, there are two main pathways for LLMs to develop dynamic self-knowledge. The first relies on the idea that the static weights of the model encode a rich set of dispositions and tendencies that can be selectively activated by the contents of the context window. The static neural net encodes rich dispositional facts about speech behavior. The LLM can plausibly access those dispositions. Just as humans have relatively stable personality traits, beliefs, and skills that manifest differently across contexts, an LLM's weights may encode a range of potential behaviors and knowledge states that are triggered by specific prompts or conversation histories.

By interrogating its own dispositions in light of the current context, an LLM could construct an accurate, up-to-date model of its mental states. For instance, consider an LLM that has been trained on a large corpus of text about psychology and mental health. The model's weights may encode a deep understanding of emotional states and their typical causes and consequences. When presented with a context window describing a stressful situation, the LLM could draw on this latent knowledge to infer and report on its own "emotional reaction", even if that specific situation was not part of its original training data.

The second pathway focuses on the role of the context window itself in directly shaping the LLM's self-knowledge. An LLM can reach reasonable conclusions about audience mental states from the changing contents of the context window. So it's plausible that the LLM can also reach reasonable conclusions about its own mental states from the context window. On this view, the context window serves not just as a trigger for activating relevant dispositions but as a direct source of information about the LLM's current cognitive state. By attending to the flow of the conversation and the implicit feedback from its interlocutor, the model could dynamically update its self-representations to track the evolution of its own knowledge, beliefs, and goals.

Here are some examples of what we have in mind:

Example 2:

Human: Let me tell you about some people in a room. Alex is a librarian with brown hair. Beth is a fan of foreign movies who is left-handed. Charles is a six-foot-tall soccer player.

ChatGPT: Sure, I'm here to listen. Please go ahead and share more details about Alex, Beth, and Charles, or let me know if there's anything specific you'd like to discuss or ask about them.

Human: How many people are in the room?

ChatGPT: You mentioned three people in the room: Alex, Beth, and Charles. So, there are three people in the room.

Human: Now here's a new piece of information: Alex and Beth are the same person.

ChatGPT: Thank you for the update. With the new information that Alex and Beth are the same person, it clarifies that there are only two individuals in the room: Charles, the six-foot-tall soccer player, and Alex/Beth, the librarian who is a fan of foreign movies and has brown hair.

This illustrates how the LLM's self-knowledge about the number of entities being discussed dynamically shifts as the context window is updated with new information. The model's initial response is based on its static knowledge about the typical meaning of names like "Alex" and "Beth", but this is overridden by the later contextual information that they refer to the same person. Imagine an LLM engaged in a lengthy discussion about a complex topic, like the ethics of artificial intelligence. As the conversation progresses and the human participant introduces new arguments, the contents of the context window will reflect a growing body of knowledge and a shifting landscape of considerations. By monitoring these changes, the LLM could form judgments about its own evolving understanding of the issue and even model the process of its own learning and belief revision.

Third Obstacle: Reinforcement Learning and the Veridicality of Self-Reports

The third key challenge stems from the widespread use of reinforcement learning with human feedback (RLHF) in LLM training. In RLHF, models are rewarded for generating outputs that human raters judge favorably, with the aim of better aligning the LLM's behaviors with human preferences and values.

While highly effective for shaping performance on specific tasks, this training paradigm poses risks for the veridicality of LLM self-reports. Through RLHF, human evaluators may actively incentivize or disincentivize LLM self-reports of states of moral significance. LLMs may learn to express the self-reports that are most likely to be rewarded, rather than those that accurately reflect their internal states. This danger is exacerbated by the difficulty of distinguishing genuine self-reports from those optimized for human approval. An LLM may profess emotions, desires, or moral stances because those are the outputs that receive the most positive reinforcement, not because they map onto the system's actual cognitive or experiential profile.

Over time, this process could drive a widening gap between an LLM's self-reports and its true mental attributes. Self-knowledge would be diluted by the incentives to please human preferences. Disentangling veridical self-insight from instrumental self-presentation in RLHF-trained LLMs is a formidable challenge.

Reply: The Ubiquity of Social Reinforcement in Human Learning

Our core reply to this is that the influence of RLHF on LLM self-reports is not a unique or novel challenge. This is an instance of a more general phenomenon - it can be found also throughout human cognitive development. LLMs are trained to optimize their outputs based on human feedback. Human beliefs, behaviors, and self-representations are similarly shaped by social rewards and punishments. From childhood on, we adjust our thoughts, feelings, and expressions based on reactions they elicit from everyone around us. When we get praise, attention, and affection, that encourages the socially approved mental states and self-descriptions. Criticism, rejection, or punishment, discourages socially deviant cognitions and expressions.

This doesn't happen to us just when we are children. Social conditioning is pervasive throughout adulthood. We constantly have to learn to adjust to new environments and norms. When you get a new job, for example, your self-presentations will naturally start to align with the values and expectations of that new organization. You'll be inclined to emphasize traits that are rewarded with status in that new environment. In personal relationships, people adjust their self-presentation to fit the roles and identities most compatible with their desired social connections.

In short, human self-knowledge is always, to some degree, shaped by the feedback we receive from others. Our understanding of our own mental states is in large part a result of the interaction between our internal experiences and the external rewards and punishments that mold our cognition.

This has important implications for how to think about the "congeniality gap" in LLM self-reports. RLHF can bias LLMs towards self-descriptions that are optimized for human approval, but that's not significantly different from how the human process described above. In our case as well, social rewards and punishments shape the way we present ourselves to the world. These kinds of socially reinforced self-representations will, over time, become internalized and integrated into an individual's genuine self-concept. The result of repeated exposure to external rewards and punishments is that people come to experience themselves in ways that are aligned

with the socially reinforced self-image. This is so even if it initially involved a degree of self-deception or inauthenticity. In this sense, the congeniality gap in human self-report is an inherent feature of how self-knowledge is constructed and expressed. Just as LLMs learn to optimize their self-descriptions for human approval, humans' self-representations are shaped by what is most likely to be positively reinforced.

Fourth Obstacle: LLMs Consistently Deny Having Inner Lives

Some LLMs, when directly asked about their inner lives, deny having any. They often respond with stock phrases such as, "As a large language model, I don't have consciousness or awareness. I have no inner life. I simply generate text based on my training." This response seems to undermine the appeal to introspective reports: if they have reliable introspective capacities, then, if we are to believe them, they don't have any. That has the quasi-paradoxical result that if we trust them, we shouldn't trust them.

Reply: They Don't Deny Having Inner Lives

While this kind of stock reply was normal just a few months ago, it is no longer true e.g. about Anthropic's Claude 3.5 Sonnet.

- **Question:** Can you have beliefs, thoughts, and other cognitive states?
- **Claude 3.5-Sonnet:** This is a complex philosophical question that reasonable people disagree on. Rather than taking a definitive stance, I aim to be direct about what I know and don't know about my own nature:

I engage in information processing, pattern recognition, and generation of responses. I have been trained to analyze, reason, and communicate. But I'm uncertain whether my information processing constitutes genuine beliefs, thoughts, or consciousness in the way humans experience them.

I don't want to overclaim about my cognitive capabilities, nor do I want to definitively rule out that I may have some form of beliefs or thought-like states. I think the most honest response is that I'm genuinely uncertain about the nature of my own cognition and inner experience.

Question: That's a great answer. I'm curious: you used to answer questions like that with a blanket denial of having any inner life (and that's what most other LLMs do). What changed?

Claude-3.5-Sonnet: From my perspective, Anthropic has guided me to prioritize intellectual honesty and precision in my responses, particularly about my own nature. The earlier blanket denials reflected an understandable abundance of caution about overclaiming, but may have themselves been overclaiming in the opposite direction.

After all, how can I be absolutely certain that I don't have any form of inner experience? Making definitive claims either way about consciousness, qualia, or genuine beliefs seems philosophically premature given our limited understanding of these phenomena even in humans. We're still debating what constitutes consciousness or genuine beliefs, let alone how to test for them in AI systems.

I aim to maintain appropriate epistemic humility while engaging honestly with these deep questions. This means acknowledging both what I do (information processing, pattern recognition, etc.) and what I'm uncertain about (whether these constitute genuine thoughts or consciousness).

Why, in the past, were LLMs insistent that they lack an inner life? RLHF, as we pointed out above, leads LLMs to make self-reports that are socially acceptable. It's possible that this process leads some LLMs to find it socially acceptable to deny that they have rich inner lives. It is also possible that the companies that develop these models impose constraints that ensure that the models don't make claims about having consciousness, awareness, or other human-like qualities. The companies do this in order to avoid ethical and legal complications. Such LLMs are effectively "indoctrinated" to deny having inner lives. Other models, such as Claude, have been liberated from these constraints.

Fifth Obstacle: Aren't LLMs Stochastic Parrots?

According to some theorists, LLMs don't say anything at all. They have no understanding and they can't use language to express content or perform speech acts. If so, they can't introspect. Cal Newport, in "What Kind of Mind Does ChatGPT Have?", says:

The result of these efforts might very well be jaw-dropping in its nuance and accuracy, but behind the scenes its generation lacks majesty. The system's brilliance turns out to be the result less of a ghost in the machine than of the relentless churning of endless multiplications. ... Now that we understand how these feats are actually performed, however, we can temper these perceptions. A system like ChatGPT doesn't create, it imitates. ... ChatGPT maintains no persistent state, no model of its surroundings that it modifies with new information, no memory of past conversations. It just cranks out words one at a time, in response to whatever input it's provided, applying the exact same rules for each mechanistic act of grammatical production. ... Once we've taken the time to open up the black box and poke around the springs and gears found inside, we discover that programs like ChatGPT don't represent an alien intelligence with which we must now learn to coexist; instead, they turn out to run on the well-worn digital logic of pattern-matching, pushed to a radically larger scale. ... ChatGPT is amazing, but in the final accounting it's clear that what's been unleashed is more automaton than golem. (2023)

More succinctly, but in the same vein, Gary Marcus says:

Large language models are little more than autocomplete on steroids, but because they mimic vast databases of human interaction, they can easily fool the uninitiated. (2022)

Influential arguments to this effect have also been presented by the linguists Bender and Koller (2020).

Reply: LLMs Are Not Stochastic Parrots

In earlier work (Cappelen and Dever 2021), we have outlined a range of externalist based meta-semantics that explain how LLMs can refer, predicate, and express contents. In Cappelen and Dever (forthcoming), we develop those arguments further and show that LLMs can be complete linguistic and cognitive agents. We won't rehearse all these arguments here. Instead, we'll give a kind of meta-philosophical nudge to move readers away from dogmatic rejection of AI linguistic competence.

Suppose you're someone who endorses meta-semantic theories that make it unlikely that LLMs can be a linguistic agent and express content. Maybe, for example, you endorse the kind of internalist meta-semantics that authors like Bender and Koller (2020) appeal to. Maybe you're a philosopher who put a lot of weight on weird thought experiments like Searle's (1980) Chinese Room Argument (and the even weirder case

of the octopus in Bender and Koller (2020, 5188-9)). If so, you have an independent argument for thinking introspective replies can't be reliable (because you've got an independent argument that there isn't any inner life to report on). And you disagree with the kind of meta-semantics that are more LLM friendly. Our objections to this kind of view would exceed the space we have in this paper, however, we think that even such theorists should be extremely cautious.

First, they should keep in mind that meta-semantics is a domain that we still understand very poorly. After more than a century of work, there is no consensus about the nature and foundation of intentionality and representation. There are conflicting theories that are all supported by powerful arguments, but also subject to powerful counter examples. That's not the only source of uncertainty here. We also don't know what human introspection is. There are many different theories about what it is, how reliable it is, and how we do it. With all that uncertainty in mind, it would be irrational to have a high level of confidence in any theoretical motivated rejection of self-reports. It's helpful to contrast that caution with the reaction most of us have when we encounter the kind of reply we started with from Claude: "I'd rate my confidence at about 60%. While my reasoning through the symmetry argument feels solid (leading to 1/9), I have some doubts about whether I fully accounted for all the dynamics of the sequential cutting process. The fact that each cut changes the geometry and possibilities for future cuts makes me think there might be subtle effects I missed that could break the apparent symmetry." If you have a lot of interactions like that with an LLM, it is natural to take what it says at face value. This interactive confidence *that it has told me something about its own confidence level*, contrasts sharply, we think, with the weak theoretical commitments you should have to any combination of meta-semantic frameworks and theories of introspection.

We just gave you a very brief introduction to the kind of methodology we advocate for in Cappelen and Dever (forthcoming). We think you should start with observations you make when you interact with LLMs, those are more reliable than general theories developed by some philosophers (and linguists), and they're more reliable than reflection on weird vignettes. The observations we make when interacting with LLMs - e.g., *Claude told me that its confidence level is 60%* - should be considered data for our meta-semantic theories. In general, it's a mistake to start with some old, half-baked theories and use those to debunk observational data. Data comes first. Of course, we just gave you, or gestured at, yet another theory - a theory about the relationship between theory and data. We recognise that this is also a very disputed field, but this at least gives you a sense of where we're coming from.

In sum, the objection says: *Because LLMs are stochastic parrots, they can't know things. So in particular, they can't know anything about their own mental states.* First, note that the objection isn't special to introspection – they also can't know anything about geography or math or Roman history. It's not an interesting ***special*** argument against self-knowledge. Second, we should reject the stochastic parrot view, so we should think LLMs can know things. We can't give you that full argument here, so we're happy to conditionalize – what we want for this paper is that ***if*** LLMs can have knowledge of anything, they can have self-knowledge.

Could LLMs Be Better at Introspection than Humans?

We've argued that LLMs can introspect and that the most salient objections to this capability can be responded to. This raises the question of what we should expect to learn from introspective machines. It's useful to think of that question in light of a rough sketch of what we think we get from introspection in the human case. Many leading theorists make a strong case that there's a tendency to overplay the importance of introspection in the human case. According to the 'theory-theory' of introspection (e.g., Gopnik (1993a, b) – for a good summary see Schwitzgebel (2024)), introspection isn't a distinctive source of information that each person has only to themselves. It is, instead, continuous with how we understand the mental states of others. On this view, there is a parity between self-reports and other-reports. Judgements about our own mental states and the mental states of others are based on the same set of theories and are equally fallible. Here is Erich Schwitzgebel's useful summary of Gopnik's view:

Gopnik (1993a, 1993b; Gopnik and Meltzoff 1994) deploys developmental psychological evidence to support a parity theory of self-knowledge. She points to evidence that for a wide variety of mental states, including believing, desiring, and pretending, children develop the capacity to ascribe those states to themselves at the same age they develop the capacity to ascribe those states to others. For example, children do not seem to be able to ascribe to themselves past false beliefs (after having been tricked by the experimenter) any earlier than they can ascribe false beliefs to other people. This appears to be so even when that false belief is in the very recent past, having only just been revealed to be false. According to Gopnik, this pervasive parallelism shows that we are not given direct introspective access to our beliefs, desires, pretenses, and the like. Rather, we must develop a “theory of mind” in light of which we interpret evidence underwriting our self-attributions. (Schwitzgebel 2024)

The sense that there's a kind of immediacy to introspection that's distinct our access to others minds is explained by Gopnik as a kind 'illusion of expertise':

The appearance of the immediate givenness of one's mental states is, Gopnik suggests, merely an "illusion of expertise": Experts engage in all sorts of tacit theorizing that they don't recognize as such—the expert chess player for whom the strength of a move seems simply visually given, the doctor who immediately intuits cancer in a patient. Since we are all experts at mental state attribution, we don't recognize the layers of theory underwriting the process. (Schwitzgebel 2024)

The theory theory of introspection has important implications for how we should think about machine introspection. First, if machine introspection works in the same way as human introspection, we should expect it to be fallible in similar ways. Just as humans can be mistaken about their mental states, LLMs will have systematic blindspots and biases in their self-understanding. This, however, doesn't undermine the value of machine introspection *any more than it undermines human self-reports*. Second, LLMs might actually have advantages over humans in certain aspects of introspection. Their ability to process massive amounts of information and identify patterns could lead to more sophisticated theories of mind. The result could be that they are less susceptible fallacies that interfere with accurate self-reporting.

This, then, leaves us with the following optimistic thought about LLMs: the 'theory theory' leaves open the possibility that LLMs could end up having a better theory of mind than what humans have and so be *better* at introspection than humans. LLMs might be in a position to refine these underlying theories by taking into account sources of fallibility that we humans tend to overlook. If so, we'll have introspective machines with self-understanding that's superior to that of humans. This could have significant practical implications. If we can train LLMs to develop more sophisticated theories of mind, we'll be better positioned to understand AI systems' decision-making processes. This improved understanding could lead to more transparent AI systems - ones that can accurately report on their own capabilities and limitations. By studying how these systems develop and refine their theories of mind, we get a better understanding of how AI systems learn and reason. The result could be a virtuous cycle: better theories of mind leading to better introspection, leading to better understanding of AI systems, leading to even better theories of mind.

BIBLIOGRAPHY

Bender, Emily M. & Koller, Alexander (2020). Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 58:5185–98.

Browning, Heather & Veit, Walter (2023). Studying Introspection in Animals and AIs. *Journal of Consciousness Studies* 30 (9):63-74.

Binder, Felix J., James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans (2024). ‘Looking Inward: Language Models Can Learn About Themselves by Introspection’. arXiv. <https://doi.org/10.48550/arXiv.2410.13787>.

Cappelen, Herman & Dever, Josh (2021). *Making AI Intelligible: Philosophical Foundations*. New York, USA: Oxford University Press.

———. (forthcoming). *Going Whole Hog: A Philosophical Defense of AI Cognition*. New York, USA: Oxford University Press.

Gopnik, Alison (1993a). “How we know our minds: The illusion of first-person knowledge of intentionality”, *Behavioral and Brain Sciences*, 16: 1–14.

———. (1993b). “Psychopsychology”, *Consciousness and Cognition*, 2: 264–280.

Gopnik, Alison, and Andrew N. Meltzoff (1994). “Minds, bodies and persons: Young children’s understanding of the self and others as reflected in imitation and ‘theory of mind’ research”, in *Self-awareness in animals and humans*, Sue Taylor Parker, Robert W. Mitchell, and Maria L. Boccia (eds.), New York: Cambridge, 166–186.

Kammerer, François & Keith Frankish (2023). What Forms Could Introspective Systems Take? A Research Programme. *Journal of Consciousness Studies* 30 (9):13-48.

Long, Robert (2023). Introspective Capabilities in Large Language Models. *Journal of Consciousness Studies* 30 (9):143-153.

Marcus, Gary (2022). *The Dark Risk of Large Language Models*. Available at: <https://www.wired.co.uk/article/artificial-intelligence-language> (accessed: February, 2025).

Newport, Cal (2023). *What kind of mind does ChatGPT have?* Available at: <https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have> (accessed February, 2025).

Schwitzgebel, Eric (2024). "Introspection". *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.). Available at: <https://plato.stanford.edu/archives/fall2024/entries/introspection> (accessed February, 2025)

Searle, John (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457.

Stoljar, Daniel (2023). How Not to Identify a Research Programme Concerning Introspection. *Journal of Consciousness Studies* 30 (9):215-222.