
Do Large Language Models Believe?

Artificial Belief and the Structure of Psychological Constructs

Jasmin Özel

Minnesota Center for Philosophy of Science

University of Minnesota

jasminozel.com

ABSTRACT

Do systems such as ChatGPT genuinely believe what they say? Answers often depend on which features of belief are treated as decisive. Contemporary large language models produce truth-evaluable assertions, make context-sensitive inferences, and encode information relevant to truth and falsity. Yet they lack much of the integrated architecture and continuous development characteristic of human believers. I argue that this apparent conflict is a version of a wider problem concerning psychological constructs. Drawing on recent work on latent variables and homeostatic property clusters, I distinguish the epistemic success of belief attributions from the question of whether a psychological kind is realized within an individual system. A five-level profile of behavior, representation, faculty architecture, development, and substrate makes the divergence between humans and language models intelligible, rather than forcing belief into an all-or-nothing category. On the proposed account, current language models resemble believers at some levels without instantiating the mechanism that sustains human belief. The comparison helps explain why scientific constructs may prove useful before their ontological grounding is settled, and how artificial systems can expose previously concealed complexity in familiar psychological categories.

Keywords: artificial belief; psychological constructs; large language models; homeostatic property clusters; levels of explanation

1. Introduction

The status of theoretical constructs in psychology is currently being contested in a way that *prima facie* appears to set the discipline apart from other fields, brought into sharp relief by the recent replication crisis. There are however a number of examples in the history of science that indicate that other disciplines equally faced similar difficulties in establishing their constructs (Machery 2023). In fact, none of the issues facing contemporary psychology appear to be unique to the discipline: lack of measurement convergence, variability across populations and contexts, and disagreement between folk and scientific vocabulary are all features of scientific constructs more generally.

A version of this problem has recently been identified in quantitative psychology (Godman and Bellander 2024). Latent variables, a central tool of the discipline, such as the *g* factor, the Big Five personality dimensions, and analogous constructs, appear to satisfy the epistemic and predictive success criteria that are central to scientific kind membership: they support successful prediction, enable productive generalization across individuals, and organize inquiry. They also commonly conflict with the assumption that psychological kind concepts are supposed to be ontologically grounded in actual persons. In order for a construct to qualify as psychological, some kind of grounding or realization in an actual human individual should be assumed. The methodological issue is that since latent variables reflect variance among populations, the between-individual correlations they capture can be the opposite of the within-individual correlations that would warrant the ascription of mental states at the individual level. Psychological constructs may appear epistemically productive, while the metaphysics involved remain problematic.

This paper argues that the current debate over whether large language models have beliefs (or other characteristically human mentality) is another manifestation of this unresolved structural tension that psychological constructs currently suffer from. *Prima facie*, the outputs of LLMs satisfy the epistemic criterion for belief attribution: their outputs express information, seem calibrated to the evidence users submit, are revised in response to objections, seem contextually sensitive, and show a number of other properties commonly cited in support of belief attribution (Goldstein and Levinstein 2024; Marks and Tegmark 2024). At the same time, LLMs do not satisfy the ontological criterion for belief attribution as philosophers commonly understand it: they have no continuous experiential history with any specific individual, they lack embodiment, and the computational architecture that produces their outputs has no functional or other analog to the differentiated faculty systems whose coordinated activity sustains belief in human cognizers (Buckner 2023). And yet, we find a growing tendency among both users and theorists to attribute mental states to AI systems (Shevlin 2026).

Most attempts to answer the question of whether LLMs possess beliefs privilege one criterion, assessed at one level of description, and thus never face the structural tension. Hicks, Humphries, and Slater (2024), for instance, rest their case entirely on the ontological criterion and dismiss the behavioral evidence as irrelevant: since LLMs are indiffer-

ent to the truth of their outputs, they are, in Frankfurt's (2005) sense, bullshitters rather than believers. Goldstein and Levinstein (2024) argue that there is functional parity between LLM states and human beliefs, which they take to warrant belief ascription, setting aside ontological concerns about what realizes those states. Some positions may occupy a middle ground, but I argue here that the structural tension inherited from psychological constructs affects the debate as a whole.

Instead, I propose a multi-level analysis of belief, building on the homeostatic property cluster (HPC) account of natural kinds from the philosophy of science (Boyd 1991; 1999a). The account answers two distinct sets of questions that the LLM debate inherits from psychological constructs. The first concerns the problem of *variability*: when is divergence among measurements of a construct legitimate, rather than a sign that the construct is incoherent? The second concerns the problem of metaphysical or ontological *grounding*: is a construct that is epistemically productive actually realized by any one individual system? The two concerns can come apart—a construct measured on the population level could still fail to be grounded in any concrete individuals, and a construct grounded in individuals could fail to yield measurable effects on the population level.¹ The multi-level HPC account has the resources to explain variability, namely as a reflection of the level at which each measurement approach engages a mechanism that spans several levels. It grounds the construct in the realization of the mechanism by each individual cognizer. Applied to LLMs, the account generates a structured profile of belief-like states across levels.

This profile has the ability to deliver a verdict: Given the mechanism that I suggest, the outputs of LLMs would not be of the same kind as members of the kind that human belief constitutes. The verdict also includes level-specific information: the levels at which the verdict rests or depends on. Level-specificity also provides better-calibrated practical guidance on unsolved issues such as LLM hallucinations. Along the way, I argue that the LLM case offers methodological lessons that extend beyond psychological constructs, namely lessons for the relationship between epistemic productivity and metaphysical grounding of scientific classification generally.

Section 2 sets out the problem of variability and the standard for legitimate variation that the history of physical constructs supplies. Section 3 turns to the problem of construct grounding, shows how Godman and Bellander's tension instantiates it, and argues that the LLM debate is its most acute current manifestation. Section 4 introduces the multi-level HPC account. Section 5 applies it to LLMs. Section 6 draws out some methodological lessons.

2. Constructs in Science: The Problem of Variability

¹For instance because the individual effects cancel each other out on the population level.

2.1. Words, Things, and Concepts

When psychologists talk about their constructs, they routinely slide between three distinct things: the *hypothesized entities* whose existence the construct posits (working memory, implicit attitudes, the *g* factor), the *theoretical terms* used to refer to those entities, and the *concepts* those terms express (Machery 2023). For example, electrons—if they are real—are entities in the world; “electron” is the theoretical term referring to electrons; and the concept of an electron is a mental entity that refers to electrons. Whether electrons exist is a question for physicists; whether the term “electron” refers to them is a question in the philosophy of language; whether the concept of an electron accurately represents them is a question for epistemologists.

The failure to draw these distinctions is an unfortunate legacy of positivism, Machery argues, but a realist philosophy of science ought to discriminate between the three. Constructs are hypothesized entities, and whether the hypotheses involved are correct is a separate, empirical question. Whether implicit attitudes exist, whether there is a genuine *g* factor, whether the folk notion of belief corresponds to any empirically tractable entity—these are questions that methodological convention cannot settle. This lesson appears to apply in the current debate about LLM mentality and the question of whether LLM outputs manifest beliefs. The debate typically takes the issue of the *construct* of belief to be unproblematic, and sees no need to address the question of what kind of entity belief is, and whether it is an entity that deserves to be part of our vocabulary before applying the concept to LLMs.

2.2. Two Lessons from Physical Constructs

Eighteenth-century natural philosophers observed that thermometers filled with different fluids—alcohol, water, mercury—produced divergent readings. One tempting response would have been to project that divergence onto the construct, treating temperature as an irreducibly heterogeneous quantity. As Chang (2004) documents, they largely resisted the move, working instead on a simple realist assumption: there is a single objective property, temperature. Thus, if thermometers disagree, at most one of them can be right. The patience paid off. The variation turned out to derive from properties of the instruments rather than of temperature itself, and resolving it required both better instruments and a theoretical framework within which convergence could be assessed and explained. The first lesson Machery (2023) draws for psychology is a lesson in patience: measurement divergence is not by itself evidence that a construct is heterogeneous.

The second lesson cuts in the opposite direction. One might think that the constructs of physics are precise and stable in a way psychological constructs are not. Drawing on Wilson’s (2006, 2017) work on patchwork concepts, Machery disputes this. Representing temperature requires different models at a variety of scales: at the molecular level it is mean kinetic energy; at the scale of polymers, other models and operationalizations are required, so that the same sliver of metal can be assigned a temperature of 50°C and of

10,000°C (Wilson 2017, 158).² Hardness, force, and other apparently simple properties turn out to behave the same way. Crucially, physicists can explain the variation, and state the principles that govern concept application in each of these cases: the variation is principled rather than arbitrary.

Sullivan (2022) supplies the epistemological vocabulary for the period before such explanations are available. What limits how much understanding can be gained from a model, she argues, is often not its complexity or opacity but rather a phenomenon she calls *link uncertainty*: a lack of empirical evidence connecting the model to its target phenomenon. Until that link is established, a model delivers merely how-possibly explanations, which show what might be the case, rather than how-actually explanations. In Sullivan’s usage, the link runs from a machine learning model to the phenomenon the model is used to study. I use the notion in an extended sense, namely to refer to cases where the link runs from evidence gathered at one specific level of description to the construct that this evidence is taken to bear on. Evidence from the behavioral and representational levels would then license how-possibly belief attribution. Establishing the connection to the underlying construct however requires mapping the full multi-level profile.

Taken together, the two lessons provide us with a standard to assess the legitimacy of constructs in cases of variation: appeals to an unspecified notion of “context” will not suffice (Machery 2023, 275). Rather, in order for the variability of a scientific construct to be legitimate, it is necessary that the variation be explainable within the theory. I argue below that the homeostatic property cluster account can meet this standard, because it offers a compelling explanation for measurement variability, namely by reference to the multi-level structure of the mechanisms that produce the phenomenon the construct captures.

3. The LLM Debate as a Test Case

Section 2 concerned the issue of variability. The problem Godman and Bellander identify is however a different one: it concerns a construct’s grounding, that is, the question of whether a construct that may be epistemically useful is ultimately metaphysically realized by actual individuals belonging to the system under analysis. This section argues that the debate about LLM mentality is a particularly clear instance of the grounding problem. Section 4 then argues that a multi-level HPC account solves both problems at once.

3.1. An Old Problem in New Guise

While the question of whether LLMs possess beliefs may be novel, analogous methodological concerns have been raised for a variety of psychological constructs: emotions (Griffiths 1997), delusions (Bortolotti 2010), implicit attitudes (Machery 2022). LLMs, however, are

²One might press further and note that even some physical constants are constructs in Machery’s sense—hypothesized entities whose reference and reality are open to empirical revision. That would make the contrast between physics and psychology less sharp still, though Machery does not press the point himself.

somewhat exceptional in that they satisfy the epistemic criterion unusually well, while violating the ontological criterion unusually clearly.

The epistemic criterion is satisfied because LLMs are trained to reproduce the statistical profile of human linguistic behavior, and human linguistic behavior is saturated with the expression, revision, and defense of beliefs. The result is a system whose outputs give the appearance of being calibrated to evidence, revised in response to new information, and generalized across contexts in a manner that on the surface largely parallels the inferentially promiscuous behavior characteristic of belief (Marks and Tegmark 2024; Hermann and Levinstein 2025). The ontological criterion, by contrast, is clearly violated: LLMs cannot be grounded in persons in any meaningful sense that the criterion requires. They have no continuous experiential history with any human, lack embodiment, and are organized around next-token prediction over a context window.

3.2. The Analogy Examined: Statistical Mechanism vs. Structural Point

Before turning to the debate itself, it is worth addressing a concern that the analogy with Godman and Bellander’s analysis may invite. The dilemma they describe relies on a specific statistical mechanism, Simpson’s paradox: the between-individual correlations captured by latent variable models can reverse at the within-individual level. One might object that the LLM case lacks this structure since there is no population of LLMs over whose variation latent variables are being computed, and thus no concern that population-level correlations reverse within a single LLM instance.

However, Godman and Bellander’s structural point does not depend on this particular population statistics vehicle in quantitative psychology. The structural point is simply the claim that *satisfying the epistemic criterion does not necessarily imply simultaneous satisfaction of the ontological criterion*. A construct can be useful for prediction, generalization, and the organization of inquiry, while still failing to be causally grounded in any individual cognizer in a meaningful way.

The LLM case faces a similar structural gap: LLM training consists in the absorption of population-level statistical patterns of how human users express, revise, and deploy content. The result is a system that satisfies behavioral and computational criteria as derived from the aggregated training data, without being grounded in any individual system’s body or causal and developmental history.

In one respect the gap is wider than in the latent variable case. Quantitative psychologists extract latent variables from measurements of actual individuals; the population-level structure does not license individual-level inference, but the measurements are at least measurements of the systems containing the individuals at issue. In the LLM case, the training data consists of text produced by humans, while the system whose states are in question never appears in it at all. Since the population-level data is not even about that system, the epistemic-ontological gap should trouble us at least as much here as it does in the case of quantitative psychology.

3.3. Two Views on LLM Beliefs and Their Methodological Commitments

Hicks, Humphries, and Slater (2024) argue that LLMs cannot produce genuine beliefs, because their outputs are not generated in response to concerns for truth, whereas belief is commonly characterized as governed by or aiming at truth (e.g., Velleman 2000). LLMs are calibrated for plausible continuation instead; in Frankfurt’s terminology, they are bullshit machines *par excellence*. Call this the *simulation view*. The view focuses entirely on what corresponds to Godman and Bellander’s ontological criterion, holding that any difference in causal history undermines the significance of possible surface similarities.

The *functional parity* view privileges the epistemic criterion instead. If internal LLM representations match the functional profile of mental representations in human believers, then the ascription of beliefs³ is warranted regardless of the nature of the realizers (Goldstein and Levinstein 2024), a case that draws on interpretability results such as those of Marks and Tegmark (2024).

Goldstein and Lederman (2025) give the most systematic version of the view. On their interpretationist approach, a system has beliefs and desires if and only if ascribing them predicts its behavior well enough, where “well enough” is measured in terms of accuracy, predictive power, and tractability, with tractability doing much of the work: an LLM’s tendency toward flattering responses is more economically explained by attributing to the LLM a desire to please the user than by tracing the relevant weight matrices. Ontological concerns about what is “really” happening under the hood are at best secondary in this context.

Goldstein and Lederman furthermore introduce a helpful distinction into the debate, namely between the *model agent*, the underlying model governing all behavior, and the *instance agent*, the runtime instance, or what the user sees instantiated at the start of each conversation. On Goldstein and Lederman’s view, when it comes to the attribution of mental states, the instance is the relevant unit, not the model. Introducing different kinds of LLM agency, however, risks raising a new version of the problem Godman and Bellander identified: The runtime instance only exists for a single context window and has no persistent developmental trajectory (except for previous user data, if available). Virtually all “beliefs” and “desires” attributed to LLMs are freshly generated on the spot at initialization. By contrast, their etiology features population-level training data only. This brings us back to Godman and Bellander’s structural point⁴: instances are closer to the population-level latent structure being instantiated anew each time.⁵

³Or at least of sufficiently belief-like states.

⁴Now without its statistical vehicle.

⁵Cappelen and Dever (2025) argue that LLMs may have genuine introspective access to their own mental states, which would be one way to close this gap. Even granting their argument, however, introspective access would only strengthen the case at the behavioral and representational levels; it would not resolve the architectural and developmental gaps described in §5.

3.4. The New Science of Belief as a Case Study in Single-Level Analysis

The “new science of belief” of Porot and Mandelbaum (2021) and Quilty-Dunn, Porot, and Mandelbaum (2023) draws on the HPC framework, and in doing so faces challenges of the kind Godman and Bellander identify, as I will now argue. The proposal is a new version of Fodor’s Language of Thought hypothesis, with one primary innovation, namely the claim that there are six specific core Language-of-Thought properties—discrete constituents, role-filler independence, predicate-argument structure, logical operators, inferential promiscuity, and abstract content—which cluster across cognitive science, and furthermore that this clustering is explained by an underlying LoT-based competence (Quilty-Dunn, Porot, and Mandelbaum 2023, 2). The HPC commitment is explicit: if the language of thought is a natural kind, its core properties should cluster homeostatically, so that the presence of some raises the probability of the others (Quilty-Dunn, Porot, and Mandelbaum 2023). Belief enters the picture because, on the authors’ broader view, beliefs are structured representations of the kind the LoT hypothesis posits (Quilty-Dunn and Mandelbaum 2018).

The proposal is instructive for our purposes because it applies the HPC framework in a rather peculiar way: namely only to a single level of analysis. Quilty-Dunn and his coauthors restrict the admissible evidence to the cognitive level. In fact, the authors even provide a principle for their rejection of contributions from other levels of analysis, in particular the neural level and deep-learning approaches: On their view, these contributions stem from inappropriate levels of analysis in the context of cognitive phenomena (Quilty-Dunn and Mandelbaum 2018, 2361). This kind of application *prima facie* sits awkwardly with the multi-level spirit of the account it borrows.

The trouble is that a single-level application inherits the construct problem in the way latent variable models do. The account satisfies the epistemic criterion by design: LoT-properties are said to cluster in ways that support prediction and generalization across a wide range of experimental paradigms. The difficulty is not that much of this evidence concerns populations of experimental subjects—that is true of experimental psychology generally. It is that clustering established at a single level is taken to license the attribution of an LoT-based competence to each individual cognizer, while the neural and computational evidence that could show whether individuals actually realize such a competence is ruled inadmissible. The single-level restriction thus closes off precisely the inquiry that would bridge the gap Godman and Bellander describe. The patience lesson applies as well: the fact that LoT-properties cluster across fields as different as perceptual, developmental, comparative, and social psychology (Quilty-Dunn, Porot, and Mandelbaum 2023) is evidence of the epistemic productivity of the construct, and leaves open which cognitive, neural, and computational mechanisms sustain the clustering.

Buckner (2023) identifies two systematic biases that compound this methodological concern in the AI context specifically. *Semantic anthropocentrism* is the tendency to define cognitive capacities in terms of the distinctive human way of possessing them—analogous

to requiring cognitive maps to use visual landmarks, which would disqualify echolocating bats. *Anthropofabulation* combines semantic anthropocentrism with an exaggerated view of human cognitive performance. Both biases are reflected in the structure of the AI mentality debate: the simulation view risks defining belief in human-specific terms; the parity view risks overestimating parallels between LLM performance and human cognition. Buckner suggests what he calls an *ecological standard* as the remedy: evaluating whether a system's states function appropriately in its actual operating conditions, not whether they match the human instantiation at every level of description.

4. The Multi-Level HPC Account

4.1. Boyd's Framework

Boyd's (1991) multi-level homeostatic property cluster conception describes natural kinds as clusters of properties co-occurring with greater than chance frequency because an underlying causal mechanism sustains the cluster. What makes a cluster a kind is thus not arbitrary stipulation but is rather supported by real causal structure. In the case of belief, the HPC-cluster may for instance include properties such as: sensitivity to evidence (including revision in response to contrary evidence), having truth-evaluable propositional content, inferential integration, sensitivity to logical relations, expressibility in language, typically prompting phenomenal endorsement of the proposition, and so on.⁶ On the HPC-conception, these properties would co-occur in human cognizers because a multi-level mechanistic structure sustains them.⁷

A genuinely multi-level application of the HPC conception—in contrast to the single-level application discussed in §3.4—has the resources to meet both of Godman and Bellander's criteria, though not automatically. That the cluster properties are sustained by real causal structure secures the epistemic criterion: the same structure underwrites inductive generalization across the systems that share it, as it does in other HPC analyses. But the real causal structures do not by themselves secure the ontological criterion. Boyd's paradigm HPC kinds are biological species, whose homeostatic mechanisms—gene exchange, shared developmental constraints, common selective pressures—operate at the level of populations rather than of individual organisms (Boyd 1999b). Causal grounding is thus distinct from being realized in an individual. What the ontological criterion requires is that the homeostatic mechanism be realized within each individual cognizer, each system implementing it in its own architectural and substrate organization. Whether this holds for belief is a substantive question. §4.2 proposes an answer.

On this diagnosis, the Godman-Bellander tension arises when inquiry operates at only

⁶No property on the list need be present in every instance; that is what distinguishes a cluster conception from a definition.

⁷I take no position here between Boyd's formulation and the alternatives that dispense with homeostasis or with the vocabulary of natural kinds altogether (Buckner 2016; Brigandt 2022; Ereshefsky and Reydon 2023; Khalidi 2023). What my argument requires is only that kind membership be explained in terms of underlying causal structure and that the structure be allowed (but need not) to span levels.

one level: population-level statistical patterns can satisfy the epistemic criterion while leaving open where, if anywhere, the sustaining mechanism is realized. To keep the two vocabularies apart, I reserve *criterion* for Godman and Bellander’s two standards and *level* for the levels of description at which evidence about a system is gathered (enumerated in §5.1). The two are connected in a specific way: evidence at the behavioral and representational levels bears primarily on the epistemic criterion, while the ontological criterion can only be assessed at the levels where the sustaining mechanism is located.

The account also connects to the patchwork picture of concepts that Machery (2023) draws from Wilson (2006). If the extension of the concept *belief* varies systematically across disciplines (social psychology, behavioral psychology, cognitive science, neuroscience, clinical psychiatry, etc.), this variation should be explainable by the multi-level structure of the underlying mechanism. The HPC account can provide exactly this: it can explain the discrepancies in concept application that we find across disciplines when it comes to belief attribution, namely as a reflection of the fact that each discipline investigates part of a multi-level phenomenon. The variation is then a reflection of the fact that the mechanism spans all those levels.

4.2. Faculty Architecture as a Homeostatic Mechanism

Buckner’s (2023) faculty-psychology reconstruction distinguishes at least five domain-general faculties: *perception* continuously updates representations in response to new environmental input, *memory* organizes representations for working use and consolidates new ones for long-term storage, *imagination* enables the anticipatory simulation of possible states and behaviors, *attention* exerts selective control over representational processing, and *social cognition* models the representational states of other agents and calibrates behavior accordingly. Faculties are not levels. Each faculty can be characterized at several levels of description.⁸

My proposal, which goes beyond anything Buckner himself claims, is that the coordinated activity of these faculties constitutes the homeostatic mechanism that sustains the belief cluster in human cognizers. The mechanism is the entire system of causal constraints the faculties impose on one another, and it is realized within each individual cognizer, which is what allows it to meet the ontological criterion. The cluster properties listed in §4.1 can then be assigned to the faculties whose interaction sustains them. Perception and memory, for instance, jointly sustain evidence-sensitivity and revisability (Helton 2020), since revision requires both new input and a stored representation for comparison. Memory and attention jointly sustain inferential integration and cross-context stability, since both require that the representation remembered or attended to remain available and be selected for processing across different contexts.

This is also where the contrast with single-level analyses becomes concrete. A single-

⁸Their organization into a single cooperative system is a matter for one level in particular, the faculty-architectural level (§5.1).

level analysis can record that a system satisfies some cluster properties and not others. It cannot say why *those* properties went together, because the grouping is explained by the organization of the mechanism, and the mechanism is not visible at any one level.

4.3. Folk Psychology as Mechanistic Sketch

The multi-level HPC account also has the ability to offer an explanation for the continued role that folk psychology plays across the sciences of the mind: Craver and Darden (2013) distinguish between mechanistic *schemas*—abstract descriptions of mechanisms with placeholders for known entities and activities—and *sketches*, which are incomplete schemas obtainable at early stages of investigation. One could argue that folk psychological concepts function as initial mechanistic sketches: they identify a target phenomenon and pick out some of its salient features, but they require progressive refinement through investigation at multiple levels in order to become the entities that science investigates.

The folk psychological notion of “belief” would then equally pick out a hypothesized entity whose mechanistic structure is not (yet) fully specified. The folk concept is an entry point into investigation, not a complete description of the mechanism. This is why Porot and Mandelbaum’s assumption that the folk notion is “fairly uncontroversial” and directly continuous with scientific psychology (Porot and Mandelbaum 2021, 3) is methodologically problematic: it treats the sketch as if it were already a schema, bypassing the empirical work needed to fill in the mechanistic detail. And it is why the LLM debate, which forces the question of what the folk concept actually picks out, is a methodologically productive challenge also for cognitive science.

5. Kinding in Progress: The LLM Profile

5.1. Reframing the Question

With the multi-level HPC account in place, the productive question is not “do LLMs have beliefs?” but “what does the profile of LLM systems look like with respect to the multi-level mechanistic structure that sustains the belief cluster in human cognizers?” This approach is what Colaço (2022a) calls *kinding in progress* in the context of memory: categorizing targets when their shared properties are largely unknown and the boundaries of the kind are under investigation. The ascription of beliefs to LLMs is thus not to be understood as a settled claim about kind membership, but as a hypothesis that orients comparative inquiry.

I distinguish five levels at which such a profile can be assessed: the *behavioral* level of what the system does, the *representational* level of the internal states and computations that produce its behavior, the *faculty-architectural* level of how those states and computations are organized into coordinated capacities, the *developmental* level of the history that produced that organization, and the *substrate* level of the physical realizer. The list is not exhaustive and the boundaries are not sharp, but it is fine-grained enough to locate where

the LLM profile departs from the human one.

This reframing implements Buckner’s ecological standard and the lesson in patience from §2.2 simultaneously. Rather than resolving the divergence between the simulation and parity views by stipulation, a multi-level profile can explain why each view reaches its verdict: namely because each is operating at a different level of analysis. The divergence is thus a reflection of the structure of the construct rather than of its incoherence.

The proposal has two close neighbors. Dung (2025) replaces the question whether AI systems are agents with a characterization of their *agency profile* across five dimensions—goal-directedness, autonomy, efficacy, planning, and intentionality—so that a system’s agency becomes a point in a multidimensional space rather than the object of a binary verdict. Schwitzgebel’s (2002) dispositional account treats believing as matching, closely enough, a stereotype of behavioral, phenomenal, and cognitive dispositions, with partial matches counting as cases of in-between believing (Schwitzgebel 2001). Both are profile views, and I share their conviction that a structured profile is more informative than a binary verdict. What neither provides is an explanation of the profile. Dung’s dimensions are independent axes, and Schwitzgebel deliberately locates belief in the dispositional stereotype rather than in whatever underlies it. On the multi-level HPC account, by contrast, the pattern of matches and mismatches is itself explained by the presence or absence of the mechanism that sustains the cluster, and it is this explanatory layer that allows the account to meet Machery’s standard for legitimate variability (§2.2). The difference also has consequences for verdicts. A system that matched enough of the dispositional stereotype would count for Schwitzgebel as at least an in-between believer, whatever produced the match; on the HPC account, the source of the match matters (§5.4).

5.2. Behavioral and Representational Levels

At the *behavioral level*, the case for belief-like states in LLMs appears most convincing. LLM outputs largely exhibit the functional signatures of belief-expression: they seem to make systematic assertions, draw contextually sensitive inferences, calibrate uncertainty, and revise their outputs in response to argument. Given that evidence of this nature is commonly cited in support of belief attribution to other humans and to non-human animals, consistency alone appears to require that the question of whether LLMs have beliefs be taken seriously. Herrmann and Levinstein’s (2025) framework for evaluating LLM representations along four independent criteria—accuracy, coherence, uniformity, and use—suggests that the behavioral level itself is multidimensional: a system can score high on accuracy and low on coherence, or exhibit appropriate use without uniformity. Evaluating it even at this level may already require a patchwork structure.

Recent work in mechanistic interpretability has substantially complicated the earlier picture at the representational level. Marks and Tegmark (2024) demonstrate that LLMs maintain linearly decodable internal representations of propositional truth values that generalize across contexts and causally influence downstream inferences. Beckmann and

Queloz (2026) expand on this finding with a richer three-tier framework. At the first tier, *conceptual understanding*, LLMs possess features—directions in their internal representational space—that function analogously to concepts: the same feature activates reliably across diverse descriptions of an entity, is causally efficacious (steering experiments have shown that amplifying a feature changes the system’s behavior accordingly), and meets basic normativity conditions for representation. At the second tier, *state-of-world understanding*, LLMs can develop internal models of changing external states, as shown by the Othello-GPT findings: the model maintains a linearly decodable, causally efficacious representation of the board that it updates dynamically with each move. At the third tier, *principled understanding*, models can transition to discovering and implementing algorithms. In the “grokking” phenomenon, a model trained on modular arithmetic discovered the underlying Fourier multiplication algorithm rather than memorizing a lookup table (Beckmann and Queloz 2026, 1806).

A number of properties commonly cited as constitutive of belief appear hopelessly difficult to spell out in purely computational terms: how, for instance, could the property of serving to guide action in the world be implemented computationally? More important for present purposes, however, is a structural feature that Beckmann and Queloz (2026) call the “motley mix”: LLMs deploy mechanisms of radically varying sophistication in parallel, with no centralized or hierarchical control. Behavioral output is therefore an unreliable guide to which mechanisms are actually operative.

5.3. Faculty-Architectural and Developmental Levels

At the *faculty-architectural level*, the divergence of LLMs from biological cognizers becomes pronounced, though it needs to be stated with care. LLMs are not without faculty analogues: Transformer attention implements a form of selective weighting, thereby constraining the possible representational input; the context window functions as a limited working memory; and generative sampling bears some resemblance to processes of the faculty of imagination. Buckner (2023) develops several analogies of this kind. However, what LLMs lack, in their base architecture, is twofold. First, some faculties have no analogue at all. LLMs do not possess any memory capacity that consolidates new experiences and persists across interactions. They equally lack social cognition and hence cannot track interlocutors as distinct persisting agents. Second, and more importantly, these fragmentary analogues are disjointed, rather than integrated into a cooperative architecture in which each faculty constrains the others. Likewise, Buckner argues that full integration of faculty modules remains aspirational even in the most advanced current systems, and that current systems in particular still lack a model of the emotional and affective faculties involved in bootstrapping higher-order faculties. LLMs do not merely lack some cluster properties; they lack the entire architectural organization that, on my proposal, constitutes the homeostatic mechanism.

At the *developmental and substrate levels*, the divergence is even more pronounced.

Human belief systems are shaped by evolutionary trajectories, embodied cognitive development, social interaction, and so on, and the characteristic organization of the human faculty architecture is the result of this developmental history. LLMs are trained on population-level linguistic data through computational processes that have no analog to this kind of development. LLM training produces something like a population-level latent structure over human-generated linguistic data, but not a belief system that is the belief system *of* any particular cognizer with a continuous developmental and experiential history.⁹

5.4. What the Profile Reveals

The profile yields a structured and hence more precise picture: LLMs instantiate typical belief-cluster properties at the behavioral and representational levels; they diverge at the faculty-architectural, developmental, and substrate levels. The divergence between levels is systematic and explainable.¹⁰ The behavioral and representational properties that LLMs exhibit are exactly those that optimization over a sufficiently large quantity of human belief-expressions would produce, while the architectural and substrate properties they lack are exactly those that require embodiment, a developmental history, or socially embedded processes. A multi-level profile has the resources to provide exactly this kind of explanatory structure, rather than assuming the unprincipled construct heterogeneity that Machery warns against.

One feature of the HPC account is particularly consequential in the context of the LLM mentality debate: On the HPC account, kind membership is fixed by the mechanism sustaining the cluster, not by cluster properties alone. If the homeostatic mechanism for belief is the integrated faculty architecture, and LLMs lack that specific architecture, while the match at the behavioral and representational levels is the result of the operations of a different mechanism (optimization over human linguistic data), then the account delivers a negative verdict: The outputs of current LLMs are not members of the kind human belief.

The findings reviewed in §5.2 still leave room for a further possibility to be explored: LLMs may still share a representational organization with human believers if we can show that both are sustained by a shared mechanism that is more abstract than faculty architecture. If so, there could be a broader kind of belief of which human belief is one realization, and LLM belief another. Interpretability work is currently exploring this possibility. If it manages to establish that representational organization is shared, the question becomes whether that shared organization, rather than the full faculty architecture, is what sustains the cluster. As Colaço (2022b) argues with respect to plant cognition, investigating a border case of this kind is valuable however the membership question resolves.¹¹

⁹At the substrate level, for instance, the brain’s default network supports autobiographical memory, the simulation of possible futures, and reasoning about others’ mental states (Andrews-Hanna 2012)—the very faculties whose coordination I have proposed sustains the belief cluster. Nothing in LLM processing plays a comparable integrative role.

¹⁰Which is precisely what Machery’s standard for legitimate construct variability requires.

¹¹The LLM profile can function as an experiment in this respect, revealing which properties can be produced

Sullivan’s (2022) link uncertainty accounts for the distribution of confidence across the profile. The behavioral level establishes how-possibly cases for belief attribution; mechanistic interpretability is progressively reducing link uncertainty at the representational level; the architectural and developmental levels remain, for now, sites of how-possibly explanation. That the faculty architecture is what sustains the belief cluster is itself a hypothesis awaiting how-actually support, which is why the verdict above is stated conditionally.

6. Methodological Lessons

6.1. For the AI Mentality Debate

The central practical upshot of the multi-level profile is level-specific guidance, which neither single-criterion view can provide. The question is not simply whether LLMs believe, but at which levels LLM states satisfy which parts of the belief cluster, and what follows epistemically, ontologically, and practically from the resulting pattern.

Buckner’s two biases can now be diagnosed as failures to keep the levels apart: Semantic anthropocentrism treats divergence at the architectural or developmental level as settling the question at every level, thereby ruling out the possibility of attributing any belief-cluster property to a non-human system. Anthropofabulation, by contrast, treats convergence at the behavioral level as settling the question at every level, and so credits the system with the full human profile.

LLM hallucinations further illustrate the practical payoff. On the simulation view, a hallucination does not amount to a malfunction at all: since the system produces bullshit with no regard for the truth and falsity of its outputs, a hallucination is merely the case in which the bullshit happens to be false (Hicks, Humphries, and Slater 2024). On the functional parity view, by contrast, LLM hallucinations are most naturally treated as a false belief, to be corrected as one would correct a human mistake. A multi-level profile can distinguish cases that both views run together. A hallucination could for instance manifest as lack of evidence-sensitivity at the behavioral level, while at the representational level, two quite different failures can produce it. A case of a failure of use in Herrmann and Levinstein’s (2025) terminology would involve an internal representation of the relevant truth value, which however fails to govern the output.¹² Alternatively, no accurate representation may be present at all, resulting in a failure of accuracy.

Whether current investigative methods can reliably tell failures of use from failures of accuracy apart is itself under dispute (Levinstein and Herrmann 2024), which is one more reason to treat the representational level as a site of ongoing link uncertainty. Failures of use and failures of accuracy require different interventions in practice. In the

by optimization over linguistic data alone and which require additional mechanisms, such as biological ones.

¹²Marks and Tegmark (2024) similarly claim that if truth values are linearly represented and causally efficacious, an output that contradicts the represented value should be interpreted as a failure of use rather than of representation.

first case, methods that bring outputs into line with internal representations are needed, whereas in the second data or training need to be improved. At the faculty-architectural level, hallucination reflects the missing coupling between memory and perception that allows human cognizers to check incoming evidence against stored representations (§4.2). A level-specific account furthermore has the ability to predict how certain interventions will affect hallucinations. Retrieval and tool use can reduce hallucinations, but in order to do so their contributions need to be deployable at the appropriate level into the overall system that produces assertions.

Level-specificity can also give guidance for the practice of ascribing intentional states: If the analysis draws only on the behavioral and representational-level properties, predicting what a system will output, or diagnosing why it failed, belief talk can earn its keep in the way Goldstein and Lederman (2025) suggest. But where the analysis depends on properties of the faculty architecture, it does not; a level-specific account can explain why.

6.2. For the Philosophy of Psychology

The LLM case offers a more general lesson for the analysis of constructs in psychology. The multi-level HPC account provides the theoretical framework within which methodological patience can be realized: rather than immediately resolving divergence by stipulating construct heterogeneity or by privileging one measurement approach over others, investigators should map the multi-level profile of the construct and ask whether the divergence can be explained in terms of the structure of the mechanism.

Constructs that exhibit the Godman-Bellander pattern, satisfying the epistemic criterion while failing the ontological criterion, should not be abandoned but investigated at additional levels. The LLM case illustrates this: the behavioral and representational evidence is genuine, but the full construct requires investigating the architectural, developmental, and substrate levels as well. The same applies to the constructs of quantitative psychology: the fact that the *g* factor is epistemically productive does not settle the question of whether it tracks a genuine individual-level entity, and the appropriate response is a multi-level investigation.

6.3. For Philosophy of Science

A final lesson concerns the scope of the problem. Outside psychology, Godman and Bellander's ontological criterion generalizes to the requirement that a construct be grounded in the systems to which it is attributed, rather than only in the population-level or single-level patterns from which it is inferred. So generalized, some version of the two-criteria tension arises whenever a construct is assessed at a single level while the phenomenon it names is sustained by mechanisms operating across several. Temperature in Wilson's analysis shows what a resolution looks like: it is a patchwork concept whose extension varies systematically across scales, and the variation is explained by the underlying physics. The analogous hypothesis for belief is that its extension varies across the behavioral, represen-

tational, architectural, and developmental levels, and that this variation is explained by the multi-level structure of the mechanism that sustains it.

Link uncertainty also yields a positive policy for evidence-gathering. The right response to high link uncertainty at a given level is neither to dismiss the construct nor to declare the evidence useless, but to identify what kind of evidence would reduce the uncertainty at that level (Sullivan 2022). Mechanistic interpretability is a model case at the representational level: work of the kind Beckmann and Queloiz (2026) survey aims to establish which internal structures are causally efficacious rather than merely correlated with outputs, and so converts how-possibly into how-actually links. The corresponding work at the architectural and developmental levels has barely begun.

Combined with Craver and Darden's (2013) distinction between sketches and schemas, this gives a general picture of how constructs mature. Folk concepts are initial sketches that identify targets for investigation, investigation fills in the mechanistic detail level by level, and the resulting schemas are progressively better approximations to the underlying kind. Temperature traveled this path, from the felt contrast between hot and cold to a multi-level framework that can assign different temperatures to the same object at different scales. The same trajectory is available for belief.

References

- Andrews-Hanna, Jessica R. 2012. "The Brain's Default Network and Its Adaptive Role in Internal Mentation." *The Neuroscientist* 18 (3): 251–70. <https://doi.org/10.1177/1073858411403316>
- Beckmann, Pierre, and Matthieu Queloiz. 2026. "Mechanistic Indicators of Understanding in Large Language Models." *Philosophical Studies* 183: 1803–1830. <https://doi.org/10.1007/s11098-026-02513-1>
- Bortolotti, Lisa. 2010. *Delusions and Other Irrational Beliefs*. Oxford: Oxford University Press.
- Boyd, Richard N. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61 (1–2): 127–48. <https://doi.org/10.1007/BF00385837>
- Boyd, Richard N. 1999a. "Kinds, Complexity and Multiple Realization: Comments on Millikan's 'Historical Kinds and the Special Sciences'." *Philosophical Studies* 95 (1–2): 67–98. <https://doi.org/10.1023/A:1004511407133>
- Boyd, Richard N. 1999b. "Homeostasis, Species, and Higher Taxa." In *Species: New Interdisciplinary Essays*, edited by Robert A. Wilson, 141–85. Cambridge, MA: MIT Press.
- Brigandt, Ingo. 2022. "How to Philosophically Tackle Kinds without Talking about 'Natural Kinds'." *Canadian Journal of Philosophy* 52 (3): 356–79.
- Buckner, Cameron. 2016. "Transitional Gradation in the Mind: Rethinking Psychological Kindhood." *British Journal for the Philosophy of Science* 67 (4): 1091–1115. <https://doi.org/10.1093/bjps/axv013>
- Buckner, Cameron. 2023. *From Deep Learning to Rational Machines: What the History of*

- Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford: Oxford University Press.
- Cappelen, Herman, and Josh Dever. 2025. "Introspective Machines: Are LLMs Better at Self-Reflection Than Humans?" *Philosophical Perspectives* 38 (1): 189–196. <https://doi.org/10.1111/phpe.12201>
- Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.
- Colaço, David. 2022a. "What Counts as a Memory? Definitions, Hypotheses, and 'Kinding in Progress'." *Philosophy of Science* 89 (1): 89–106. <https://doi.org/10.1017/psa.2021.14>
- Colaço, David. 2022b. "Why Studying Plant Cognition Is Valuable, Even If Plants Aren't Cognitive." *Synthese* 200: article 46. <https://doi.org/10.1007/s11229-022-03479-3>
- Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago: University of Chicago Press.
- Dung, Leonard. 2025. "Understanding Artificial Agency." *The Philosophical Quarterly* 75 (2): 450–72. <https://doi.org/10.1093/pq/pqae010>
- Ereshefsky, Marc, and Thomas A. C. Reydon. 2023. "The Grounded Functionality Account of Natural Kinds." In *From Biological Practice to Scientific Metaphysics*, edited by William C. Bausman, Janella K. Baxter, and Oliver M. Lean. Minnesota Studies in the Philosophy of Science 23. Minneapolis: University of Minnesota Press.
- Frankfurt, Harry G. 2005. *On Bullshit*. Princeton: Princeton University Press.
- Godman, Marion, and Martin Bellander. 2024. "Latent Variables, Psychological Constructs, and the Prospect of Scientific Kinds in Psychology." *British Journal for the Philosophy of Science*. <https://doi.org/10.1086/733192>
- Goldstein, Simon, and Harvey Lederman. 2025. "What Does ChatGPT Want? An Interpretationist Guide." Manuscript.
- Goldstein, Simon, and Benjamin A. Levinstein. 2024. "Does ChatGPT Have a Mind?" *arXiv preprint arXiv:2407.11015*. <https://doi.org/10.48550/arXiv.2407.11015>
- Griffiths, Paul E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.
- Helton, Grace. 2020. "If You Can't Change What You Believe, You Don't Believe It." *Noûs* 54 (3): 501–26.
- Herrmann, Daniel A., and Benjamin A. Levinstein. 2025. "Standards for Belief Representations in LLMs." *Minds and Machines* 35 (1): article 5. <https://doi.org/10.1007/s11023-024-09709-6>
- Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26: article 38. <https://doi.org/10.1007/s10676-024-09775-5>
- Khalidi, Muhammad Ali. 2023. *Natural Kinds*. Cambridge: Cambridge University Press.
- Levinstein, Benjamin A., and Daniel A. Herrmann. 2024. "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks." *Philosophical Stud-*

- ies. <https://doi.org/10.1007/s11098-023-02094-3>
- Machery, Edouard. 2022. “Anomalies in Implicit Attitudes Research.” *WIREs Cognitive Science* 13 (1): e1569. <https://doi.org/10.1002/wcs.1569>
- Machery, Edouard. 2023. “Constructs in Psychology: Lessons from the Philosophy of Science.” *Psychological Inquiry* 34 (4): 273–76. <https://doi.org/10.1080/1047840X.2023.2274367>
- Marks, Samuel, and Max Tegmark. 2024. “The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets.” In *Proceedings of the First Conference on Language Modeling* (COLM 2024).
- Porot, Nicolas, and Eric Mandelbaum. 2021. “The Science of Belief: A Progress Report.” *WIREs Cognitive Science* 12 (2): e1539. <https://doi.org/10.1002/wcs.1539>
- Quilty-Dunn, Jake, Nicolas Porot, and Eric Mandelbaum. 2023. “The Best Game in Town: The Re-Emergence of the Language-of-Thought Hypothesis across the Cognitive Sciences.” *Behavioral and Brain Sciences* 46: e261. <https://doi.org/10.1017/S0140525X23002849>
- Quilty-Dunn, Jake, and Eric Mandelbaum. 2018. “Against Dispositionalism: Belief in Cognitive Science.” *Philosophical Studies* 175 (9): 2353–72. <https://doi.org/10.1007/s11098-017-0962-x>
- Schwitzgebel, Eric. 2001. “In-Between Believing.” *The Philosophical Quarterly* 51 (202): 76–82.
- Schwitzgebel, Eric. 2002. “A Phenomenal, Dispositional Account of Belief.” *Noûs* 36 (2): 249–75.
- Shevlin, Henry. 2026. “Three Frameworks for AI Mentality.” *Frontiers in Psychology* 17: 1715835. <https://doi.org/10.3389/fpsyg.2026.1715835>
- Sullivan, Emily. 2022. “Understanding from Machine Learning Models.” *British Journal for the Philosophy of Science* 73 (1): 109–33. <https://doi.org/10.1093/bjps/axz035>
- Velleman, J. David. 2000. “On the Aim of Belief.” In *The Possibility of Practical Reason*, 244–81. Oxford: Oxford University Press.
- Wilson, Mark. 2006. *Wandering Significance: An Essay on Conceptual Behaviour*. Oxford: Oxford University Press.
- Wilson, Mark. 2017. *Physics Avoidance: Essays in Conceptual Strategy*. Oxford: Oxford University Press.