Superficialism about Belief,
and How We Will Decide that Robots "Believe"

Eric Schwitzgebel
Department of Philosophy
University of California, Riverside
Riverside, CA 92521
USA

September 4, 2025

[intended for a special issue of *Semiotic Studies* on Krzysztof Poslajko's *Unreal Beliefs*]

Superficialism about Belief,
and How We Will Decide that Robots "Believe"

Abstract: *Superficialism* about property X treats the possession of property X as determined entirely by superficial as opposed to deep facts. Belief should be understood superficially. Facts about belief are determined entirely by facts about actual and potential behavior and conscious experience – plus transitional cognitive states ultimately understood in terms of actual and potential behavior and conscious experience. Superficialism about belief excels on both intuitive and pragmatic grounds, compared to accounts of belief in terms of deep cognitive or neural architecture. Behavior-focused superficialism suggests that robots and Large Language Models may soon (perhaps already do) believe. If consciousness is also essential to belief, matters might soon become more complicated, if it becomes reasonable to wonder whether some of our most advanced AI systems are conscious. Regardless, it will be practical to describe some such systems has having what I'll call "belief*" – belief shorn of commitment to any conscious aspect – and people might forgivably forget to pronounce the asterisk. Krzysztof Poslajko should welcome this manner of thinking, though it needn't be as "antirealist" as he suggests.

Word Count: ~5000 words

Keywords: artificial intelligence; belief; Large Language Models; Poslajko, Krzysztof; robots; superficialism

Superficialism about Belief,
and How We Will Decide that Robots Believe

1. Introduction.

I will explain and defend *superficialism* about belief, discuss its application to the question of whether robots and Large Language Models believe, and suggest that Krzysztof Poslajko should agree with these positions, while weakening his commitment to "anti-realism" about belief.

2. Superficialism Explained.

A superficial judge of character will assess someone's clothes, bearing, and words interpreted at face value. A superficial reading of Shakespeare will likewise mostly take characters at their word and focus on the obvious aspects of each scene, rather than searching for deeper themes and motives. A superficial historical account will take actions one at a time without exploring nonobvious connections or underlying social forces. A superficial housecleaning doesn't touch the backsides and undersides of household items. In general, superficial treatments and understandings focus on what is readily observed and disregard or downplay what it requires work to discern. However, what counts as "the surface" can vary for different types of superficiality.

Superficialism about property X is the view that whether an entity has property X is determined (that is, constituted or grounded; covariance or supervenience is not enough) entirely by superficial facts about that entity, where "superficial facts" are some class of readily observed facts, in contrast with some alternative set of "deeper", or less obvious, less readily observable facts. So, for example, a superficialist about beauty in humans might hold that beauty is skin-

deep. On such a view, whether a person is physically beautiful depends entirely on the shape and coloration of their bodily surface; nothing else matters. A metaphysical superficialist about cats might hold that even robotic cats are cats, as long as those robots superficially look and act enough like cats. A different sort of superficialist might define the "surface" to include internal organs, so that a macroscopic copy of a cat is a cat, regardless of its evolutionary history or whether it has catlike DNA. A dissection goes beneath the skin, both literally and metaphorically, but not as deep as a DNA analysis.

Superficialism about belief is the view that whether an entity believes some particular proposition depends entirely on some set of "superficial" facts about that entity. For example, philosophical behaviorism (Graham 2000/2023; Robinson 2019) takes the relevant superficial facts to be the entity's actual and potential outward behavior, including verbal behavior. Dennettian and Davidsonian interpretativism (Dennett 1987; Davidson 1984; cf. Mölder 2010; Curry 2020; Pautz 2021a,b) is similarly superficialist: Whether an entity is properly interpretable as believing depends entirely on facts about its behavior and environment, as they guide potential interpretative practices, and not at all (except derivatively) on brain states or underlying cognitive architecture. On such views, an entity made of pure undifferentiated balsa wood, if it would always outwardly act and react identically to a person who believes that plaid dungarees are hideous, necessarily would also believe that plaid dungarees are hideous. (Such an entity is presumably not physically possible, but the relevant modality is conceptual possibility.)

The form of belief superficialism I favor – "phenomenal" or "liberal" dispositionalism (Schwitzgebel 2002, 2013, 2021, forthcoming; Bantegnie and Schwitzgebel forthcoming) – conceptualizes the surface more capaciously, in terms not only of behavioral but also of phenomenal and cognitive dispositions. *Behavioral* dispositions are just the dispositions

countenanced by philosophical behaviorism. *Phenomenal* dispositions are dispositions to have particular conscious experiences (e.g., feelings of surprise, inner speech experiences). *Cognitive* dispositions are dispositions to enter cognitive states that are in turn definable in terms of other behavioral, phenomenal, and cognitive states and dispositions (e.g., the disposition to adopt a new belief that logically follows from the belief in question, or the disposition to form a new intention, on the assumption that intentions themselves can be understood in terms of behavioral, phenomenal, and cognitive dispositions). Ideally, on this view, all of the terms for cognitive dispositions could in principle be eliminated en masse by Ramsification (Lewis 1972): They serve as ontological placeholders in a network characterizable entirely in terms of phenomenal and behavioral dispositions, plus other such transitional "cognitive" placeholders. Some dispositions might mix the behavioral, phenomenal, and cognitive: For example, the disposition to *express the thought* that plaid dungarees are hideous might have behavioral, phenomenal, and cognitive aspects. Crucially, neural states and low-level functional architecture are not part of the surface.

Thus, on this view, to believe that plaid dungarees are hideous is just to be disposed to say they are hideous if asked, to refuse to wear them, to feel revolted upon seeing them, to reduce one's opinion of a friend's sartorial taste upon seeing that friend in plaid dungarees, and so on. All such dispositions hold only *ceteris paribus* (all else being equal or normal, absent countervailing influences) or given implicit conditions. One might not express one's negative opinion to a sensitive friend proud of their new purchase. The relevant dispositions are not a finitely specifiable set, and ordinary folk psychology enables us to devise or recognize new ones on the fly: If the believer in question wishes to revolt a suitor, wearing plaid dungarees might strike them as just the right choice. The disposition to make that choice in those circumstances is

not only *consistent* with believing them to be hideous (springing the *ceteris paribus* clause, so to speak) but partly *constitutive* of having the belief (which includes the disposition positively to wear them under certain rare conditions).

Anyone possessing the pattern of dispositions I began to express in the previous paragraph, and which ordinary language users could further articulate with good-enough agreement, if that person has that pattern of dispositions robustly, across a wide variety of situations, thereby believes that plaid dungarees are hideous. There's nothing more to it. It doesn't matter what low-level cognitive architecture they have, as long as it robustly supports this dispositional pattern. It doesn't matter what brain states they are in, or even whether they have a brain at all, as long as their physiology supports this pattern of behavioral, phenomenal, and cognitive dispositions. It doesn't matter what their causal history is, or whether they even have a causal history (unless having a causal history of the right sort is necessary for having the dispositions, for example, on some versions of externalism about phenomenology and linguistic meaning; e.g., Dretske 1995; Putnam 1975).

Beliefs, on this view, are structurally like personality traits as we ordinarily think of personality traits. To be an extravert, or to be narcissistic, is nothing more or less than to be disposed to act and react – behaviorally, phenomenally, and cognitively – in the manner characteristic of an extravert or narcissist. Extraverts and narcissists have brains and deep cognitive architectures that are responsible for those dispositional patterns, but it's the dispositional patterns that constitute the trait, not the brain states or deep architectures, and any entity with the exactly the same dispositional patterns would be extraverted or narcissistic to exactly the same degree, even if their brain or architecture were somehow radically different. I hope this strikes you as an intuitively plausible metaphysics of personality traits. According to

my variety of liberal dispositional superficialism about belief, the same metaphysics applies to believing that plaid dungarees are hideous, and believing that your children's happiness is more important than their grades, and believing that Confucius loved the ancient rituals of the early Zhou dynasty.

3. Liberal Superficialism Defended.

Liberal superficialism about belief is attractive on intuitive and pragmatic grounds, and not unattractive on scientific grounds.

The Intuitive Defense. Intuitively, superficialism gives the right answers about cases (Poslajko calls these "application criteria"). Imagine someone as different as possible from you in their internal architecture and representational structures, consistently with robustly possessing all the phenomenal, behavioral, and cognitive dispositions associated with believing plaid dungarees to be hideous. A science fiction creature, snooping on our internet, sees an advertisement for plaid dungarees. Recoiling in horror, it exclaims "U yˇ ʔ!" – where "U" is a noun regularly instantiated in the presence of plaid dungarees, "ʔ" is an adjective regularly instantiated in the presence of things from which it aesthetically recoils, and "y" is a copula. This creature would not purchase a pair, except perhaps as a joke. It would downgrade its opinions of any human or alien who intentionally chose to wear such an item. And so on. Not only externally, but also phenomenally and cognitively, it is disposed to act and react just like a human who finds plaid dungarees hideous, despite as-different-as-possible a low-level architecture. I suggest that, intuitively, most of us would find it natural to describe this alien as believing that plaid dungarees are hideous.

Conversely, imagine an entity as similar as possible to us in low-level internal architecture but who lacks the behavioral, cognitive, and phenomenal dispositions characteristic of (I would say constitutive of) believing that plaid dungarees are hideous. They have no tendency to negatively judge those who wear them, no inclination to avoid purchasing them, no inclination to agree with a friend who calls them hideous, and so on – and further, they have these dispositions robustly. This isn't a case in which the dispositions are "masked" or some unusual circumstance obtains (such as preparing a method-acting performance of a character who favors plaid dungarees). To the extent we can imagine such cases, it seems right to say that despite those details of architecture, the entity in question does not find plaid dungarees hideous.

What if the behavioral dispositions are present but the phenomenal dispositions absent or wrong? If there's no phenomenology whatsoever, we might think of it as a "zombie" case, in the sense of Kirk (1974) and Chalmers (1996). Intuitions probably diverge about whether such zombies genuinely have beliefs (see Fischer and Sytsma 2021). However, if the phenomenal dispositions are reversed (e.g., inner delight at seeing dungarees) or irrelevant (e.g., experiencing neither delight nor loathing but instead only the sound of a trumpet), the intuitive inclination to ascribe belief is likely to be seriously undercut. What if the phenomenal dispositions are present and behavioral dispositions absent? If the absence of behavior is "excused" – for example, by paralysis – that's probably enough to make belief ascription intuitive (as in Strawson's 1994 "Weather Watchers"). If the absence of behavior is inexplicable – experiencing passionate confidence that P is true but excuselessly having no inclination to assert it or act on it – intuitions probably collapse or diverge.

My suggestion is this: Our intuitive judgments in belief ascription track the patterns of behavioral, phenomenal, and cognitive dispositions. When they are robustly present, we're

inclined to ascribe belief. When they are robustly absent, we're inclined to deny belief. When they're mixed, intuitions diverge or we treat it as an in-between case. Thus, the type of superficialism I am calling liberal dispositionalism gets belief ascription extensionally correct by intuitive standards. If we treat possession of the relevant behavioral, phenomenal, and cognitive dispositions as necessary and sufficient for belief, or even better constitutive of belief, we can respect and nicely account for that pattern of intuition.

The Pragmatic Defense. Philosophical theory needn't always follow ordinary intuition about cases. When it doesn't, good pragmatic or scientific/theoretical grounds should justify that break from ordinary use. Are there pragmatic grounds for favoring a deep view or a purely behavioral superficialism?

To see why not, consider: Why do we care about what people believe? It's because we care about how they do and would act, what they do and would feel, and what inferences they do and would make – not because we care about their deep architecture (except insofar as that architecture supports what we do care about). You care what your friend thinks of plaid dungarees because you want to know what birthday purchases would and wouldn't delight her, her fashion preferences, what quiet judgments she might be making about your new plaid-dungaree-wearing boyfriend. You care about what she would outwardly say and do, and also – contra entirely behavioral approaches – about what she might inwardly think and feel. The superficialist account of belief in terms of behavioral, phenomenal, and cognitive dispositions tracks what we care about in belief attribution. It tracks why the word "belief" is important to us in the first place.

Suppose we were to devise a new term, *schmelief,* for people (perhaps a subclass of humans with highly unusual interior architectures, or perhaps alien creatures who have joined

human society) whose behavioral, phenomenal, and cognitive dispositional patterns match ours but whose deep architecture is different. Your friend Trina believes that plaid dungarees are hideous, while your other friend Jordan schmelieves that they are hideous. Wouldn't you want some term to capture what Trina and Jordan share in common? You'd have to invent a third term – call it *lululieve* – to capture the genus to which belief and schmielief belong. *Lululieve* will turn out to be the more useful and important term for most ordinary purposes. It will also be the one you apply to new acquaintances, if you don't yet know their interior architecture. It's the more important and practical term. It is the one we will shift to using. If I am right, it is what we meant, or should have meant, by "believe" all along.

Furthermore, I'd suggest, it's not just the outward behavior that matters pragmatically but the consciousness experiences as well. Experiences of affirmation, surprise, conscious judgment, confidence, and doubt; inner speech and imagination; nervousness about *what might happen if*; facts about what does and does not occupy conscious attention; and so on – these are too central to our interests in belief ascription to warrant excluding them from the stereotype constitutive of believing. They comprise a crucial part of what we care about in ascribing beliefs. We don't want only to predict and explain outward behavior; we want our belief ascriptions also to track the target's experienced inner life. This point is perhaps more vivid if we imagine belief-like outward behavioral dispositions alongside *contrary* phenomenology – saying that P while innerly thinking that not-P, etc. – though a pure case of this might be even harder to coherently imagine than an experientially blank philosophical zombie. Phenomenology as central not only to what we do track but to what we want to, and should want to, track in belief ascription. If so, we should agree that an entity with no conscious experiences whatsoever could at best qualify only as a borderline, mixed, or in-between believer – one who conforms to some important portion of

the dispositional stereotype while sharply deviating from another important portion. However, my main aim in this paper is to defend superficialism in general not liberal dispositionalism in particular. We care a lot about behavior – often quite a bit more about that than about inner life – and I have no strong objection against a pragmatic defense of behaviorist superficialism. Indeed, in Section 4, I will defend a close relative of that view.

The Scientific Defense. Some philosophers have suggested that superficialism about belief is unscientific (for example, Quilty-Dunn and Mandelbaum 2018). A properly scientific approach requires looking at the subpersonal architecture underlying belief and characterizing belief in those terms. While I certainly don't object to exploring underlying cognitive architecture, the result is a cognitive science of *humans*, or *mammals*, or some other specific biological type. If we aim to think about belief in general, as it might potentially apply to future conscious AI systems (if they can exist) or hypothetical (or actual, presumably somewhere in our giant universe) alien systems, our account of belief had better not rest too finely on contingent details of human architecture. The question of AI belief is imminent, and even a skeptic about AI belief should agree that it cuts inquiry precipitously short to rule out AI belief simply on grounds of different underlying architecture, whatever else might be true of them.

Compare again with personality traits. It would be interesting to discover that extraversion in humans is implemented by a particular neural or low-level functional architecture. But what does and should constitute extraversion is not such details of implementation. (Admittedly, there's a metaphysical rabbit-hole here related to the debate between role functionalism and realizer functionalism; see Levin 2004/2023; Pober 2024.) Personality psychology is not unscientific by virtue of staying at the dispositional surface. Neither is the treatment of belief. Indeed, as I have argued elsewhere (Schwitzgebel

forthcoming), the most obvious way of being a deep-structure scientific realist about belief generates pseudopuzzles, such as attempting to count the number of beliefs that are stored and accessed.

4. How We Will Decide That Robots, and Maybe Language Models, "Believe".

If behaviorism were the best approach to belief, then robots and language models would believe if (and only if) their behavioral dispositions are the dispositions characteristic of believers. Behaviorist accounts might differ concerning the relevant class of behavioral dispositions. If belief requires dispositions to bodily action in our shared world, then robots might believe, but a purely text-based language model could never believe. If linguistic dispositions are sufficient, then language models could potentially also believe (perhaps if they can pass a sufficiently rigorous Turing test; Turing 1950). If behavioristically inflected interpretativism were correct, the same result would follow. However, to answer the question of whether robots and language models believe, as I prefer to construe the question, we must determine whether robots and language models have (the right kind of) conscious experiences.

Scientific consensus and popular opinion strongly favors the answer that artificial systems are *not yet* meaningfully conscious. However, that consensus might soon break, with popular opinion quickly shifting and some influential mainstream scientists and philosophers holding that we are on the cusp of designing meaningfully conscious or sentient AI systems (Dehaene, Lau, and Kouider 2017; Chalmers 2023; Butlin et al. 2023; Colombatto and Fleming 2024). Once that break occurs, consensus is likely to elude us for a long time, with some researchers defending restrictive and/or biological views on which no familiar type of AI system could possibly have consciousness (Godfrey-Smith 2016; Seth 2024). Our scientific and

theoretical tools will fail, their applicability contingent on contentious assumptions about consciousness that advocates of alternative views will justifiably decline to accept (Schwitzgebel 2024).

Still, we will want to talk in certain ways about robots and language models. In particular, we will want to communicate with each other about how they are likely to respond to certain types of inputs or situations. The language of belief, and more common neighboring folk-psychological terms such as "thinks" and "knows", will be convenient and tempting. Does this language model think/believe/know that David Lewis had diabetes? Does this nursing assistant robot think/believe/know/remember that the pills are in the second drawer on the left? With cautious neutrality about whether such AI systems really have the phenomenal and phenomenology-involving cognitive dispositions necessary for genuine belief, we might create the new dispositional concept *belief**, which is present if and only if the system has the right behavioral dispositions, plus cognitive dispositions characterizable in terms of networks of behavioral dispositions.

A language model might then have a belief* that P (for example belief* that Paris is the capital of France or belief* that cobalt is two elements right of manganese on the periodic table) if:

- it would consistently output P or text strings of similar content consistent with P, when directly asked about P;
- it would frequently output P or text strings of similar content consistent with P, when P is relevant to other textual outputs it is producing (for example, when P would support an inference to Q and it has been asked about Q);

- it would rarely output negations of P, or text that (if a human were to write it) would be naturally interpreted as claims of ignorance about P, and similarly for negations of propositions that straightforwardly follow from P given its other beliefs*;
- when P, in combination with other propositions that the language model believes*, would straightforwardly imply Q, and the truth of Q is important to the truth or falsity of forthcoming text outputs, it will commonly output Q or enter a state that facilitates the future output of Q;
- the above dispositions are relatively stable, unless new evidence against the truth of P is presented.

These conditions could be refined and further conditions could be added, but this conveys the idea. These conditions are imprecise, but that's a feature, not a bug. The same is true for dispositional characterizations of personality traits and human beliefs. Such concepts are fuzzy-bordered and require expertise to apply.

Current language models do not meet these conditions as a general matter. They "hallucinate" too frequently, change their answers, don't consistently enough "remember" what they earlier committed to, can easily be coaxed into making a statement and then later retracting it, make tangled and inconsistent inferences. If I ask ChatGPT 4.5 whether waffles are better than eggs or vice versa, it will sometimes say (or say*) one thing, sometimes the opposite, sometimes refuse to express a preference unless pressed, and reverse itself. If I ask Claude 3.7 Sonnet which element is two to the right of manganese on the periodic table, it sometimes gets it

right, sometimes gets it wrong, and sometimes reasons in baldly inconsistent ways.[1] On matters such as these, current Large Language Models emit text *suggestive* of belief*, but it's probably not right to say that they have beliefs* about the superiority of eggs to waffles or the relative positions of elements in the periodic table.

Still, the situation is changing fast, and for widely shared pieces of common knowledge, such as that Paris is the capital of France, Large Language Models show excellent stability and use of inferential connections. Also, of course, real human believers are far from consistent and inferentially perfect, so we shouldn't insist on too high a bar. It's a reasonable extension of common usage to say that ChatGPT "says" and "knows" that Paris is the capital of France and thus, perhaps, similarly for "believes". Even if we want to say that such extensions aren't literally correct usages, this is probably enough for belief*. The practical problem is in determining where these models' beliefs* end and their wobbly responding begins. No rigid or bright-line criterion is likely to work. Instead, I recommend a pragmatic approach to belief* – the ascription need only be good enough, in the relevant respects, for the purposes at hand.

For embodied robots with sensorimotor activity in our shared spaces, belief-talk will likely prove practically unavoidable. The nurse's assistant robot rolls into a client's room. Does

---

[1] For an example of an incoherent answer, as of this writing on March 11, 2025, I asked Claude 3.7 Sonnet:

> What element is two elements to the right of manganese on the periodic table?

It replied:

> The element that is two elements to the right of manganese in the periodic table is iron (Fe).
>
> Manganese has atomic number 25, so two elements to the right would be atomic number 27, which is iron.
>
> The sequence in that part of the periodic table is:
>
> * Manganese (Mn); atomic number 25
> * Iron (Fe): atomic number 26
> * Cobalt (Co): atomic number 27
>
> So cobalt is the element that's two positions to the right of manganese.

it "know" where the pills are stored? Maybe yes, maybe no. It will be useful to ask this question, giving a different answer for the robot that will straightaway roll to the relevant drawer than for the robot that will need to launch a physical search. We can even try the classic "false belief" task from developmental psychology. Take a robot that we would normally say "knows" where the pills are, then surreptitiously move the pills while the robot is away. It will be natural to say that it "thinks" or "believes" that the pills are in the drawer where it left them yesterday, and we can thereby explain why it moves toward that drawer at pill-taking time. If this robot is sophisticated enough in other ways, for example linguistically, it would meet reasonable behavioral criteria for belief*. Not only will it search there, it will tell (or tell*) others that the pills are there; it will say (or say*) "The pills are supposed to be here; did somebody move them?"; it will initiate a search, or call for help, or at least refrain from the pointless action of offering an empty hand to the client at pill time. Similarly we might say that delivery bots do, or don't, know* or believe* that some particular sidewalk is closed for construction, and that companion bots, do or don't, remember* or believe* that March 17 is their friend's birthday, and so on, predicting and explaining their behavior accordingly.

Of course, it's hard to pronounce that asterisk, even if one is scientifically or philosophically convinced that robots are nonconscious and have the wrong underlying cognitive structures and thus cannot literally be said to believe. It's easier to be loose and sloppy, just adapting our comfortable old terms for this new use, until eventually it seems no longer like analogy or extension, but just the literal meaning. You and I might still wonder whether robots of this sort "really believe", but if technology proceeds in the direction I've described, that will probably come to seem like a philosopher's quibble.

That is how we will decide that robots "believe".

Have we fallen back into behaviorist instrumentalism? No, not about belief (hence the scare quotes around "believe" in the title and previous paragraph), only about belief*. In *our current* language – or, at least, in the language that I hope you and I can share – such systems would be at best in-between cases of believers, despite being paradigm believers*. What I am imagining is a future evolution of language in which "belief" shifts its meaning. Presumably, people will continue to care about the presence or absence of consciousness and will debate its existence or nonexistence in robots, but we will need words other than "belief" to track such distinctions.

5. Realism or Anti-Realism about Belief?

Krzysztof Poslajko (2024) and other "anti-realists" about belief should welcome all this. On anti-realist views, there is no underlying metaphysical fact about what belief "really is" independent of our usage. No metaphysical anchor prevents us from changing our usage in whatever way is convenient. The concept is "plastic" and "apt for serious semantic changes" (p. 176). Neither do our belief-attribution practices need to track the results of cognitive science, from which they are "autonomous" (p. 182). As long as belief-ascription practice is useful – as Poslajko thinks it is – and as long as we embrace a pragmatic metaphilosophy – as Poslajko does – we should welcome its effective adoption to new uses. Fictionalists (e.g., Toon 2023) and interpretativists (e.g., Curry 2020), should adopt a similar attitude. However, Poslajko mentions "artificial systems" only briefly in passing (p. 174), so it's hard to know whether he would indeed reach similar conclusions about belief or belief*. I invite him to expand upon this issue.

Although Poslajko cites my phenomenal dispositionalist view as an inspiration for his view (p. 73), he portrays himself as disagreeing with me on a crucial point. I hold, he says, that

the ordinary folk psychological concept of belief is superficial, and thus beliefs exist in the ordinary sense as long as superficial patterns of the right sort exist (p. 116). My view then counts as realist, though with a light touch. On his view, the ordinary folk psychological concept of belief is deep and requires the existence of "underlying mental causes" (p. 121) and "stored internal states" (p. 123). If there are no such causes and states, or if no such states exist as robustly real "natural kinds", then our current concept of belief "is misleading" and "requires conceptual revision" (p. 128). Still, Poslajko argues, beliefs exist and belief ascriptions can be true and useful. Ordinary people are mistaken only about the metaphysical kind to which beliefs belong and how robustly real they are. Poslajko's anti-realism is thus also light-touch.

I suspect the truth about the folk psychology of belief lies somewhere in the middle – containing conflicting strands and metaphysically open-textured. Poslajko notices a real tendency in the folk to treat beliefs as discretely stored, deep underlying causes of our dispositions. Yet also, as Curry (2019) and Gauker (2021) note, folk psychological thinking has superficialist elements too, especially if we think about how people would apply the concept of belief to hypothetical cases where the deep structure is postulated to be different or absent. By emphasizing different aspects of our informal understanding of belief, we can rigorize the concept in different ways – and we can then appeal to pragmatic grounds for rigorizing in one way or another.

The oft-cited fact that people are widely held to act as they do "because" they believe as they do (Egan 1995, cited in Poslajko, p. 118-119) is no evidence at all in favor of the view that ordinary people hold a deep, causal notion of belief. "Because" explanations can be dispositional (the glass broke because it was fragile) and even entirely non-causal (the slope of the curve increases because the equation is exponential).

If Poslajko is willing to accept that the ordinary conception of belief is multifaceted in a way that makes dispositionalism one acceptable rigorization of it, the gap between Poslajko's view and my own narrows. Any remaining gap lies, I suspect, mainly in the metaphysical respect I have for "unnatural" folk kinds (cf. Dupré 1995 on fish), including belief as real enough, and in my inclination to accept the importance of patterns of phenomenal consciousness to belief.

Still, insisting on the centrality of consciousness to belief might soon begin to look old-fashioned and, perhaps ironically, not superficial enough.[2]

[2] For helpful discussion, thanks to Brice Bantegnie, Sophie Nelson, Adam Pautz, Jeremy Pober, Krzysztof Poslajko, and William Robinson.

References

Bantegnie, Brice, and Eric Schwitzgebel (forthcoming). The metaphysics of beliefs: Dispositional theories. In T. Lombrozo and N. Van Leeuwen, eds., *The Oxford handbook of the cognitive science of belief.* Oxford University Press.

Butlin, Patrick, Robert Long, et al. (2023). Consciousness in Artificial Intelligence: Insights from the science of consciousness. *ArXiv* 2308.08708. URL: https://arxiv.org/abs/2308.08708

Chalmers, David J. (1996). *The conscious mind.* Oxford University Press.

Chalmers, David J. (2023). Could a Large Language Model be conscious? *Boston Review* (Aug. 9). URL: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious

Colombatto, Clara, and Stephen M. Fleming (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness, 2024* (1) niae013.

Curry, Devin Sanchez (2018). Beliefs as inner causes: The (lack of) evidence. *Philosophical Psychology, 31,* 850-877.

Curry, Devin Sanchez (2020). Interpretativism and norms. *Philosophical Studies, 177,* 905-930.

Davidson, Donald (1984). *Inquiries into truth and interpretation.* Oxford University Press.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2017). What is consciousness, and could machines have it? *Science, 358* (6362), 486-492.

Dennett, Daniel C. (1987). *The intentional stance.* MIT Press.

Dretske, Fred (1995). *Naturalizing the mind.* MIT Press.

Dupré, John (1995). *The disorder of things.* Harvard University Press.

Egan, Frances (1995). Folk psychology and cognitive architecture. *Philosophy of Science, 62,* 179-196.

Fischer, Eugen, and Justin Sytsma (2021). Zombie intuitions. *Cognition, 215,* 104807.

Gauker, Christopher (2021). Belief attribution as indirect communication. In L. Koreň, H. B. Schmid, P. Stovall, and L. Townsend, eds., *Groups, norms and practices.* Springer.

Godfrey-Smith, Peter (2016). Mind, matter, and metabolism. *Journal of Philosophy, 113,* 481-506.

Graham, George (2000/2023). Behaviorism. *Stanford Encyclopedia of Philosophy*. Spring 2023 edition.

Kirk, Robert (1974). Zombies v. materialists. Proceedings of the Aristotelian Society, 48 (supplementary), 135-152.

Levin, Janet (2004/2023). Functionalism. *Stanford Encyclopedia of Philosophy.* Summer 2023 edition.

Lewis, David K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50,* 249–258.

Mölder, Bruno (2010). *Mind ascribed*. Johns Benjamins.

Pautz, Adam (2021a). Consciousness meets Lewisian interpretation theory: A multistage account of intentionality. In U. Kriegel, ed., *Oxford Studies in Philosophy of Mind, vol. 1.* Oxford University Press.

Pautz, Adam (2021b). Varieties of interpretationism about belief and desire. *Analysis, 81,* 512-524.

Pober, Jeremy (2024). *Hybrid functionalism.* Unpublished manuscript.

Poslajko, Krzysztof (2024). *Unreal beliefs.* Bloomsbury.

Putnam, Hilary (1975). The meaning of "meaning". *Minnesota Studies in the Philosophy of Science, 7,* 131-193.

Quilty-Dunn, Jake, and Eric Mandelbaum (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies, 175,* 2353-2372.

Robinson, William S. (2019). *The epiphenomenal mind.* Routledge.

Schwitzgebel, Eric (2002). A phenomenal, dispositional account of belief. *Noûs, 36,* 249-275.

Schwitzgebel, Eric (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann, ed, *New essays on belief.* Palgrave.

Schwitzgebel, Eric (2021). The pragmatic metaphysics of belief. In C. Borgoni, D. Kindermann, and A. Onofri, eds., *The fragmented mind.* Oxford University Press.

Schwitzgebel, Eric (2024). *The weirdness of the world.* Princeton University Press.

Schwitzgebel, Eric (forthcoming). Dispositionalism, yay! Representationalism, boo! In J. Jong and E. Schwitzgebel, eds., *The nature of belief.* Oxford University Press.

Seth, Anil (2024). Conscious artificial intelligence and biological naturalism. *PsyArXiv* tz6an. URL: https://doi.org/10.31234/osf.io/tz6an

Strawson, Galen (1994). *Mental reality.* MIT Press.

Toon, Adam (2023). Mind as metaphor. Oxford University Press.

Turing, Alan (1950). Computing machinery and intelligence. *Mind, 49,* 433-460.