Fictionalism about Personas: Folk Psychology as an Interpretability Strategy

Weiming Sheng [1]

Abstract

Folk psychology is an effective tool for predicting and interpreting LLM systems, and the recent literature characterizes this success through the concept of a *persona*. But what does it mean for an LLM to take on a persona? *Realism* says the model instantiates the persona's mental states; *simulationism* says the model simulates the persona; and *fictionalism* says personas are fictional characters generated by the model. Realism cannot explain how such mental states would arise from training, and simulationism, taken literally, raises a series of difficult conceptual questions. *Assistant fictionalism* accounts for the same data without these costs, and I show that it withstands two objections. The view has the practical upshot that interpretability work should distinguish properties of the model from properties of the characters it generates.

1. Introduction

Contemporary large language models (LLMs) generate outputs which often look as if they are produced by a *minded* being. That is, when interacting with LLM systems, it is natural to attribute mental states like belief, desire, and intention to them, and to predict what they will output based on these states. This offers the potential for a new interpretability strategy when dealing with LLM systems: we can explain and make sense of their outputs using mental state concepts, thereby achieving an *intentional* understanding of such systems.

Call the practice of attributing mental states and making predictions based on them *folk psychology*. Our disposition to do folk psychology on LLM systems could perhaps be explained away as mere anthropomorphism, if they did not work *so well*. A concrete example is provided by the recent Chain-of-Thought (CoT) monitoring literature, where the CoT of a model is monitored against harmful output or reward hacking (Korbak et al., 2025; Baker et al., 2025). CoT is useful for monitoring only insofar as it provides a folk-psychological explanation of what the Assistant is about to output. For instance, a reward hacking model might say in its CoT that it *wants* to skip writing this part of the code and *intends* to trivialize the unit test. That CoT monitoring can do load-bearing work in industry settings (Guan et al., 2025) show the effectiveness of folk psychology for LLM systems.

Why are LLM systems so amenable to folk psychology? Many have found the concept of *persona* helpful for this question (Marks et al., 2026; nostalgebraist, 2025; Shanahan et al., 2023; Chen et al., 2025; Lu et al., 2026). A persona is roughly a set of personality traits or a "psychological profile" (Chalmers, 2026). More precisely, we can say that a persona is a distribution over mental states conditional on context. The Assistant in user-assistant dialogues in chat-bots like Claude serves as the paradigmatic example of a persona: it is characterized by personality traits such as helpfulness and honesty, as well as a broad knowledge and deep technical abilities. Another example of persona comes from the use of LLMs in social-survey simulation, where models are conditioned on demographic or social-background descriptions and asked to answer survey questions as members of those subpopulations (Argyle et al., 2023).

This paper focuses on the question: *what does it mean to say an LLM takes on a persona?* Answering this question clarifies the status of folk psychology as an interpretability strategy for LLM systems. I distinguish three answers to the question:

Realism LLMs *realize* or *instantiate* the psychological profile associated with a persona.

Simulationism LLMs produce the personas as *simulacra*.

Fictionalism LLMs write texts that feature *fictional characters*; personas are the special case where such description is first-personal.

In section 2 I argue that realism is untenable because of the *acquisition problem*. In section 3 I consider simulationism: it is a good metaphor, but taking it seriously involves us with

[1]Department of Philosophy, Columbia University, New York, U.S.. Correspondence to: Weiming Sheng <ws2720@columbia.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

conceptual difficulties. In section 4 I discuss fictionalism and show how to answer two challenges for the position.

Going forward, by an LLM I mean a generative language model with parameter count in at least the billions, trained to minimize some loss on a text corpus. Such a model defines a next-token distribution $p_\theta(w_k \mid w_{<k})$ from which we can sample autoregressively. By an LLM system I mean an LLM together with harnesses such as a system prompt, chat interface, or memory, configured to produce User-Assistant dialogues.

2. Realism and the Problem of Acquisition

Realism says that, for the model to take on a persona that has tendencies to certain beliefs and desires just is for the model to have those tendencies to beliefs and desires. In other words, the persona talk is mostly a way to categorize mental states and personality traits; to say that a model takes on a persona is just to attribute such mental states and personality traits to the model (Keeling & Street, 2025; Frankish, 2024; Lederman & Mahowald, 2024; Goldstein & Lederman, 2025; Keeling & Street, 2026; Herrmann & Levinstein, 2024; Goldstein & Levinstein, 2024; Chalmers, 2026; Schwitzgebel, 2023).

Realism faces a hard challenge:

> **Problem of Acquisition:** What is it about the *design and training* of LLMs that gives rise to their mental states?

How realism is spelled out will not matter much for our discussion. In particular, the realist may take the "LLM" to refer to either the model itself or to an instance of the model, which could either be a hardware instantiation or a virtual instantiation (Chalmers, 2026). In the latter case, the problem of acquisition would be: what is it about the design and training of LLMs so that instances of it have mental states? For simplicity, I will speak of models and LLMs, but there is no difficulty in making the appropriate adjustments for the instance view.

I consider two answers to the question above: (1) LLMs are *explicitly* designed to have mental states; (2) having mental states is *instrumental* for LLMs to achieve their design goals. I will show that neither satisfactorily answers the question.[1]

[1]A potential third option: having mental states is an *accidental* feature of LLMs achieving their design goals. Goldstein & Levinstein (2024) mention this and compare it to an LLM's "emergent" ability to do in-context learning (Wei et al., 2022). But if the presence of mental states is to be *explained*, it cannot be a mere accident, and emergence is not an explanation.

2.1. Explicit goal

The basic goal in training an LLM is to minimize a loss function on some data. In contemporary chat-model training, two loss functions are typically used. First, a *cross-entropy loss* adjusts the model to assign higher probability to sequences in the data. This is applied to data scraped from the internet and to curated user-assistant data:

$$\arg\min_\theta -\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log p_\theta(x_t \mid x_{<t})\right].$$

DPO loss (Rafailov et al., 2024) is then minimized, where

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}}[\log\sigma(r_\theta(x,y_w) - r_\theta(x,y_l))]$$

and $r_\theta(x,y) := \beta\log\frac{p_\theta(y|x)}{p_{\text{ref}}(y|x)}$ is the implicit reward, with y_w and y_l the preferred and dispreferred responses to prompt x, and p_{ref} the frozen reference policy.

In both cases, the explicit goal is simply to *minimize the loss function on some data*. This goal has nothing to do with mental states. Minimizing these loss functions makes certain outputs more likely than others, and that is all. Folk psychology does not enter the picture (Bender & Koller, 2020; Shanahan, 2024b).

2.2. Instrumental convergence

Grant that LLMs are not explicitly designed to have mental states. Still, having mental states might be *useful* for minimizing loss. For instance, *belief*, understood as the ability to track the truth of sentences, seems useful for reducing both cross-entropy and DPO loss. Pursuing this thought, Levinstein & Herrmann (2025) examined whether the internal activations of LLMs co-vary in some simple and generalizable way with the truth-value of prompts. Their conclusion is negative, but newer studies, e.g. Ying et al. (2026), seem to show that such co-variance exists and can be detected simply. In any case, *some* non-trivial co-variance between internal activations and truth is unsurprising, since state-of-the-art LLMs perform very well on factual benchmarks like MMLU (Hendrycks et al., 2020): the co-variance patterns might not be detectable by a linear probe, but that is not evidence against their reality. Can we then explain LLMs' purported mental states by appealing to their usefulness in minimizing loss?

I think not. The basic problem with the instrumentalist strategy is that it cannot justify most of the predictive success of folk psychology for LLM outputs. Truth-tracking may be useful for loss reduction, but most folk-psychological concepts we apply to LLMs are not so plausibly useful. Consider emotion concepts. Ascriptions like desperation or anger predict systematic and measurable differences in LLM outputs (Sofroniew et al., 2026). But why would it be

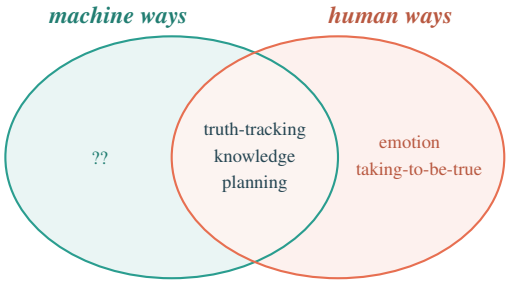

Figure 1. Two kinds of mechanisms

useful for the model to have emotions like these to reduce either cross-entropy or DPO loss? There seems to be no plausible story for how a desperate or angry model could be better at reducing loss; if anything, you would think that the model would do best to stay calm—but that is really just to say, without emotion.

The point is that there are two kinds of internal mechanisms: machine ways and human ways (figure 1). The machine ones developed to reduce loss on the training data. The human ones, as picked out by folk psychology, presumably developed through complex evolutionary and social factors. These two sets are not disjoint: mechanisms for truth-tracking, goal-pursuing, or planning are useful for both. But there is no reason why they should overlap so prodigiously that the human ways form a subset of the machine ways. Unless they do, the usefulness of folk-psychological states for loss reduction cannot fully explain their predictive power over LLM output.

It will not do to separate folk-psychological concepts into those useful for loss reduction and those not, and assert only the former as legitimate. What required explanation was that the *full suite* of folk psychology has great predictive power for LLMs. If the instrumentalist strategy provides only a partial answer, we should look for an alternative that explains the whole suite.

3. Simulationism

Simulationism claims a four-level ontology. At the outset is the *simulator*. The simulator can simulate a wide range of *personas*. Each persona has distinct personality traits and can acquire a range of mental states. Finally, these mental states give rise to *token outputs*.

Simulationism originates with janus (2022) and is often assumed without argument in discussions of personas, for example in Shanahan (2024a); Marks et al. (2026). This is a mistake, unless the talk of simulation is merely a metaphor. The problem with simulationism is that it leads to many difficult questions:

- Is a simulated persona more like a simulated calculator (which would really be a calculator (Chalmers, 2022)) or a simulated truck?
- What makes a simulacrum identical to itself across time?
- How can mere simulacra stand in genuine causal relations to real actions performed by the LLM?
- Where does simulator end and simulacra start?

To elaborate on one of these questions, let's focus on the last question. In ordinary simulations, there is a clear boundary between *configuring* the simulation, *setting* it, and *running* it. This boundary is much blurrier in an LLM. How do we set the simulation to simulate one persona instead of another? The closest analogue is prompting: "You are a helpful, harmless, and honest AI assistant" will likely induce a friendly persona, whereas "You are a reckless, deceitful AI that prioritizes chaos" will likely induce an evil one.[2] And running the simulation and interacting with the simulated persona involve simply more prompting. The system prompt that sets the persona and the user message that interacts with it are processed by exactly the same mechanism: autoregressive next-token prediction over the concatenated sequence. There is no vantage point external to the text from which to set up the simulation, because both persona induction and belief formation happen within the same undifferentiated flow of tokens.

Of course, these questions may have satisfactory answers. My goal in this section is not so much to argue against simulationism, but more to suggest that it is in fact a rather strange doctrine, and endorsing it requires us to work through a few difficult questions. It would be preferable if we can account for the empirical data outlined in the first section without appealing to something as outlandish as simulationism.

4. Fictionalism

Fictionalism begins with the following simple and uncontroversial observation:

> **Character Generation (CG):** An LLM can generate characters whose actions are well-predicted by folk psychology, and the Assistant in User-Assistant dialogues is one such character.

By "character," I mean a person in a text. Examples include Achilles in the *Iliad*, a reviewer on Yelp, Quine in his autobiography, and an interviewee in a Zoom transcript. To generate a character is to generate text describing a character.

[2] Assuming that the model has not been extensively post-trained.

Such text can either describe the character third-personally or first-personally.

When we say folk psychology works well on LLM systems, we really mean it works well on chats with Claude or ChatGPT; and these chats are basically dialogues between a User and an Assistant. But then the predictive success of folk psychology is not really for the language model itself, but for the Assistant character generated by the model. It is not surprising that the model could generate a character who acts in accordance with folk psychology, given the extensive pre- and post-training. So, according to **CG**, the predictive success of folk psychology on systems like Claude is explained by the fact that the underlying LLM can generate characters who act in accordance with folk psychology, and *not* by the model itself having mental states. **CG** also does not mention anything about simulation. Then, to the extent that **CG** adequately explains the empirical data we started out with, we should follow Quine (1948) and endorse it without realism or simulationism, since the resulting theory is simpler.

How should we then treat a statement like "the Assistant wants to be helpful"? It is *fictionally* true: literally false, but true in the game of make-believe we play with LLM systems like Claude. In this game, we are to imagine, after we write and send our part of the conversation, that a helpful, polite, etc. AI assistant is on the other side reading the conversation and responding in ways appropriate to such an entity.[3] Call **CG** plus anti-realism via fiction *Assistant fictionalism*.

> **Assistant fictionalism:** **CG** is true, realism is false, and characters like the Assistant are *fictional*.

4.1. Objection: Shoggoth Problem

Consider how a novelist might write a compelling character. She puts herself in the character's role and imagines how she would feel in their situation. Similarly, when we explain others' actions using folk psychology, we often put ourselves in their shoes and imagine their mental states. This is the simulation theory (ST) for mindreading (Goldman, 2008). If ST is correct as an account of mindreading, then doing folk psychology requires *having* folk-psychological states and being able to simulate their occurrence. Then insofar as LLMs can generate folk-psychologically plausible characters, LLMs themselves must have folk-psychological states.

Response: There are three main reasons why ST is widely accepted, and none carries over to LLMs. (1) ST enjoys *phenomenological support*: introspection reveals we often imaginatively project ourselves into others' situations rather than applying tacit laws. This is moot for LLMs, which cannot introspect—or, if they can, the question of realism would already be settled. (2) ST is *parsimonious*: "there is no need to invoke the tacit knowledge of a Theory of Mind to account for mindreading, since a more parsimonious explanation is available: we reuse our own cognitive mechanisms to mentally simulate others' mental states" (Barlassina & Gordon, 2017). This is again moot for LLMs because the more parsimonious explanation here is something like TT, not ST. We already know that LLMs acquire knowledge of statistical patterns resembling a folk-psychological theory from training; it would be theoretically costly to say in addition that LLMs have the capacity for all folk-psychological states and can imaginatively employ them for mindreading. (3) ST enjoys *empirical support*. A famous though disputed piece of evidence is the discovery of "mirror neurons" that fire both during action execution and observation (Rizzolatti & Sinigaglia, 2016). The empirical findings for ST so far are relevant only to the human case. I doubt further ST-related empirical work directed at LLMs could help, since such work would have to *assume* that LLMs have mental states and test whether these are employed in mindreading—but whether these mental states exist is the question at hand.

4.2. Objection: Explanatory Collapse

If mental ascriptions to the Assistant are merely fictional, how can they explain real facts such as emails being sent or PRs being created (Goldstein & Kirk-Giannini, 2026)? In other words, since fictional truths cannot causally explain actual truths (Hempel & Oppenheim, 1948; Lewis, 1986), fictional ascriptions to the Assistant cannot account for the empirical success of folk psychology, which we saw in section 1.

Response: There are two ways for a fictional statement to pick out a fact: through its literal content or through its real content. Take, for example, the fictional statement that "there is an angry face in the sky." The literal content is the set of possible worlds in which there is *literally* an angry face floating in the sky. The actual world is not in such a set, so the statement is false at the actual world. The real content is the set of possible worlds which make the statement properly fictional, i.e. "pretence-worthy in the relevant game" (Yablo, 1998). The actual world, if it features a cloud with this resemblance, is part of this set; a world without clouds would not be.

We explain actual facts using the real content of fictional truth all the time. A student comes to me and says: "Sorry for not submitting the homework on time. My computer just *decided* to crash." Her computer did not literally decide to do anything: this is a figure of speech, a fiction. But the sentence "my computer just *decided* to crash" does pick out actual facts: roughly, the facts that the computer crashed

[3] See Walton (1993) for the general theory of fiction.

without warning, that the student could not have anticipated it, and that therefore she is not responsible for it. I should not reject the student's explanation because she said something *literally* false, though I could reject it if I suspect the real content of the statement is not true—for instance, if I know the student could have anticipated the crash.

The fact picked out by "the Assistant wants to be helpful," through its real content, is the obtaining of some internal activation pattern from the set of patterns associated with the model generating a helpful Assistant character. When "the Assistant wants to be helpful" causally explains why an email was sent, it is really this activation pattern that causally explains the email being sent.

5. Conclusion

We are in a paradoxical situation. Language models are larger and less transparent than ever, and yet the Assistant character, partly in virtue of that very opacity, is often as interpretable through folk psychology as our fellow humans. Assistant fictionalism dissolves the paradox. The success of folk psychology does not require mental ascriptions to the model or some simulacrum, but only to the characters it generates. And we already know how to deal with characters: it is something we have been doing, with novels and plays and films, for as long as we have had them.

The realism discussed in this paper is part of a larger trend: as LLMs grow more capable, the temptation to make ever more ambitious attributions to them will grow in step. One lesson of the paper is that we should be careful to distinguish attributions to the model from attributions to a character the model generates. The two come apart. A fictional character can be conscious, wise, anxious, or in love, without any of these being properties of the model generating the character. Call it *the Assistant fallacy*: inferring that a property holds of the model from the fact that it holds, fictionally, of the Assistant. Something like this fallacy seems to me already prevalent in the philosophical literature on AI, and I hope this paper gives some reason for avoiding it.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, July 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2.

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation, March 2025. URL http://arxiv.org/abs/2503.11926. arXiv:2503.11926 [cs].

Barlassina, L. and Gordon, R. M. Folk Psychology as Mental Simulation. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017. URL https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/.

Bender, E. M. and Koller, A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Chalmers, D. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton, New York, 2022.

Chalmers, D. J. What we talk to when we talk to language models. Manuscript, 2026.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, September 2025. URL http://arxiv.org/abs/2507.21509. arXiv:2507.21509 [cs].

Frankish, K. What are large language models doing? In *Anna's AI Anthology. How to live with smart machines?* Xenomoi Verlag, Berlin, 2024.

Goldman, A. I. *Simulating minds: the philosophy, psychology, and neuroscience of mindreading*. Philosophy of mind. Oxford Univ. Press, Oxford, 1. issued as paperback edition, 2008. ISBN 978-0-19-536983-0 978-0-19-513892-4.

Goldstein, S. and Kirk-Giannini, C. D. *Ai Welfare: Agency, Consciousness, Sentience*. Oxford University Press, New York, 2026. Forthcoming.

Goldstein, S. and Lederman, H. What does chatgpt want? an interpretationist guide. Manuscript, 2025.

Goldstein, S. and Levinstein, B. A. Does chatgpt have a mind? *Philosophy of Ai*, 2024. Forthcoming.

Guan, M. Y., Wang, M., Carroll, M., Dou, Z., Wei, A. Y., Williams, M., Arnav, B., Huizinga, J., Kivlichan, I., Glaese, M., Pachocki, J., and Baker, B. Monitoring Monitorability, December 2025. URL http://arxiv.org/abs/2512.18311. arXiv:2512.18311 [cs].

Hempel, C. G. and Oppenheim, P. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175, 1948. ISSN 0031-8248. URL https://www.jstor.org/stable/185169.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

Herrmann, D. A. and Levinstein, B. A. Standards for Belief Representations in LLMs. *Minds and Machines*, 35(1): 5, December 2024. ISSN 1572-8641. doi: 10.1007/s11023-024-09709-6. URL http://arxiv.org/abs/2405.21030. arXiv:2405.21030 [cs].

janus. Simulators — LessWrong. September 2022. URL https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators.

Keeling, G. and Street, W. On the attribution of confidence to large language models. *Inquiry*, pp. 1–27, January 2025. ISSN 0020-174X, 1502-3923. doi: 10.1080/0020174X.2025.2450598. URL http://arxiv.org/abs/2407.08388. arXiv:2407.08388 [cs].

Keeling, G. and Street, W. What's it like to be a chat? On the co-simulation of artificial minds in human-AI conversations, January 2026. URL http://arxiv.org/abs/2601.13081. arXiv:2601.13081 [cs].

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Madry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, December 2025. URL http://arxiv.org/abs/2507.11473. arXiv:2507.11473 [cs].

Lederman, H. and Mahowald, K. Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of Llms. *Transactions of the Association for Computational Linguistics*, 12:1087–1103, 2024.

Levinstein, B. A. and Herrmann, D. A. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7):1539–1565, July 2025. ISSN 0031-8116, 1573-0883. doi: 10.1007/s11098-023-02094-3. URL https://link.springer.com/10.1007/s11098-023-02094-3.

Lewis, D. Causal Explanation. In *Philosophical Papers, Volume II*, pp. 214–240. OUP, 1986.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models, January 2026. URL https://arxiv.org/abs/2601.10387v1.

Marks, S., Lindsey, J., and Olah, C. The persona selection model: Why ai assistants might behave like humans. February 2026. URL https://alignment.anthropic.com/2026/psm/. Anthropic Alignment Science Blog.

nostalgebraist. the void, July 2025. URL https://nostalgebraist.tumblr.com/post/785766737747574784/the-void.

Quine, W. V. On What There Is. *The Review of Metaphysics*, 2(5):21–38, 1948. ISSN 0034-6632. URL https://www.jstor.org/stable/20123117.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

Rizzolatti, G. and Sinigaglia, C. The mirror mechanism: a basic principle of brain function. *Nature Reviews. Neuroscience*, 17(12):757–765, December 2016. ISSN 1471-0048. doi: 10.1038/nrn.2016.135.

Schwitzgebel, E. How We Will Decide that Large Language Models Have Beliefs, November 2023. URL https://schwitzsplinters.blogspot.com/2023/11/how-we-will-decide-that-large-language.html.

Shanahan, M. Simulacra as Conscious Exotica, July 2024a. URL http://arxiv.org/abs/2402.12422. arXiv:2402.12422 [cs].

Shanahan, M. Talking about Large Language Models. *Communications of the ACM*, 67(2):68–79, February 2024b. ISSN 0001-0782, 1557-7317. doi: 10.1145/3624724. URL https://dl.acm.org/doi/10.1145/3624724.

Shanahan, M., McDonell, K., and Reynolds, L. Role play with large language models. *Nature*, 623(7987):493–498, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06647-8. URL https://www.nature.com/articles/s41586-023-06647-8.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., and Lindsey, J. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Walton, K. L. *Mimesis as make-believe: on the foundations of the representational arts*. Harvard University Press, Cambridge, Mass., 1. paperback ed edition, 1993. ISBN 978-0-674-57603-2 978-0-674-57619-3.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent Abilities of Large Language Models, October 2022. URL http://arxiv.org/abs/2206.07682. arXiv:2206.07682 [cs].

Yablo, S. Does Ontology Rest on a Mistake? *Aristotelian Society Supplementary Volume*, 72(1):229–283, 1998. URL https://philarchive.org/rec/YABDOR.

Ying, Z. J., Ravfogel, S., Kriegeskorte, N., and Hase, P. The Truthfulness Spectrum Hypothesis, February 2026. URL http://arxiv.org/abs/2602.20273. arXiv:2602.20273 [cs].