Don't Believe the Belief Hype!

Alessandro Corona Mendozza
Department of Nordic Studies and Linguistics
University of Copenhagen
Copenhagen, Denmark
alessandro.mendozza@hum.ku.dk

Abstract

Recent work in interpretability has suggested that language models contain 'belief-like' internal representations, based on evidence that intermediate activations encode true from false statements separately. We argue that this belief narrative is misleading. First, the concept of belief is theoretically opaque: philosophy and cognitive science offer fragmented, incompatible accounts, leaving no stable source for a robust implementation. Second, intermediate activations are not privileged 'internal states' in a sense that would justify belief ascriptions. Third, probing results show at best that model states are *truth-sensitive* but this does not imply they are *truth-representing*. Sheer sensitivity is ubiquitous in simple systems and does not warrant talk of belief. Finally, we show how careless belief-talk risks both theoretical and practical confusion. As an alternative, we advocate a conceptual engineering approach: coin new, precise terms for truth-sensitive representations, rather than stretching an overloaded, obscure notion of belief.

Introduction

Recent interpretability studies have shown that language models can separate prompts by truth or falsity [Burns et al., 2024, Li et al., 2024, Marks and Tegmark, 2024, Bürger et al., 2024]. The idea that language models can have an internal representation of truth has been conflated with the claim that their internal activations have some form of 'belief-likeness', and a set of tests was operationalized to strengthen these claims [Herrmann and Levinstein, 2024]. We argue against framing these findings in terms of 'belief-likeness'. Our observations span (1) theoretical concerns rooted in the opacity of the concept of belief, which is fragmented to date both in philosophy of mind and cognitive science; (2) concerns about the adequacy of intermediate activations as a proxy for beliefs; (3) theoretical and practical risks that we incur when carelessly ascribing beliefs to language models. After considering these issues, (4) we advocate for a different approach that we should pursue for a better conceptual clarity regarding the study of this property that was found in intermediate activations.

1 The opacity of belief

No science for beliefs Herrmann and Levinstein [2024], when looking for their set of *desiderata* that intermediate representations have to satisfy to count as belief, ground their first intuitions in a set of mathematical models of human beliefs and behavior such as decision theory [Jeffrey, 1965], formal epistemology [Genin and Huber, 2022] or radical interpretation [Davidson, 1973]. These capture aspects of human belief behavior, but say little about what beliefs *are*. Unfortunately, cognitive science lacks a unified account of belief. While the concept does appear in multiple domains, such as attribution theory, theory of mind, predictive coding or bias study, a thorough science of human belief is still in its infancy [Porot and Mandelbaum, 2021]. For now, the best view that we have is a

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability Workshop.

fragmented landscape, with isolated implementations of the notion in human agents but no consensus definitions for mapping that to non-human systems [Schwitzgebel, 2024]. This leaves any empirical inquiry into other systems lacking theoretical guidance other than mere folk psychology [Goldstein and Levinstein, 2024], which is arguably the most common level of analysis that justifies bold claims about AI cognition and other metaphorical ascription of mental capabilities like 'knowledge' or 'reasoning'. We should not expect folk physics to have the same insightfulness of actual physics for machine learning practices: in the same way, folk psychology is more a liability than an asset when we want to assert that a certain property is visible in a system.

A bouquet of philosophical theories Expert disagreement flourishes even more when we consider competing philosophical theories on propositional attitudes. As shown by a recent PhilPapers report [Bourget and Chalmers, 2023] the current orthodoxy has a balanced split between *representationalists* (46.5%), claiming that beliefs are relations to internal representational states (mental content with propositional structure, e.g. sentences in the language of thought [Fodor, 1975] or mental maps of the world [Lewis, 1994]), and *dispositionalists* (31.5%), understanding beliefs as dispositions to behave, reason, and feel in characteristic ways, without positing internal representational vehicles [Schwitzgebel, 2013]. Furthermore, if we aim to map mental states to other systems, some shade of functionalism [Levin, 2023] should be adopted too. The problem becomes obvious when we explore how each of those frameworks bud off: say, for example, that we want to adopt a simple version of functionalism [Bratman, 1999] where belief is co-defined by its interaction with desires and intentions. This move would exclude a large set of legitimate and well-argued philosophical theories that would deny that this simple scaffolding is enough for an entity to count as an outright belief, for instance by pointing that it lacks the right representational structure. The scientific fragmentation of the notion of belief has a philosophical counterpart that already led to strong eliminativist pushbacks [Churchland, 1981] and actively impedes empirical analysis on non-trivial, non-metaphorical levels.

Going with the intentional stance Given this outlook, a possible fallback is to adopt a very thin notion of belief. The most liberal theory that we have for propositional attitudes is interpretationism, with Dennett's framework being the most popular choice [Dennett, 1971, 1981]. In short, interpretationists claim that when a system is too complex to be fruitfully understood by considering its design (e.g. what kind of algorithm it implements) or its physical properties (e.g. how many molecules are there), our best shot at interpreting it is by adopting the *intentional stance*, i.e., understanding its behavior through mindreading. By this account, beliefs are useful abstractions that we cast on the system for our epistemic purposes. There are several problems with this form of interpretationism [Searle, 1992]. We can seal the case by looking at the most relevant for our purposes: interpretationism it is *too loose* of a fit to accurately capture the kind of phenomenon that we saw in probing studies on truth-sensitive activations. The intentional stance can be applied to chess-playing symbolic algorithms or simple measuring devices. This leaves out the specificity of finding sentential representations that are truth-sensitive, and places latent stream activations in the same category as a thermostat. If we can ascribe beliefs by simply looking at the models' outputs, then probing the internals becomes an uninteresting task. From this theoretical fragmentation, we turn to more empirical limitations of treating intermediate activations as candidate bearers of belief.

2 The inadequacy of intermediate representations

Where truth is measured The subject of analysis for the studies that are concerned to find a direction of truth expressed by language models is typically the residual stream [Marks and Tegmark, 2024, Burns et al., 2024] or other intermediate states such as attention head pre-output representations [Li et al., 2024]. The working hypothesis is that if we can discriminate true from false prompts by looking at that level through a mix of probing, interventions or visualizations, and if we are successful on a series of further tests [Herrmann and Levinstein, 2024], then we can say that those representations, that treat the source sentences as being true or false, are, in a non-trivial sense, belief-like. This is because they seem to be *internal* to the model, as opposed to its utterances, and *truth-representing*, i.e., directly encoding whether the prompt is true or false. Each of these assumptions has some problems.

Misleading internalism Beliefs are usually taken to be *internal* to one's mind, with a notable exception [Clark and Chalmers, 1998]. By analogy, some interpretability work treats intermediate

activations as *internal* in a way that logits and outputs are not, making them better candidates for belief-like states. We think this is a mistake. The problem with this view is that intermediate representations are not *internal* to the model's weights, but merely mark intermediate states of processing from prompt embedding to inferred logits. They are not more intrinsic to the model than the logits themselves. Probing them is still an external analysis of the system's behavior, just at a different layer. To see this, imagine attaching a LayerNorm and unembedding matrix to an intermediate layer. The model would output logits at that point, and we would call them 'outputs' rather than 'internal states'. Conversely, we could take the final logits, re-embed them, and run them through more transformer layers. In that case, the outputs would become 'intermediate' states: the distinction between 'internal' and 'external' breaks down once we see that both are points in the same pipeline. Logits and earlier representations are the same type of entities, so if we want to ascribe belief-likeness to the latter, we should do the same for the former. If belief requires a genuinely internal property, these activations do not qualify. Calling them 'internal' in the sense of mental states misleads more than it clarifies.

Truth-sensitive vs truth-representing Another formidable obstacle in classifying intermediate activations as being belief-like lies in the severe underdetermination that probing datasets naturally embody: this has been acknowledged both by Marks and Tegmark [2024] and by Levinstein and Herrmann [2025]. While other problems from using probing classifiers can be mitigated or overcame, for example by coupling probing studies with control tasks [Hewitt and Liang, 2019] or interventions [Belinkov, 2022], this issue seems fundamentally inescapable. As Levinstein and Herrmann [2025] note, every dataset that is explainable in terms of a single feature is also explainable in terms of a set of correlated features. Such features, in the case of truth, can be easily inferred: 'commonly believed', 'commonly found in the training data', 'asserted in textbooks', and so on. This suggests that the best way to interpret language model representations is by seeing them as *truth-sensitive* more than *truth-representing*.

The problem with truth-sensitivity alone is that it is clearly not enough to qualify something as a belief, since it can be easily found in a lot of instruments that do not exhibit any form of cognition. To see why, imagine a simple and reliable polygraph, a lie detector responding to blood flow, small muscular twitches, and so on. In a sense, such a device is inherently truth-sensitive, since it is able to discriminate between false and true assertions from its users. Yet, the device clearly lacks the representational structure that relates those assertions to their truth or falsity, and thus fails to be truth-representing. The polygraph merely exploits correlations between bodily cues and deception, without encoding or representing what makes a statement true or false. Ascribing beliefs to such a system would therefore be, at best, purely metaphorical. Similarly, when a language model's intermediate representations respond to correlational features of text rather than encoding relations to truth conditions, they should not be understood as genuinely truth-representing. Belief, by contrast, presupposes some form of representational commitment to truth: an internal structure that purports to capture how things are and can be evaluated as accurate or inaccurate [Schwitzgebel, 2024]. Conflating truth-sensitivity with such commitment, in this underdetermined case, is a bade case of overreach. As we will see next, this conflation leads to both theoretical and practical complications in our received notions and our relation with AI agents.

3 Consequences of believing

The good There may be a number of reasons to continue speaking of intermediate activations as having a *belief-like* quality. For instance, doing that may provide a good abstraction that helps us understanding the inner workings of a transformer: beliefs have, in fact, a lot of explanatory power for our interpretation about other human agents, despite the theoretical fragmentation of the concept. Similarly, framing interpretability results in terms that resonate with human cognitive concepts can be an effective strategy for communicating relevant bits of information for broader audiences, particularly in discussions of AI alignment.[1] For example, knowing that certain activations may track truth in ways that diverge from the model's sampled outputs invites a healthy skepticism toward the assertion-like character of those outputs. In this sense, the belief-like framing offers the most compelling explanatory basis for efforts to build 'lie' detectors [Azaria and Mitchell, 2023, Burns et al., 2024].

[1]Thanks to one of the reviewers for pointing this out.

Some examples of overreaching However, opening the discourse to the possibility that language models have anything akin to human beliefs has some consequences from our analysis of such systems. First of all, the evidence collected for truth-sensitivity on intermediate activations opened the door for some bold claims about language model cognition and agency. For instance, Goldstein and Levinstein [2024] use the results to advocate for a mentalistic treatment of language models that can eventually bring support of an analysis of AI wellbeing and moral patiency Goldstein and Kirk-Giannini [2025]. In another domain, Williams and Bayne [2024] use the notion of *proto-belief* to set up the idea that language models could be understood as *proto-asserters*, performing speech acts and being legitimate testifiers in our social community. The ambitious thought-logging project that Chalmers [2025] laid down can also be listed across these lines. The problem here is not that accounts of AI cognition or agency are flawed in the first place; the problem is that a truth-sensitive representation that does not fit the bill except for the thinnest, metaphorical sense in which we talk about beliefs, should not be taken as evidence of sparks of cognition to construct theories that are even more loaded and risky.

The problem with conceptual borrowing This problem arises because belief is a rich concept from the inferential perspective: it has connections with epistemology [Dutant and Littlejohn, 2024], speech act theory [Lackey, 2008], ethics [Chignell, 2018], and so on. Being this connected, mapping the concept of belief to non-human systems leads to a paradigm case of conceptual borrowing [Floridi and Nobre, 2024]. This phenomenon happens when we graft a concept from a domain to another one and retain all the additional baggage and implications that the notion carries. While an operative notion of belief as a truth-sensitive representation may map well in language models, the implications may not be this easy to support, and we may develop a theoretically flawed entity. This also has the practical effect of spreading undue anthropomorphization without proper theoretical support. This can have vicious consequences for people who interact with language models and lack access to a clear-cut distinction between full-fledged human beliefs and 'beliefs' in the thinnest sense [Mlonyeni, 2024, Holbrook et al., 2024].

Theoretical clarity and pareidolia We can now weigh advantages and disadvantages of this intentional talk. On the more liberal side, we have a communication strategy that simplifies a complex object of inquiry and that fosters interest around a possibly relevant character of language models. On the other hand, a more conservative approach would claim that this oversimplification carries a baggage of bad consequences and that, being in its infancy, interpretability should stick to clear and theoretically sound explanations before extending its threads by rushed analogies. We stand by this last claim: interpretability researchers have the aim of providing explanations and tools to better understand the processing and behavior of black-box models. This is a different task than trying to fit large neural network in an AI narrative that specifically looks for human-like patterns in the systems. We lack the possibility of flagging intermediate activations as beliefs, both because the source theory is too scarce or confused to capture a single notion and because the evidential support from our studies is limited itself. If we want to adhere to the aim of clarity and adequate explanations, we should not give in to the temptation of simplicity and interesting upshots: a lesson that has already made clear by instant classics in our literature [Nanda et al., 2023, Wang et al., 2022].

4 Two conceptual engineering strategies

Amelioration by adapting human concepts The act of redefining the extension of old concepts to let them adapt to new entities is a matter of *conceptual engineering* [Isaac et al., 2022, Himmelreich and Köhler, 2022]. More specifically, it is the way in which we *ameliorate* deficitary concepts and correct the set of entities they range over for practical and theoretical reasons. The problem with amelioration, when it comes down to mapping anthropocentric notions such as belief to AI systems, is that those reasons should be stronger than the loss in theoretical clarity that we undergo due to conceptual borrowing. This is a historical challenge of the whole 'artificial intelligence' program, and it becomes more relevant the more hard to interpret and analyze AI systems become. For a healthy mapping between human systems and artificial systems we have to be conservative with each step, to avoid the obstruction and fundamental confusion that arises when we get stuck in the web of inferential relations that theory-laden terms bring with them. If we do not have a clear, orthodox theory of belief that we can implement without major cuts, then we should refrain to call something a 'belief' because some minimal properties of beliefs (not even exclusive to language

model representations) can be spotted in the activations. However, the research enterprise is not for naught: we can use what we have to do much better than capture 'belief-like' representations.

De novo engineering A better alternative has been labeled as *De Novo Engineering* (DNE) [Chalmers, 2015], the act of coining new concepts for the entities that we observe. DNE is ubiquitous, both in philosophical and scientific practice: we can cite the concepts of *explication* or *supervenience* for the first and the concepts of *feature* or *grokking* for the second, to bring examples which are relevant to the context. Amelioration has the rhetorical advantage of preserving the old concept's semantic punch, and in the context of AI leads to the most fascinating avenues of comparative cognition while staying on an intuitive level for laypeople. DNE, on the other hand, has stronger *theoretical* advantages. First, by constructing brand new concepts, researchers can avoid unsatisfiable accountability for competing theories, rich inferential networks and the misuse of the conceptual role of the borrowed concept. This leads to theoretical freedom, swiftness and simplicity. Second, note that if an entity like a truth-sensitive representation seems interesting from an interpretability perspective, then our conceptual toolkit should be as much a perfect fit as possible for such entity. Reducing concepts goes both ways: we trivialize beliefs when we range over the latent stream, but we trivialize the latent stream when we ascribe it a belief-like quality, too. If we think that truth-sensitivity can be theoretically fruitful, we should find our own terms to declare this property. We would lose a lot of insight, for instance, if tokens were simply called 'words' because they are 'word-like'. The same reflection applies to beliefs and intermediate activations.

The epistemic role of interpretability Which strategy to follow boils down to what we understand the epistemic role of interpretability to be: what kind of niche does interpretability occupy in a larger scientific backdrop? If interpretability should provide instruments and rhetorical devices to adapt our human concepts to language models, then amelioration is a legitimate strategy. If, on the other hand, we are concerned with theoretical clarity and proper, neutral description of phenomena that lie inside transformers, then DNE gives us more freedom and rigor. Even if we want to proceed with amelioration, however, we need crisp and correct concepts to revise and map to new systems: this is not the case for belief, to date. We should wait for the science of belief to catch up with us. In the meantime, if truth-sensitivity in representation is theoretically fruitful, we can work on our own conclusions for that property. If we do a good job, we may merge at the crossroads with cognitive science and philosophy in the future: detaching our notions from their domain is the best shot that we have for meeting those fields with a strong track record of contributions.

Competing Interests

The author declares no competing interests.

References

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL `https://aclanthology.org/2023.findings-emnlp.68/`. TLDR: Evidence that the LLM's internal state can be used to reveal the truthfulness of statements is provided, highlighting its potential to enhance the reliability of LLM-generated content and its practical applicability in real-world scenarios.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL `https://aclanthology.org/2022.cl-1.7/`.

David Bourget and David J. Chalmers. Philosophers on philosophy: The 2020 philpapers survey. *Philosophers' Imprint*, 23(11), 2023. doi: 10.3998/phimp.2109.

Michael Bratman. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press, New York, 1999.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. 2024. doi: https://doi.org/10.48550/arXiv.2212.03827. URL https://arxiv.org/abs/2212.03827.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. Truth is Universal: Robust Detection of Lies in LLMs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 138393–138431. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f9f54762cbb4fe4dbffdd4f792c31221-Paper-Conference.pdf.

David J. Chalmers. What is conceptual engineering and what should it be? *Inquiry*, pages 1–18, 2015. ISSN 0020-174X. doi: 10.1080/0020174X.2020.1817141. URL https://doi.org/10.1080/0020174X.2020.1817141.

David J. Chalmers. Propositional Interpretability in Artificial Intelligence. 2025. doi: https://doi.org/10.48550/arXiv.2501.15740. URL https://arxiv.org/abs/2501.15740.

Andrew Chignell. The Ethics of Belief. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018. URL https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/.

Paul M. Churchland. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy*, 78(2):67–90, 1981. doi: 10.5840/jphil198178268.

Andy Clark and David J. Chalmers. The Extended Mind. *Analysis*, 58(1):7–19, 1998. doi: 10.1093/analys/58.1.7.

Donald Davidson. Radical interpretation. *Dialectica*, 27(1):313–328, 1973. doi: 10.1111/j.1746-8361.1973.tb00623.x.

Daniel C. Dennett. Intentional Systems. *Journal of Philosophy*, 68(February):87–106, 1971. doi: 10.2307/2025382.

Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1981.

Julien Dutant and Clayton Littlejohn. What is rational belief? *Noûs*, 58(2):333–359, June 2024. ISSN 0029-4624. doi: 10.1111/nous.12456. URL https://doi.org/10.1111/nous.12456.

Luciano Floridi and Anna C. Nobre. Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, 34(1):5, April 2024. ISSN 1572-8641. doi: 10.1007/s11023-024-09670-4. URL https://doi.org/10.1007/s11023-024-09670-4.

Jerry Fodor. *The Language of Thought*. Harvard University Press, 1975.

Konstantin Genin and Franz Huber. *Formal Representations of Belief*. Metaphysics Research Lab, Stanford University, fall 2022 edition, 2022. URL https://plato.stanford.edu/archives/fall2022/entries/formal-belief/.

Simon Goldstein and Cameron Domenico Kirk-Giannini. Ai wellbeing. *Asian Journal of Philosophy*, 4(1):1–22, 2025. doi: 10.1007/s44204-025-00246-2.

Simon Goldstein and Benjamin A. Levinstein. Does chatgpt have a mind?, 2024. URL https://arxiv.org/abs/2407.11015.

Daniel A. Herrmann and Benjamin A. Levinstein. Standards for belief representations in llms. *Minds and Machines*, 35(1):5, December 2024. ISSN 1572-8641. doi: 10.1007/s11023-024-09709-6.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks, 2019. URL https://arxiv.org/abs/1909.03368.

Johannes Himmelreich and Sebastian Köhler. Responsible AI Through Conceptual Engineering. *Philosophy & Technology*, 35(3):60, July 2022. ISSN 2210-5441. doi: 10.1007/s13347-022-00542-2. URL https://doi.org/10.1007/s13347-022-00542-2.

Colin Holbrook, Daniel Holman, Joshua Clingo, and Alan R. Wagner. Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. *Scientific Reports*, 14(1):19751, September 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-69771-z. URL https://doi.org/10.1038/s41598-024-69771-z.

Manuel Gustavo Isaac, Steffen Koch, and Ryan Nefdt. Conceptual Engineering: A Road Map to Practice. *Philosophy Compass*, 17(10):1–15, 2022. doi: 10.1111/phc3.12879.

Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, New York, NY, USA, 1965.

Jennifer Lackey. Norms of Assertion and Testimonial Knowledge. In *Learning from Words: Testimony as a Source of Knowledge*, pages 103–140. 2008. ISBN 978-0-19-921916-2. URL https://doi.org/10.1093/acprof:oso/9780199219162.003.0005.

Janet Levin. Functionalism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2023 edition, 2023. URL https://plato.stanford.edu/archives/sum2023/entries/functionalism/.

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

David K. Lewis. Reduction of mind. In Samuel D. Guttenplan, editor, *A Companion to the Philosophy of Mind*, pages 412–431. Blackwell, 1994.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, 2024. URL https://arxiv.org/abs/2306.03341.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. 2024. doi: https://doi.org/10.48550/arXiv.2310.06824. URL https://arxiv.org/abs/2310.06824.

Philip Maxwell Thingbø Mlonyeni. Personal AI, deception, and the problem of emotional bubbles. *AI Soc.*, 40(3):1927–1938, May 2024. ISSN 0951-5666. doi: 10.1007/s00146-024-01958-4. URL https://doi.org/10.1007/s00146-024-01958-4.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

Nicolas Porot and Eric Mandelbaum. The science of belief: A progress report. *WIREs Cognitive Science*, 12(2):e1539, 2021. doi: https://doi.org/10.1002/wcs.1539. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1539.

Eric Schwitzgebel. *A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box*, page 75–99. Palgrave Macmillan UK, London, 2013. ISBN 978-1-137-02652-1. doi: 10.1057/9781137026521_5. URL https://doi.org/10.1057/9781137026521_5.

Eric Schwitzgebel. *Belief*. Metaphysics Research Lab, Stanford University, spring 2024 edition, 2024. URL https://plato.stanford.edu/archives/spr2024/entries/belief/.

John R. Searle. *The Rediscovery of the Mind*. MIT Press, 1992.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Iwan Williams and Tim Bayne. Chatting with bots: Ai, speech-acts, and the edge of assertion. *Inquiry: An Interdisciplinary Journal of Philosophy*, 2024. doi: 10.1080/0020174x.2024.2434874.