Belief Attributions Require Behavioral Evidence

Syed Ali Haider [1]

Abstract

Machine learning routinely attributes beliefs to its systems, identifying them with verbal outputs, log probabilities, probing-classifier directions, or internal circuits. These operationalizations disagree, and the field has no principled way to say which, if any, is the belief. We argue the question is misposed at the level of structural form. Drawing on Ramsey's functionalist account of credence (Ramsey, 1931; de Finetti, 1937) and a parallel move in consciousness research (Teng, 2026; Michel, 2023; Morales & Lau, 2022), we defend a criterion: what makes something a belief is not its form but its functional role in guiding action across perturbation. Two recent results illustrate the criterion's bite. Farquhar et al. (2024) show that contrast-consistent search recovers prominent features rather than knowledge, exactly the failure a structural method would be expected to exhibit. Hase et al. (2023) show that causal-tracing localization is uncorrelated with edit success, yet locate-then-edit methods continue to dominate the field. The second case is not a counterexample but an instance of the methodological pattern we diagnose. Treating functional grounding as constitutive would unify three subfields that are converging on the same philosophical problem without recognizing it as one.

Position. Belief-attribution to machine learning systems is warranted by functional integration with behavior across perturbation, not by the structural form of any internal feature. Methods that identify beliefs by structural form alone are vulnerable to a predictable failure mode: they recover correlates of belief rather than belief itself.

[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA. Correspondence to: Syed Ali Haider <syed.ali.haider.gr@dartmouth.edu>.

Preprint. May 17, 2026.

1. Introduction

Three research programs in contemporary machine learning rest, often without saying so, on a single philosophical question. Honesty evaluation in AI safety supposes a fact about what a model truly believes, a fact that can come apart from what it says (Burns et al., 2023; Pacchiardi et al., 2024). Knowledge probing in LLM evaluation asks "does the model know X?" and operationalizes the question through verbal report, log probabilities, probing classifiers, or activation patching, often producing verdicts that disagree (Hu & Levy, 2023). Mechanistic interpretability identifies circuits and features as representing the belief that P (Meng et al., 2022; Geva et al., 2023). Across all three, a single system admits several structural operationalizations of belief, and the operationalizations come apart in cases that matter.

The standard response has been more careful operationalization. Better probes, more granular interpretability, more behavioral consistency. The unspoken assumption is that the right structural feature, identified with enough care, will simply *be* the belief. We argue that this assumption is wrong, and that the failure of structural identification to do the work it is asked to do is not a methodological accident but the predictable consequence of a category error.

The Ramseyan tradition in formal epistemology gives the diagnosis directly. For Ramsey (1931), a degree of belief is constituted not by its form but by its function in an agent's practical economy, by its capacity to guide action under stakes and to survive Dutch-book exposure. What makes a probability distribution a credence is not its mathematical shape but its integration with the agent's commitments. A representation that lacks this integration may have the form of a credence without being one.

Recent consciousness research has converged on a parallel insight. Teng (2026), building on work by Michel (2023) and Morales & Lau (2022), argues that visibility measures (which assess the structural clarity of a perceived stimulus) and confidence measures (which assess whether second-order judgments reliably track first-order correctness) can dissociate, and that the latter is the more diagnostic for whether a representation is conscious in any task-relevant sense. The methodological lesson generalizes beyond consciousness. When an epistemic kind admits both structural and functional criteria, and when the two can come apart, it

is the functional criterion that does the philosophical work.

Applied to belief-attribution in machine learning, this yields a single criterion across three subfields that currently develop their own piecemeal methods. Belief-claims should be cashed out in terms of functional second-order tracking across perturbation, not in terms of whichever structural operationalization is convenient. The contribution is not a technique but a unification: a principled reason to treat methodological commitments the field has been developing in isolation as instances of one underlying constraint.

Two recent results illustrate what the criterion buys. One predicts a methodological failure; the other diagnoses the field's response to one.

2. The Ramseyan Standard

Three claims from Ramsey and de Finetti matter for what follows. First, credences are constituted by their functional role in an agent's practical economy. For Ramsey (1931), a degree of belief is operationally measurable through betting behavior, because what makes it a degree of belief is its disposition to guide action under stakes. Second, the normative force of probabilism derives from Dutch-book vulnerability. Incoherent credences expose the agent to systems of bets each acceptable in isolation but jointly guaranteeing loss; the argument presupposes an agent whose practical commitments can be exploited. Third, de Finetti's representation theorem shows that exchangeable credences can be represented as if the agent were learning an objective chance from data, but only for agents whose credences satisfy the relevant functional conditions. On this tradition, what makes a representation a credence is not its form but its functional integration with action and updating. A probability distribution that lacks this integration has the form of a credence without being one, and the same point generalizes to belief.

We bracket two questions. First, the agent gap. Machine learning systems do not straightforwardly satisfy the Ramseyan presupposition of a believing subject, and we argue conditionally: if belief-attribution to such systems is meaningful at all, the tradition tells us what would make it so. Second, the choice between Ramseyan and rival accounts of credence (objectivist Bayesianism, primitivism, dispositional accounts). The functional emphasis is shared widely across these traditions, and the argument does not require that the betting interpretation be right in detail.

3. Visibility and Confidence

Consciousness research has developed a methodologically mature version of the structural-versus-functional distinction. Visibility measures, of which the Perceptual Awareness Scale is canonical, ask participants to rate how clearly a stimulus is perceived. Confidence measures ask participants to first perform a perceptual task and then rate confidence in their first-order response (Norman & Price, 2015). The two normally align but can dissociate. Rausch & Zehetleitner (2016) found that visibility ratings track stimulus contrast even when discrimination is at chance, which is to say that visibility can be inflated by features that have nothing to do with the task. The interpretive move, developed by Michel (2023) and Morales & Lau (2022) and applied to aphantasia by Teng (2026), is that confidence measures more reliably target task-relevant conscious representation, while visibility measures can be distorted by phenomenal content that does no epistemic work.

What matters for our argument is the abstract lesson the consciousness literature has worked out in detail. When an epistemic kind admits both structural and functional criteria, the functional criterion bears the philosophical weight. Structural features can come apart from functional ones, and when they do, it is the functional features that track what the analysis of the kind requires. This lesson is not specific to consciousness. It applies wherever a structural feature is being asked to do epistemic work that the philosophical tradition has reserved for a functional property. Belief-attribution in machine learning is such a case.

The transposition is structural rather than literal. We are not claiming that probing-classifier accuracy is metacognitive sensitivity, or that LLM belief is consciousness in any meaningful sense. We are claiming that the methodological shape of the consciousness literature's argument, that an epistemic kind admits both structural and functional criteria, that the two can dissociate, and that the functional one is the philosophically loaded one, generalizes beyond its original domain.

4. Two Worked Examples

If the criterion is doing real work, it should produce verdicts that diverge from current practice in identifiable cases. Two recent results illustrate this in different ways. One is a clean predictive success: a method built on structural identification fails in exactly the way the criterion predicts. The other is more telling: a structural failure has been demonstrated and acknowledged, and yet the field has continued to attribute beliefs to the failing structures, exhibiting the methodological pattern the criterion is designed to expose.

4.1. Contrast-Consistent Search

Burns et al. (2023) introduced contrast-consistent search (CCS), an unsupervised method intended to discover latent knowledge inside language models. The method identifies a direction in activation space whose values for a statement

and its negation approximately sum to one, a consistency property knowledge satisfies, and treats this direction as the model's representation of truth. The reasoning is structural throughout. Knowledge has certain formal properties (negation-consistency); CCS finds directions with those properties; therefore CCS finds knowledge. No behavioral grounding is required at any step; the consistency property is the criterion of identification.

Farquhar et al. (2024) showed the inference fails. They proved theoretically that arbitrary features, not just knowledge, can satisfy the consistency structure CCS searches for. Empirically, they demonstrated that CCS recovers different features depending on prompt and inductive bias. In some settings it recovers a character's stated preferences rather than the model's knowledge. In others, the contrast-pair mapping itself. In others again, features that correlate with knowledge but are dissociable from it. The summary verdict: existing unsupervised methods recover whatever feature of the activations happens to be most prominent.

This is the structural-functional gap displayed empirically. The structural property (negation-consistency) underdetermines the functional one, namely whether the feature actually plays the role of the model's commitment to truth in guiding behavior. Without functional grounding, the structural identification cannot pick out knowledge from its near-correlates. The Ramseyan criterion predicts this directly. Knowledge, like credence, is constituted by its functional role; structural form alone cannot identify it. CCS was a structural method given a functional name, and the failure follows.

4.2. Localization, Editing, and the Field's Response

ROME (Meng et al., 2022) and its successor MEMIT (Meng et al., 2023) edit factual associations in language models by modifying weights in specific MLP layers identified via causal tracing. The interpretability claim is structural: certain weights are picked out as the belief's location, based on their causal role in producing relevant outputs. The vocabulary of "storage" and "location" does real philosophical work, suggesting that the methods have identified where beliefs live.

Hase et al. (2023) tested whether the localization claim does the work it appears to. If causal tracing genuinely identifies where a fact is stored, then editing in the localized layer should outperform editing elsewhere. They found the opposite: the correlation between causal tracing results and edit success is statistically near zero across several editing methods, and a substantial fraction of factual knowledge can be successfully edited outside the layers tracing identifies as its location. Structural localization and functional editability dissociate.

What followed is more revealing than the original finding. Rather than retracting the localization claim or treating the dissociation as a constitutive challenge, the locate-then-edit paradigm has continued to dominate. Subsequent methods ((Li et al., 2024), (Fang et al., 2025),(Li & Chu, 2025)) have extended ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) with improved optimization or regularization, but the structural attribution at their core has remained intact. Recent "precise localization" work frames Hase et al. (2023) as showing that causal tracing offers limited insight into *which* layer to edit, and responds by proposing finer-grained neuron-level localization. The reaction is uniformly "do localization better," not "the localization claim was the wrong question."

The field's response to Hase et al. (2023) is itself evidence of the pattern the functional criterion is designed to diagnose. A subfield that has acknowledged dissociation between its structural attributions and the functional behavior those attributions were supposed to explain, but that continues to license belief-attribution claims as though no dissociation had been demonstrated, is operating with a kind of unprincipled compartmentalization. The functional criterion forces a choice: either the localization vocabulary is making weak claims about useful interventions on weights, in which case the rhetoric needs tightening, or it is making strong claims about belief storage, in which case the dissociation is a problem that has not been answered.

The pragmatist response (that the methods *work* for editing in practice, so the dissociation is a footnote rather than a foundation) is coherent only on the weak reading. On the strong reading, where ROME really does identify where the fact is stored, the dissociation is fatal. The field has tacitly chosen the weak reading in practice while continuing to speak the strong reading aloud, and the result is methodological double-vision the functional criterion is designed to bring into focus.

5. What the Criterion Demands

Two commitments follow if the criterion is taken seriously. The first is that belief-attribution requires perturbation-robust behavioral evidence. To attribute the belief that P to a system is to commit to demonstrating that P guides the system's behavior across paraphrase, context shift, and counterfactual probing. A purely structural identification, whether log probabilities, probing accuracy, or a causal-tracing peak, is necessary but not sufficient. The second is that disagreement between operationalizations is evidence rather than noise. When verbal output, internal probe, and behavioral consistency disagree, the response should be diagnostic (which representation, if any, plays the functional role of belief?) rather than triangulation toward an averaged structural signal. Some existing methods, including causal

scrubbing, behavioral consistency tests, and perturbation-based probing, partially satisfy these commitments. The contribution of the Ramseyan reframing is to show why functional grounding is not optional methodological hygiene but constitutive of the kind being attributed.

6. Discussion

The criterion does not reduce belief to behavior. It identifies the philosophical kind by functional role while remaining neutral on what realizes that role. A system can have internal states that constitute beliefs because of their functional integration, and the position does not deny this; what fails is the inference from structural identification alone to belief-attribution.

This does not rules out interpretability. Interpretability claims become valuable when they identify structures that play the right functional role, namely when intervention shows the structure actually drives the relevant behavior across perturbation. The criterion constrains warranted attribution; it does not exclude one source of evidence.

7. Conclusion

The question of what makes something a belief in machine learning has been treated as a methodological problem to be solved by better operationalization. We have argued it is a philosophical problem that admits no solution at the level of structural form. The Ramseyan tradition gives a principled answer: functional integration with action across perturbation. Farquhar et al. (2024) illustrate the criterion's predictive force; Hase et al. (2023) and the field's response illustrate the methodological pattern the criterion makes visible. Treating functional grounding as constitutive rather than as a hygienic afterthought would do real work where ad hoc methods currently struggle, and would unify three subfields that are converging on the same problem without seeing it as one.

References

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.03827.

de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1): 1–68, 1937.

Fang, J., Jiang, H., Wang, K., Ma, Y., Jie, S., Wang, X., He, X., and seng Chua, T. Alphaedit: Null-space constrained knowledge editing for language models, 2025. URL https://arxiv.org/abs/2410.02355.

Farquhar, S., Varma, V., Kenton, Z., Gasteiger, J., Mikulik, V., and Shah, R. Challenges with unsupervised LLM knowledge discovery. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2312.10029.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Hase, P., Bansal, M., Kim, B., and Ghandeharioun, A. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. Spotlight; arXiv:2301.04213.

Hu, J. and Levy, R. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5040–5060, 2023. doi: 10.18653/v1/2023.emnlp-main.306.

Li, Q. and Chu, X. AdaEdit: Advancing continuous knowledge editing for large language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4127–4149, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.208. URL https://aclanthology.org/2025.acl-long.208/.

Li, X., Li, S., Song, S., Yang, J., Ma, J., and Yu, J. Pmet: Precise model editing in a transformer, 2024. URL https://arxiv.org/abs/2308.08742.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2202.05262.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.07229.

Michel, M. Confidence in consciousness research. *WIREs Cognitive Science*, 14(2):e1628, 2023. doi: 10.1002/wcs.1628.

Morales, J. and Lau, H. Confidence tracks consciousness. In Weisberg, J. (ed.), *Qualitative Consciousness: Themes from the Philosophy of David Rosenthal*,

pp. 91–105. Cambridge University Press, 2022. doi: 10.1017/9781108768085.007.

Norman, E. and Price, M. C. Measuring consciousness with confidence ratings. In Overgaard, M. (ed.), *Behavioral Methods in Consciousness Research*. Oxford University Press, 2015. doi: 10.1093/acprof:oso/9780199688890.003.0010.

Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., Evans, O., and Brauner, J. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.15840.

Ramsey, F. P. Truth and probability. *The Foundations of Mathematics and other Logical Essays*, 1931. Written 1926, posthumously published 1931, edited by R.B. Braithwaite, Routledge & Kegan Paul, London.

Rausch, M. and Zehetleitner, M. Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, 7:591, 2016. doi: 10.3389/fpsyg.2016.00591.

Teng, L. Metacognition in aphantasia: Taking the "conscious" view seriously. *Neuropsychologia*, 221:109331, 2026. doi: 10.1016/j.neuropsychologia.2025.109331.