Can we propositionally interpret AI Agents?

Abstract

In recent work on AI interpretability, David Chalmers introduces the notion of *propositional interpretability*: the condition under which an artificial system can be said to have beliefs, knowledge, or reasoning states with propositional contents that are interpretable. The proposal is especially attractive in the context of increasingly agentic AI systems (Agentic AI), for which questions about what the system *takes to be the case* or *aims to bring about* appear central to explanation and safety. For instance, a system's being trustworthy seems essentially fixed by the correct application of propositional attitude ascriptions—that the Agentic AI has beliefs and desires of a trustworthy system. Propositional interpretability promises a way of making sense of such systems at the level of reasons, not merely mechanisms—all while side-stepping complex questions of AI consciousness.

This presentation notes and explores one old challenge to such interpretation—namely, its indeterminacy. If propositional interpretability is situated within the interpretivist tradition (drawing on radical interpretation and rationality constraints) then propositional content is not fixed without appeal to strict constraints (Davidson 1974). On these views, propositional attitudes are fixed only by identifying the interpretation that best rationalises an agent's behaviour and internal organisation. The familiar problem, originating in Quine and developed by Davidson, is that even these interpretive constraints can often fail to determine a unique assignment of propositional content to an agent(Quine 1960). Indeed, multiple incompatible sets of propositional attitudes may equally satisfy all interpretive constraints placed on an agent. Accordingly, the ascription of propositional attitudes to Agentic AI systems may be radically underdetermined by the evidence of interpretation —even with *a priori* constraints such as appeals to ideal rationality.

This poses a problem for the ideal of *thought logging* that Chalmers presents as a guiding aim of propositional interpretability: a systematic mapping from an agentic system's internal processes to a stream of propositional attitudes (Chalmers 2025). If propositional content is radically underdetermined, then there may be no unique sequence of beliefs, desires, or intentions to recover. What thought logging produces will depend on prior interpretive choices rather than on facts fixed by the system itself. In that case, propositional interpretability risks reduction to an interpreter-relative heuristic, rather than a method for uncovering determinate internal attitudes.

The conclusion is not that propositional interpretability has no role in understanding agentic AI. Rather, the claim is that, we must either reject the interpretivist paradigm or motivate strict constraints of ideal rationality or behaviour to fix unique propositional attitude ascriptions to AI systems.

Bibliography

Chalmers, David J. 2025. "Propositional Interpretability in Artificial Intelligence." *arXiv preprint* arXiv:2501.15740.

Davidson, Donald. 1974. “Belief and the Basis of Meaning.” *Synthese* 27 (3–4): 309–23. https://doi.org/10.1007/BF00485052.

Lewis, David. 1974. “Radical Interpretation.” *Synthese* 27 (3–4): 331–44. https://doi.org/10.1007/BF00485053.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.