

Chain-of-Thought: Epistemic Flaws and Fictional Explanations

FABIO MORREALE, JOAN SERRÀ, and YUKI MITSUFUJI, Sony AI

Chain-of-thought (CoT) aims to verbalise how AI systems arrive at their outputs. However, previous work proved that CoTs are unfaithful and necessarily non-causal, as they are not causally anchored to the mechanisms that produced the action and thus cannot claim with any certainty that what they display is grounded in real reasoning. In this paper, we first surface a number of epistemic fallacies underlying CoT. These are based on the assumptions that there must exist a mapping between human and machinic abstractions, and that such a mapping can be intercepted and communicated via language and automation. We then show that CoT replaces causality with discursive coherence, and that palatable fictional narrative sustains the illusion of intelligibility.

CCS Concepts: • **Computing methodologies** → **Intelligent agents**; • **Human-centered computing** → **HCI theory, concepts and models**.

Additional Key Words and Phrases: CoT, Agentic AI, Alignment, Agential Realism, XAI

1 Introduction

The incipit for this paper is one of the questions proposed by the workshop organisers about chain-of-thought (CoT):

Does chain-of-thought contain useful insights that help users interact more effectively with the agent?

Our response, which we will elaborate on in the paper, is sceptical.

CoT denotes a family of practices in which a model emits intermediate textual reasoning. CoT promises to show *how* the system arrived at an answer as seen in many LLM chatbots (Fig. 1). CoT techniques are thus intended to be utilised by a human observer to find explanatory evidence of the inner *reasoning* of a machine learning model. In general, these techniques aim to intercept some traces of such reasoning, transfigure them into human-legible form (language), and eventually display them to the user.

Previous work has shown that CoT provides a false sense of intelligibility, as the human-legible reasoning does not necessarily correspond to the model’s internal logic. Turpin et al. [10] demonstrate that CoT is affected by forms of unfaithfulness: models produce seemingly narratives that are consistent with the final answer but do not actually track the mechanisms that produced actions; they are unfaithful to the causal drivers of model behaviour. Barez et al. [3] sharpen the epistemic consequence for XAI. When a rationale is not causally anchored, coherence becomes a weak proxy for explanation, and CoT should be treated as an artefact of linguistic plausibility rather than a window into “how the model works”. CoT thus functions as a post-hoc justification for the output rather than a description of it; it is a plausible-sounding, step-by-step explanation that matches the answer but did not produce it. Manuvinakurike et al. [6] also reported that CoT does not reliably improve outcomes and may fail to deliver dependable explanatory support in multi-step workflows. In this short paper, we join this line of critique by identifying three possible epistemic fallacies embedded in CoT.

2 Three Epistemic Flaws with CoT

First, CoT - as all the broader XAI field - is committed to bridging human and machinic abstractions. However, the attempt to map these two forms of abstraction might be grounded on epistemically-tenuous grounds [1, 4, 5]. The reasons, as we have previously discussed [7, 8], are related to the fact that human cannot have epistemic access to

Are Chain-of-thought proper scientific objects or palliatives?



⊗ Thought for 7 seconds ▾

- Hmm, this is a nuanced philosophical and scientific question about the nature of Chain-of-Thought reasoning. The user is asking whether CoT is a legitimate object of scientific study or merely a temporary fix for AI limitations.
- The question requires addressing both the technical and epistemological dimensions. I should start by defining CoT clearly, then systematically analyze both perspectives. The "scientific object" view can be supported by its measurability, psychological parallels, and mechanistic interpretability. The "palliative" view should acknowledge its practical utility while examining its limitations as a proxy for true reasoning.
- A dialectical approach would work best—presenting both sides before synthesizing them. The conclusion should recognize CoT as a transitional epistemic tool that's currently scientifically valuable but may evolve. The user seems interested in deeper implications, so I'll avoid oversimplification and emphasize the philosophical tension.
- The response should maintain academic rigor but remain accessible, using concrete examples like LLM reasoning tasks and cognitive science comparisons. I'll structure it to first validate both perspectives, then show their synthesis, leaving room for the user's own reflection.

Fig. 1. Chain-of-thought in Deepseek.

features that relevant for machines, and their abstractions cannot be measured against each other. XAI is thus a process of imposing human reasoning on non-human processes, and the epistemic validity of this mapping process is debatable.

Second, CoT takes this questionable idea a step further in an attempt to automate such mapping. In CoT, humans are relegated to being spectators of explanations of machinic reasoning, which thus do not require any humans in the epistemic explanatory process. Contrasting this view, and inspired by Baradian's agential realism metaphysics [2] metaphysics, we recently proposed that only partial and situated explanations can ever be generated, and that they depend on an entanglement of models, observers, tools, and contexts [8]. We thus challenged the assumptions implicit in mainstream XAI research that some hidden *explananda* exist within AI models that can be recovered by an observer who is external to the interpretative process. A similar conclusion has been reached by Davide Picca from a semiotic perspective: model outputs need active interpretation from human users [9]. Picca argued that AI models "manipulate linguistic forms without genuine understanding, and their outputs require active interpretive cooperation from users". He discussed that, while AI models can produce the sign, their inability to have direct access to the reference means they cannot be an independent interpreter. Explanations are thus not inherent within the model but emerge through sustained engagement and therefore depend on the human interpreter (among other objects of configuration). Thus, any *automatic interpretation*, which, by definition, excludes humans from the loop, might be a faulty epistemic object.

Third, CoT relies on *language* as a transparent vehicle for making the model intelligible. Weatherby recently argued that the recent success of generative AI forces us to rethink the relationship between language and referentiality [11]. Language, in his account, does not cleanly *represent* an already-formed inner cognition. Rather, language is "is

complex, cultural, and even poetic first, and referential, functional, and cognitive only later”, reversing the common assumption that semantic content pre-exists expression and is merely encoded into words. From this perspective, rendering *reasoning* in natural language is not a neutral translation step. Linguistic form does not simply report a latent process but reorganises what can count as a reason by imposing human semantic norms (coherence, intentionality, justificatory order) onto *traces* that were never produced under those norms. CoT, therefore, is epistemically unstable: the trace reads as if it carried a stable reference to internal mechanisms, while in fact it might not.

3 CoT is Fictional and Implications for HCXAI

CoT succeeds in resembling human reasoning and thereby invites users to believe that it reveals the model’s inner logic. However, as we have seen, this resemblance is not evident. CoT outputs are *fictional* narratives whose coherence is decoupled from any objective referent. As such, CoT sustains an illusion of understanding but does not provide any access to model behaviour. Not only can CoT make a system feel legible because the trace offers a familiar narrative form, but even more profoundly, it encourages users to treat narrative plausible coherence as evidence that inner machine working can be 1) unequivocally understood without human input and 2) correctly reported in human language.

The critique of CoT has direct consequences for *alignability*. CoT traces can make a system appear value-consistent because the narrative resembles the kind of reasoning a value-aligned agent would be expected to follow. However, as we mentioned, this appearance rests on plausibility rather than on evidential access to the conditions that produced the action. Alignment risks being assessed at the level of narrative coherence rather than at the level of accountability. The issue is that a coherent rationale can signal that an action is justifiable within a familiar normative framework, but it is not evidence that the action was generated in accordance with that framework.

Going back to the initial question posed by the workshop organisers, we are sceptical that CoT can offer any useful insights in either an agentic or a non-agentic setting. We instead recommend further study to identify and challenge the assumptions embedded in CoT practice. Specifically, we invite empirical and theoretical scrutiny of the proposition that CoT, at least in its current implementations, functions primarily as a not just plausible but palatable narrative. Such narratives operate by producing a psychological sense of trust and understanding in machinic processes and decisions, even when no corresponding explanatory grounding is available.

References

- [1] Ramón Alvarado. 2023. AI as an Epistemic Technology. *Science and Engineering Ethics* 29, 5 (Oct. 2023), 32. doi:10.1007/s11948-023-00451-3
- [2] Karen Barad. 2007. Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. (2007), 542.
- [3] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v1.
- [4] Jenna Burrell. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (June 2016), 2053951715622512. doi:10.1177/2053951715622512
- [5] M. Beatrice Fazi. 2021. Beyond Human: Deep Learning, Explainability and Representation. *Theory, Culture & Society* 38, 7-8 (Dec. 2021), 55–77. doi:10.1177/0263276420966386
- [6] Ramesh Manuvinakurike, Emanuel Moss, Elizabeth Anne Watkins, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2025. Thoughts without Thinking: Reconsidering the Explanatory Value of Chain-of-Thought Reasoning in LLMs through Agentic Pipelines. doi:10.48550/arXiv.2505.00875 arXiv:2505.00875 [cs].
- [7] Fabio Morreale, Marco A. Martinez-Ramirez, Raul Masu, WeiHsiang Liao, and Yuki Mitsufuji. 2025. Reductive, Exclusionary, Normalising: The Limits of Generative AI Music. *Transactions of the International Society for Music Information Retrieval* (Sep 2025). doi:10.5334/tismir.256
- [8] Fabio Morreale, Joan Serrà, and Yuki Mitsufuji. 2026. Emergent, not Immanent: A Baradian Reading of Explainable AI. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (2026).
- [9] Davide Picca. 2025. The Semiotic Channel Principle: Measuring the Capacity for Meaning in LLM Communication. *arXiv preprint arXiv:2511.19550* (2025).

- [10] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *NeurIPS 2023*.
- [11] Leif Weatherby. 2025. *Language machines: cultural AI and the end of remainder Humanism*. Vol. 74. U of Minnesota Press.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009