


## Letter

## LLMs don't know anything: reply to Yildirim and Paul

Mariel K. Goddu<sup>1,2,3</sup>,  
Alva Noë<sup>4,5</sup>, and  
Evan Thompson <sup>6,\*</sup>

In their recent Opinion in *TICS* [1], Yildirim and Paul propose that large language models (LLMs) have 'instrumental knowledge' and possibly the kind of 'worldly' knowledge that humans do. They suggest that the production of appropriate outputs by LLMs is evidence that LLMs infer 'task structure' that may reflect 'causal abstractions of... entities and processes in the real world' [1]. While we agree that LLMs are impressive and potentially interesting for cognitive science, we resist this project on two grounds. First, it casts LLMs as agents rather than as models. Second, it suggests that causal understanding could be acquired from the capacity for mere prediction.

The map does not know the way home, and the abacus is not clever at arithmetic. It takes knowledge to devise and use such models, but the models themselves have no knowledge. Not because they are ignorant, but because they are *models*: that is to say, tools [2,3]. They do not navigate or calculate, and neither do they have destinations to reach or debts to pay. *Humans* use them for these epistemic purposes. LLMs have more in common with the map or abacus than with the people who design and use them as instruments. It is the tool creator and user, not the tool, who has knowledge.

Despite these considerations, Yildirim and Paul [1] entertain the hypothesis that LLMs could build 'worldly knowledge' out of what they call 'instrumental knowledge', analyzed as success in the next token generation, in particular 'task

domains'. This is a misleading characterization of 'instrumental knowledge'. The term is typically used to refer precisely *not* to statistical or associative learning, but rather, by contrast, to knowledge achieved by humans and other animals on the basis of appreciating contingencies between their actions and outcomes (e.g., operant conditioning) [4]. Whereas instrumental knowledge, properly understood, is 'difference-making' knowledge that supports causal intervention [5–9], the capacity to guess what comes next, no matter how robust, is no such ability. No amount of prediction, from a set of observations no matter how large, can support the grasp of the notion of 'making', 'generating', or 'doing' that is the basis of our world knowledge.

Imagine seeing a sequence of lights on the wall: green, yellow, red; green, yellow, red. Given these observations, you may successfully predict the next color (green). The capacity to make this prediction, a form of statistical inference, is distinct from the capacity to understand what *generates* the sequence, or what would change it (e.g., knowing it is made by the traffic light outside, without which it would not occur). Evidence from developmental psychology suggests that the human capacity to appreciate this type of dependency (i.e., to have a generative concept of 'cause') develops from our experience of our own and others' goal-directed actions [5]. Human learning occurs via active participation in world affairs, by making the differences that we need and want to make in our lives as social, biological organisms.

Indeed, Yildirim and Paul acknowledge that the idea that LLMs could acquire causal abstractions of the world requires 'a leap of faith' [1]. We suggest that three leaps of faith are required, all of which plunge us into the absurd.

The first leap is the idea that a model of something could become the thing that is

being modeled. Models are tools, not agents, and they are *our* tools, constructed to serve our interests and values. LLMs do not perform any of their own tasks; they perform *our* tasks. The answers that LLMs return to our prompts are 'approximately truth-preserving and relevant' [1] only by our lights [10]. ChatGPT really appears to 'write' that first draft, although this should be no surprise, since this is what we designed it for: ChatGPT is a pretend subject engineered by real subjects to seem like a real subject. It takes nothing away from the potential utility (or risks [11]) associated with such powerful technologies of pseudo-agency to insist on what we already know: even the best models do not become what they are so effectively used to model [2,3].

The second leap is the idea that computational processes enabling statistical pattern detection and token generation can be an 'instrument' for acquiring the causal knowledge involved in understanding 'task structure' [1]. However, linking empty tokens based on probabilities (even in ways that we are in a position to know does reflect the truth of a given domain, be it a summarization task, physics, or arithmetic) does not warrant attributing knowledge of that domain to the token generator itself.

We said above that LLMs do not perform any tasks of their own, they perform our tasks. It would be better to say that they do not really *do* anything at all. Hence the third leap: treating LLMs as agents. However, since LLMs are not agents [12], let alone epistemic ones, they are in no position to do or know anything.

<sup>1</sup>Department of Philosophy, Stanford University, 450 Stanford Way, Main Quad, Building 90, Stanford, CA 94305, USA

<sup>2</sup>Institut für Philosophie, Freie Universität Berlin, Habelschwerdter Allee 30, 14195, Berlin, Germany

<sup>3</sup>Centre for Advanced Study in the Humanities: 'Human Abilities', Schönhauser Allee 10-11, 10119, Berlin, Germany

<sup>4</sup>Department of Philosophy, University of California Berkeley, 314 Philosophy Hall #2390, Berkeley, CA 94720, USA

<sup>5</sup>Einstein Research Group: 'Reorganizing Ourselves', Graduierten Kolleg: Normativity, Critique, Change, Freie Universität Berlin, Altensteinstraße 15, 14195, Berlin, Germany

<sup>6</sup>Department of Philosophy, University of British Columbia, 1866 Main Mall, Vancouver, BC V6T 1Z4, Canada

\*Correspondence:

evan.thompson@ubc.ca (E. Thompson).

<https://doi.org/10.1016/j.tics.2024.06.008>

© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

## References

1. Yildirim, I. and Paul, L.A. (2024) From task structures to world models: what do LLMs know? *Trends Cogn. Sci.* 28, 404–415
2. Noë, A. (2023) *The Entanglement: How Art and Philosophy Make Us Who We Are*, Princeton University Press
3. Frank, A. et al. (2024) *The Blind Spot: Why Science Cannot Ignore Human Experience*, MIT Press
4. Dickinson, A. (1994) Instrumental conditioning. In *Animal Learning and Cognition* (Mackintosh, N.J., ed.), pp. 45–79, Academic Press
5. Goddu, M. and Gopnik, A. (2024) The development of human causal learning and reasoning. *Nat. Rev. Psychol.* 3, 319–339
6. Pearl, J. and Mackenzie, D. (2018) *The Book of Why: The New Science of Cause and Effect*, Basic Books
7. Pearl, J. (2009) *Causality*, Cambridge University Press
8. Woodward, J. (2007) Interventionist theories of causation in psychological perspective. In *Causal Learning: Psychology, Philosophy, and Computation* (Gopnik, A. and Schulz, L., eds), pp. 19–36, Oxford University Press
9. Woodward, J. (2021) *Causation with a Human Face: Normative Theory and Descriptive Psychology*, Oxford University Press
10. Smith, B.C. (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment*, MIT Press
11. Birhane, A. (2021) The impossibility of automating ambiguity. *Artif. Life* 27, 44–61
12. Roli, A. et al. (2022) How organisms come to know the world: fundamental limits on artificial general intelligence. *Front. Ecol. Evol.* 9, 806283