Are LLMs "just" next-token predictors?

Zak Hussain[1,2*], Rui Mata[1] and Dirk U. Wulff[2,1]

[1*]University of Basel.
[2]Max Planck Institute for Human Development.

*Corresponding author(s). E-mail(s): z.hussain@unibas.ch;
Contributing authors: rui.mata@unibas.ch; wulff@mpib-berlin.mpg.de;

Abstract

Large language models (LLMs) are often labeled as "just next-token predictors" devoid of cognitive capacities such as "thought" or "understanding". We argue that these deflationary claims are premature. Drawing on prominent theoretical and philosophical frameworks in cognitive science, we critically evaluate different forms of "Justaism" that dismiss LLM cognition by labeling LLMs as "just" simplistic entities without specifying or substantiating the critical capacities they supposedly lack. Our analysis highlights the need for a more nuanced discussion of LLM cognition, aiming to better inform future research and the development of machine intelligence.

Keywords: large language models, cognition, artificial intelligence

Over 70 years ago, Alan Turing posed a question that has since captivated computer scientists, cognitive scientists, and philosophers alike: "Can machines think?" (Turing, 1950). With the recent proliferation of increasingly powerful artificial intelligence systems—in particular, large language models (LLMs)—variants of this question have made their way far beyond the confines of academic departments.

Some have asserted that LLMs cannot be said to "think" because they are "just...": "next-token predictors", "function approximators", or "stochastic parrots", and thus lack some essential capacity necessary for cognition. Unfortunately, in their most facile form, such deflationary claims fail to state what exactly this capacity is. These more facile claims have consequently been given the pejorative label "Justaism" (pronounced "just-a-ism") due to the confident self-evidence with which they are wielded (Aaronson, 2023). Thus, Justaism refers not to deflationary positions against LLM cognition *per se*, but specifically to *unsubstantiated* deflationary positions.

Crucially, opinions on the reality of LLM "thought", "reasoning", or "understanding" (henceforth, "cognition") have implications both for people's willingness to trust such systems (Mitchell & Krakauer, 2023) and, ultimately, for their socio-economic impact.

In what follows, we present two flavors of Justaism, and provide a critical analysis of these positions based on prominent views in cognitive science. We refer to the flavors' prototypical forms but also provide specific examples found in the literature and public discussion on LLM cognition in a companion webpage (github.com/Zak-Hussain/againstJustaism).

1 Flavors of Justaism

1.1 Anti-simple-objectives Justaism

"It's just a next-token predictor."

Perhaps the most common form of Justaism, which we dub *anti-simple-objectives Justaism*, takes issue with how LLMs are pre-trained. The assertion is that because the LLM pre-training objective is simply to predict the masked or next token, LLMs cannot be doing something as complex as cognition.

Assuming proponents of this view believe that humans can be said to possess cognition, anti-simple-objectives Justaism can be refuted by making the following facetious analogy to humans and other creatures shaped by evolution: We humans are *just* "next-child producers", stumbling forward in pursuit of the all-encompassing base objective of inclusive fitness maximization—such a simple objective cannot possibly lead to cognition. To most, this analogy's conclusion is absurd, thus questioning the same reasoning when applied to LLM cognition.

Of course, there are important differences between next-token prediction and next-child production. For instance, the ancestral environment from which we evolved was potentially richer than the online text corpora used to train LLMs. Combined with a sufficiently complex nervous system and other distinguishing factors (e.g., resource competition), biological evolution may lead to the development of *instrumental objectives* that are more conducive to cognition than next-token prediction.

However, even if it were the case that these distinguishing factors are pivotal in the development of instrumental objectives *in humans*, it is nevertheless plausible that cognition-enabling representations and even instrumental objectives could be acquired via simple LLM pre-training objectives. Indeed, research suggests that models trained with simple base objectives (e.g., next-token prediction) could develop complex instrumental strategies similar to those of humans in order to achieve high performance on the base objective (through a process of "mesa-optimization" Hubinger, van Merwijk, Mikulik, Skalse, & Garrabrant, 2019). Furthermore, these instrumental objectives may not need to be especially complex so as to be on par with those of human beings. After all, many foundational theories in cognitive science posit relatively simple (instrumental) objectives as fundamental components, with prominent examples including predictive brain theories (e.g., "Bayesian brain", "predictive coding", "active

inference" Clark, 2013). If simple objectives are overall thought insufficient for the development of cognition, researchers need to clarify why humans are not in a similar position to LLMs in this regard.

Ultimately, we would argue that it is by no means self-evident that an LLM seeking to predict the next token could not acquire representations and even instrumental objectives akin to humans. The onus is on the objector to substantiate why this is (likely) untrue.

1.2 Anti-Anthropomorphism

"It's just a machine."

A second prominent form of Justaism, which we dub *Anti-anthropomorphic Justaism*, claims that attributing cognition to machines constitutes a fundamental error. In its strongest form, it argues that such thinking commits a category error because cognition is *by definition* a human capacity. On this view, the essential capacity that LLMs lack and humans possess is just that: humanness.

Although logically valid, we would argue that this view is unproductively restrictive. Advances in scientific theory are often based on the realization that a concept is more general than previously believed. One instructive example comes from animal cognition research, where, in response to a growing body of empirical evidence, researchers began to see great utility in ascribing capacities previously thought to be uniquely human, including emotion, self-awareness, or consciousness, to non-human animals (De Waal, 2016). We believe it should be *in principle* acceptable to consider such conceptual generalizations for information processing systems more broadly.

There are, of course, more moderate forms of anti-anthropomorphic Justaism. For instance, one might take the view that although it is not a problem *in principle* to talk about LLM cognition, the burden of evidence for doing so should be set very high. One reason for this would be to guard against the Eliza effect (Mitchell & Krakauer, 2023), which refers to the human propensity to all-too-liberally ascribe "thought" to even the simplest of machines.

Although we agree that it is important to reject naive anthropomorphism, we note that running counter to anthropomorphism is another, perhaps more infamous, human tendency: anthropocentrism. Regarding cognition, anthropocentrism is the tendency to view capacities such as "thought" as so unique that it would not make sense to ascribe them to "lesser" systems, such as non-human animals (see, e.g., De Waal, 2016). In the context of artificial intelligence, it can be observed in the well-documented phenomenon of algorithmic aversion—the human tendency to rely more on human advisors over equally good or better-performing algorithms (Jussupow, Benbasat, & Heinzl, 2022). Anthropocentrism may ultimately have implications for the adoption of novel technologies that have the potential to contribute to human wealth and well-being.

In light of humans' countervailing tendency to view their own cognition as exceptional, we would advocate for specifying more precisely the forms of cognition in question and the evaluative criteria to be employed. We believe this will enable more

substantive discussions of and comparisons between the capabilities of humans and other information-processing systems.

1.3 Other criticisms of LLM cognition

Beyond *anti-simple-objectives* and *anti-anthropomorphic* forms of Justaism are more moderate forms of LLM deflationism. Perhaps most prominent in the literature are arguments that emphasize the importance of the distinction between *meaning* (semantics) and *form* (syntax) (e.g., Bender & Koller, 2020; Searle, 1980). One argument is that LLMs do not have access to *meaning* because of their lack of real-world grounding and, consequently, human-like understanding. These positions are considerably more substantial than their Justaic cousins mentioned above and thus do not deserve to be labeled as "Justaic". We are also more sympathetic to these views and believe that LLMs that are more physically and socially grounded are also likely to be cognitively more advanced.

Nevertheless, we would caution against confidently concluding that LLMs lack cognition on these bases. Not only do the arguments draw on notoriously hard-to-define concepts (e.g., *meaning* or *grounding*), but they also rely heavily on thought experiments of a specific nature. These thought experiments make an appeal to our intuitions by showing that another system in an analogous but more intuitively transparent position to an LLM—such as the human in Searle (1980)'s famous "Chinese Room"—appears cognitively lacking (e.g., Bender & Koller, 2020; Searle, 1980). However, the intuitions evoked by such thought experiments can be misleading, especially when applied to complex systems that have been trained on more data than a human could process in a lifetime. As a testament to the limits of intuition in this context, consider the to-many-astonishing effectiveness of "simply" scaling-up model and training set sizes for improving LLM performance (Kaplan et al., 2020). Although we share the intuitions evoked by such thought experiments, we thus resist putting too much weight on them.

2 Toward a more measured discussion

In support of a more measured discussion of LLM cognition, we would like to advance three guiding principles: (i) modesty regarding human cognition (and our understanding of it), (ii) consistency for future work comparing humans and LLMs, and (iii) a focus on empirical benchmarks.

Regarding modesty, we would reiterate that human history is littered with delusions of human exceptionalism (De Waal, 2016). This is despite our limited understanding of the mechanisms underlying cognition. Thus, although we fully support cautioning against the dangers of (naive) anthropomorphism, we see the need for a backstop against the opposite tendency: viewing human cognition as too special to also be ascribed to LLMs.

Regarding consistency, we would reiterate the need for consistent goalposts: Are we applying the same standards to LLMs as we would to humans? For instance, if we wish to reduce LLM cognition to its pre-training objective (i.e., next-token prediction), we must show why the same reductionism should not apply to humans as well.

Similarly, when LLMs commit errors that appear so elementary to us as to discredit LLM cognition, it is important to recall the host of fallacies and illusions that humans are susceptible to and consequently may not so easily identify or view as significant. These considerations not only help guard against certain biases (e.g., algorithmic aversion), but they can also provide a new perspective on human cognition by helping identify aspects of cognition that are, in fact, uniquely human.

Finally, we are sympathetic to Turing (1950)'s view (among others, e.g., Niv, 2021; Zhang, Bengio, Hardt, Recht, & Vinyals, 2021) that discussions of cognition should focus on observables. As Trott, Jones, Chang, Michaelov, and Bergen (2023) note, axiomatic rejections of LLM cognition can lead to positions that have no empirically testable implications. Not only does this run contrary to good scientific practice, but it can also lead to investigations of LLM cognition that lack practical relevance. After all, it is predominantly the behavior of a system that impacts the world. Consequently, we believe in the need for clear and consistent empirical benchmarks that allow for direct evaluations of the cognitive capacities of humans and LLMs.

Ultimately, the jury is still out on the existence and extent of LLM cognition. Empirical research has demonstrated interesting cognitive deficits in LLMs (e.g., Berglund et al., 2023), but also impressive feats (e.g,. Bubeck et al., 2023). Given the limitations of Justaic reasoning, we believe the ball is in the skeptic's court to show why these feats do *not* constitute evidence of cognition.

Acknowledgements. We thank Ralph Hertwig, Anne-Marie Nussberger, Lucius Caviola, and Thomas T. Hills for their helpful feedback and Laura Wiles for editing the manuscript. We recognize funding to Dirk U. Wulff (197315) and Rui Mata (204700) from the Swiss National Science Foundation.

References

Aaronson, S. (2023). *The problem of human specialness in the age of ai.* https://scottaaronson.blog/?p=7784. (Accessed: 2024-03-31)

Bender, E.M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198).

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., Evans, O. (2023). The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arxiv preprint arxiv:2309.12288.*

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arxiv preprint arxiv:2303.12712.*

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204, https://doi.org/https://doi.org/10.1017/S0140525X12000477

De Waal, F. (2016). *Are we smart enough to know how smart animals are?* WW Norton & Company.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arxiv preprint arxiv:1906.01820.*

Jussupow, E., Benbasat, I., Heinzl, A. (2022). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. *Proceedings of the 28th european conference on information systems.*

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., ... Amodei, D. (2020). *Scaling laws for neural language models.*

Mitchell, M., & Krakauer, D.C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120, https://doi.org/https://doi.org/10.1073/pnas.221590712

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*(5), 601-609, https://doi.org/https://doi.org/10.1037/bne0000471

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424, https://doi.org/https://doi.org/10.1017/S0140525X00005756

Trott, S., Jones, C., Chang, T., Michaelov, J., Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309, https://doi.org/https://doi.org/10.1111/cogs.13309

Turing, A.M. (1950, oct). Computing machinery and intelligence. *Mind*, *59*, 433–460, https://doi.org/10.1093/mind/LIX.236.433

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115, https://doi.org/https://doi.org/10.1145/3446776