Against Justaism: A call for more measured discussions on LLM cognition

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

Are large language models (LLMs) "*just* next-token predictors" devoid of cognitive capacities such as "thought" or "understanding"? We argue from a cognitive science perspective that it is far from self-evident that such perspectives hold despite their prominence in current discussions surrounding LLMs. We make our case by critically evaluating different flavors of *Justaic* reasoning and end with a call for more measured discussions on LLM cognition.

Keywords: large language models, cognition

Over 70 years ago, Turing (1950) posed a question that has captivated computer scientists, cognitive scientists, and philosophers alike: "Can machines think?". With the recent proliferation of increasingly powerful AI systems—in particular, large language models (LLMs)—variants of this question have made their way far beyond the confines of academic departments. Turing warned that this is a poorly formulated question prone to verbal dispute and that we should instead focus on behavioral operationalizations of the problem. Whilst we are sympathetic to this view (that is, to focusing on a system's behavior), 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 take their potential socio-economic impact seriously.

Despite coming in different flavors, many assertions against LLM cognition share two common claims. The first, usually more explicit, claim is that LLMs lack some crucial capacity necessary for cognition. For instance, one might assert that LLMs

cannot "think" because they are "just...": "next-token predictors", "function approximators", or "stochastic parrots", and thus lack the essential capacity. Unfortunately, in their most facile form, deflationary claims along these lines fail to state what exactly this capacity is and have thus been given the pejorative label "Justaism" (pronounced "just-a-ism") due to the confident self-evidence with which they are sometimes wielded (Aaronson, 2023).

The second, usually more implicit, claim is that whilst LLMs may lack this cognition-enabling capacity, humans likely possess it. It is, after all, the human possession of this capacity that gives meaning to folk psychological concepts such as "thought", "reasoning", or "understanding". It is also what makes claims such as "LLMs are *just* next-token predictors" interesting: if humans were also "just" next-something predictors (as we argue below, this view is not uncommon in cognitive science), saying the same about LLMs would lose its LLM-diminishing bite. Ultimately, claims about LLM cognition are only deflationary when compared to some baseline, and, in the case of LLMs (and AI systems more generally), this baseline is often us.

Since strong Justaic claims provide little justification for their deflationary implications, it can be challenging to criticize them beyond objecting that they beg the question. In what follows, we thus attempt to sketch two different prototypical flavors of Justaism, before responding to each in turn. For examples of Justaism in the literature and public discussion, please refer to github.com/Zak-Hussain/againstJustaism.

1 Flavours of Justaism

1.1 Anti-simple-objectives

"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 appears to be that since the LLM pre-training objective is simply to predict the missing (masked language modeling) or next (causal language modeling) token, LLMs cannot be doing something as complex as cognition.

Assuming proponents of this view still believe that humans do cognition, the claim that "it's just a next-token predictor" can be addressed by making the following facetious analogy to humans (and other creatures of 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 could not possibly lead to cognition.*

To most, this conclusion is clearly absurd, thus putting into question the same (Justaic) reasoning applied to LLMs. Of course, one might (rightly) object that there are important dis-analogies between next-token prediction and "next-child production." For one, our ancestral environment was far more complex than the online text corpora used to train LLMs. Combined with a flexible nervous system, competition for scarce

resources, and so on, a creature of evolution could develop *instrumental objectives* (and abilities) that are more conducive to cognition than next-token prediction.

However, consistency dictates that such arguments also consider the possibility of similar instrumental objectives in AI systems such as LLMs—a prospect researchers have been taking seriously (e.g., Hubinger, van Merwijk, Mikulik, Skalse, & Garrabrant, 2019). These instrumental objectives need not be especially complex: many foundational theories in psychology and neuroscience posit relatively simple objectives as a fundamental component (e.g, "predictive brain", Clark, 2013), especially in the domain of language processing (Ryskin & Nieuwland, 2023). According to these views, humans can be represented, for instance, as pursuing activities that increase the release of positively valenced neurotransmitters such as dopamine and serotonin. If simple objectives are not sufficient for cognition, perhaps humans are in a similar position to LLMs.

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

1.2 Anti-Anthropomorphism

"It's just a machine."

The second form Justaism we call *Anti-anthropomorphic Justaism*. It is the claim that attributing cognition to machines is a fundamental error. In its strongest form, it argues that such thinking commits a category error since cognition is *by definition* a human capacity. On this view, the essential capacity that LLMs lack and humans possess is just that: humanness.

Whilst logically valid, we would argue that this view is unproductively restrictive towards the use of concepts in science. Advances in scientific theory are often based on the realization that a concept is more general than previously believed. One related example comes from animal cognition research, where, in response to a growing body of empirical evidence, researchers began to see great utility in ascribing cognitive capacities previously thought to be uniquely human, including consciousness (New York University, 2024), to other animals as well (**?**)—parrots notwithstanding! 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 whilst it is not a problem *in principle* to talk about LLM cognition, the burden of evidence for doing should be set very high. One reason for this would be to guard against the Eliza effect, which refers to the human propensity to all-too-liberally ascribe "thought" to even the simplest of machines (Weizenbaum, 1966). In this view, saying, "It's just a machine," is tantamount to saying, "Let's not make this mistake again".

While we agree that it is important to reject naive anthropomorphism, it is easy to overcompensate. After all, 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 and

impressive that it would not make sense to ascribe them to "lesser" systems (see, e.g., De Waal, 2016). In the AI context, 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, 2020).

Anti-anthropomorphic Justaism cautions against prematurely extending cognition to non-human systems. However, its restrictive position may overlook the power of conceptual generalization. Against the countervailing tendency to view human cognition as exceptional, we would advocate for a more substantive discussion of the theoretical merits of extending cognitive capacities to 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. These views may be related but do not form an obvious flavour, nor are they necessarily question-begging enough to be considered instances of Justaism. 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)—the former of which, it is argued, LLMs do not have access to due to, for instance, their lack of real-world grounding. To clarify, these positions are considerably more substantial than their Justiac cousins mentioned earlier. We are thus more sympathetic to these views in the sense that we believe that LLMs that are more physically and socially grounded are also likely to be more cognitively advanced.

Nevertheless, we would caution against confidently concluding that LLMs lack cognitive capacities on these bases. Not only do the arguments draw on notoriously hard-to-defined concepts (e.g., *meaning* or *grounding*), they also rely heavily on thought experiments of a specific nature: these thought experiments make an appeal to our intuitions by showing that a system in an analogous but more intuitively transparent position to an LLM appears cognitively lacking (e.g., Bender & Koller, 2020; Searle, 1980). However, the intuitions evoked by such thought experiments can be misleading (Dennett & Dennett, 1993), 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 model performance (Kaplan et al., 2020). Although we share the intuitions evoked by these perspectives, we thus resist updating too strongly on them.

Finally, we wish to turn to the wider public discussion of LLM cognition. This has revealed other Justaic positions, such as those emphasizing the importance of consciousness, self-awareness, emotion, life, agency and other capacities that LLMs are believed to lack (e.g., Machine Learning Street Talk, n.d.). Unfortunately, it is beyond the scope to properly address these concerns here. Future work will be needed to spell out why each of these capacities is *necessary* for cognition (as opposed to just being incidentally the case in humans), and *which aspects* of cognition they are necessary for. For instance, certain conceptions of *thinking* might by definition require

consciousness, in which case the question of whether LLMs are conscious becomes crucial to LLM *thought*, but perhaps less so for *understanding*, *reasoning*, and so on.

2 Conclusion: A more measured discussion

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

Concerning empiricism (i), 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 observables: that is, on system *behavior* (and, to a lesser extent, internal states). As Trott, Jones, Chang, Michaelov, and Bergen (2023) note, axiomatic (*a priori*) rejections of LLM cognition can lead to positions that have no empirically testable implications. Not only does this run contrary to good scientific practice, it can lead to discussions of that are socially or otherwise inconsequential. After all, it is predominantly the behavior of a system that impacts society.

Regarding modesty (ii), we would reiterate that human history is littered with delusions of human exceptionalism, especially when it comes to cognition De Waal (2016). This is despite the fact that our understanding of the mechanisms underlying cognition is still emerging. Thus, although we fully support cautioning against the dangers of (naive) anthropomorphism, we see the current work as advancing a backstop against the opposite tendency: viewing human cognition as too special (and sufficiently well understood as to know that it is too special) to ascribe to LLMs.

Finally, we recommend consistency (iii). When investigating cognition in LLMs, we find it helpful to ask: are we applying the same standards to LLMs as we would to the same kinds of investigations in humans? For instance, if we wish to reduce LLM cognition to next-token prediction, the onus is on us to 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 we as humans are susceptible to, and consequently may not so easily identify or view as significant. Not only do these considerations 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.

Ultimately, the jury is still out on the existence and extent of LLM cognition. Research has demonstrated interesting cognitive deficits in LLMs (e.g., Berglund et al., 2023), as well as impressive feats (e.g,. Bubeck et al., 2023). To date, LLMs are perhaps the most predictive, general models of human behavior and neural language data available (Tuckute, Kanwisher, & Fedorenko, 2024). The ball is now in the skeptic's court to show why this observation does *not* constitute evidence of cognition.

Supplementary information. If your article has accompanying supplementary file/s please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
- Ethics approval and consent to participate
- Consent for publication
- Data availability
- Materials availability
- Code availability
- Author contribution

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

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,

Dennett, D.C., & Dennett, D.C. (1993). *Consciousness explained*. Penguin uk.

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. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, ,

Machine Learning Street Talk (n.d.). *Machine learning street talk.* YouTube channel. Retrieved from https://www.youtube.com/c/MachineLearningStreetTalk ([Accessed: 23 May 2024])

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,

New York University (2024, April 19). *The new york declaration on animal consciousness.* Retrieved from https://sites.google.com/nyu.edu/nydeclaration/declaration (Accessed: 2024-05-06)

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*(5), 601,

Ryskin, R., & Nieuwland, M.S. (2023). Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*, ,

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424,

Trott, S., Jones, C., Chang, T., Michaelov, J., Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309,

Tuckute, G., Kanwisher, N., Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*, ,

Turing, A.M. (1950, October). I.—COMPUTING MACHINERY AND IN LIGENCE. *Mind*, *LIX*(236), 433–460, https://doi.org/10.1093/mind/LIX.236.433 Retrieved from https://doi.org/10.1093/mind/LIX.236.433 (_eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf)

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45,

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,