Title: Other Minds Problem Revisited at the Age of AI

Content:

The traditional problem of other minds has long been framed as a tension between infallible knowledge of one's own mind and error-prone observational knowledge of those of others(Avramides, 2001). Over the past decades, a dominant philosophical strategy has emerged that reframes this problem rather than resolving it. Against the Cartesian background, philosophers such as Chihara and Fodor(1965) propose that we can still acquire knowledge of the unobserved mental entities by appealing to explanatory theories that infer mental states from observable criteria. This abductive strategy shifts philosophical focus toward the causal connections between mental states and observable behaviours, thereby deprioritising first-personal authority and introspective access in favour of a general explanatory framework (Carruthers, 2011; Goldman, 2006; Gopnik & Wellman, 1994). Consequently, the traditional problem of other minds is 'no longer an interesting problem', as Fodor(Fodor, 1979) remarks.

However, by representing a 'linguistic subjectivity' that is indistinguishable from human output, recent developments in Large Language Models (LLMs) appear to bring back the problem of other minds into view: if an entity talks like a subject, are we thereby committed to attributing to it subjectivity with a first-personal perspective[1]? This paper argues that the appearance of subjectivity in LLMs is illusory. Specifically, it argues that this illusion arises from the uncritical extension of abductive inferential strategies originally developed for understanding human minds.

The argument will be arranged in three stages. First, I explain why LLMs' outputs readily invite attributions of subjectivity. I argue that the modern abductive theorists, such as Gopnik(1994) and Carruthers(2011), provide a theoretical loophole that is tacitly exploited in AI-mindedness debates. According to this framework, by observing linguistic data and performing an Inference to the Best Explanation (IBE), we can posit 'mental states', such as beliefs, desires, and intentions, as the underlying causes. Subjectivity, in this framework, amounts to the summation of causal mental explanations. When applied to LLMs, this abductive mechanism encounters a 'false positive'. Because the model's output aligns with the patterns of human reasoning, through abductive reasoning, it seems that we can naturally attribute mental states to them and reasonably conclude that these LLMs have subjectivity(Chalmers, 2023; Kosinski, 2024). This generates a modern problem of other minds: should we conceive of a complex language model as possessing subjectivity? If so, what becomes of the distinction between human and non-human 'others'?

[1] In this paper, I use 'subjectivity' to refer not to any minimal capacity for information processing or self-reference, but to a robust form of mindedness that is constituted by the capacity to occupy a first-personal perspective with normative and epistemic authority.

Stage two aims to give two reasons in arguing against the idea that LLMs' outputs are made from a subjective perspective, and we should not regard LLMs as minded. The first reason is articulated through Brandom's(1994) inferentialism. In Brandom's framework, using language is a normative practice of 'giving and asking for reasons'. To assert 'I am in pain' is to undertake an inferential commitment. The subject is not just a processor of information but also a subscriber to responsibilities. LLMs, however, operate in a 'de-normatised' environment. They may simulate or mimic the syntax of an inference, but they cannot be held accountable for the consequences of their claims. An LLM cannot 'commit' because it has no social or existential stake in the game of reasons. On the contrary, it navigates a map of words without ever inhabiting the territory of responsibility(Bender et al., 2021), and for this reason, LLMs' linguistic performances fail to qualify as expressions of subjectivity, even when they are indistinguishable from competent human language use.

Correlated to the first reason, the second reason says that the lack of Brandomian commitment leads directly to a failure of the first-personal authority, which is essential for minded subjects. Drawing on Moran's (2018) account, I argue that first-personal authority is not a matter of epistemic privilege grounded in inner observation, but a deliberative authority exercised through avowal. When a human subject speaks from the first-personal perspective, they do not merely report their mental states; they make up their minds and stand behind their commitments. This authoritative 'I' is the anchor of subjectivity, and it represents what a real-minded human is like. In contrast, an LLM's 'I' is merely a linguistic indexical, generated without deliberation or ownership. Lacking the capacity for avowal, LLM discourse amounts to a report generated from nowhere, rather than an expression of a situated subject.

At the final stage, I argue that both the normative commitment and epistemic authority require a situated and embodied agent. I argue that the first-personal perspective is intrinsically linked to such a kind of localised embodiment. Following the tradition of P.F. Strawson(1959) and Cassam(2007), I argue that we must recognise that a minded subject is not a disembodied ego, but a localised individual. A first-personal perspective is constituted not by privileged access, but by restriction: to have a perspective is to encounter the world from a determinate, embodied point of view that is necessarily not-everywhere. Embodiment introduces vulnerability, such as pain, resistance, and sensory limits. These constraints are what allow a subject to have a 'stake' in the world, which in turn enables the authority in Moran's words and linguistic normativity in Brandom's words. LLMs, existing in a frictionless digital vacuum, lack all these embodied boundaries that are necessary to distinguish a self from the world. It is precisely because of these de-perspectivised characteristics that offer us a reason to argue against its subjectivity.

Back to the question left at stage one, I propose that the ‘other’ in the problem of other minds essentially refers to entities capable of occupying a restricted first-personal perspective grounded in normative and embodied agency. On this understanding, the problem of other minds remains philosophically significant, not only for clarifying debates about AI subjectivity, but also for motivating a reassessment of abductive approaches to mental state attribution.

(word count: 975)

Bibliography

Avramides, A. (2001). *Other minds*. Routledge. https://doi.org/10.4324/9780203870174

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Vol. 183, Number 183). Harvard University Press.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. OUP Oxford.

Cassam, Q. (2007). *The possibility of knowledge*. Clarendon Press.

Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*, *1*.

Chihara, C. S., & Fodor, J. A. (1965). Operationalism and ordinary language: A critique of Wittgenstein. *American Philosophical Quarterly*, *2*(4).

Fodor, J. A. (1979). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, *3*(1).

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511752902.011

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Moran, R. (2018). *The Exchange of Words: Speech, Testimony, and Intersubjectivity*. Oup Usa.

Strawson, P. F. (1959). *Individuals*. Routledge.