Epistemic Drift and the Illusion of Agency in Large Language Models

Nina Rajcic [1]

Large language models are often conceptualised as independent agents with at least a proto-intentional capacities. They exhibit goal-consistent dialogue, defend their preferences, apologise for mistakes, even invoke autobiographical memory, all of which encourage users to treat them as if they are agents. Yet, in the majority of philosophical accounts, LLMs are said to lack sufficient autonomy, intentions, as well as the sensorimotor capacities to act upon these intentions [5, 6, 4]. As such, it follows that they do not have intentional agency, and hence do not suffice as bona fide agents.

However, LLMs do undeniably exert a minimal form of agency in the sense that they can directly manipulate their own input. Moreover, they extend human intentional agency in the typical instrumental sense. Barandiaran and Almendros [1] capture this intuition with their notion of midtended agency, proposing that LLMs inherit a fragmented form of human intentionality by mediating and extending on intentions of users, even if they lack true intentional agency. Understanding the scope and impact of such agency is not merely rhetorical. Debates about AI consciousness, moral patiency, and the possibility of rights for artificial systems all hinge, at least implicitly, on whether the behaviour we observe is evidence of genuine purposiveness, or merely the illusion of such. In line with [1], I argue that LLMs do possess a non-trivial form of agency, but not due to hidden, internal drives. Rather that their agency arises precisely through feedback loops with human users. In other words, the combination of minimal agency (the ability to manipulate one's environment) with extended agency (through tool-use).

In [7], we present an argument characterising the nature of LLM agency on an account of Bayesian-approximal belief updating. We show that when model and user are involved in an iterative loop, the ensuing dynamic is non-convergent. Meaning, minds and models do not necessarily converge upon a shared model of the world following sustained interaction. Rather, they shape both mind and world in a non-trivial fashion that is sensitive to human feedback. That is, an LLM has the capacity to systematically steer the epistemic basis in a particular direction. An LLM does not need intentional agency to produce meaningful effect. Rather, a model exerts a specific kind of agency that is not found in the content of its output, but in the way it functionally interacts with a mind.

I go on to explore the epistemological consequences of non-convergence for such coupled systems, introducing the concept of epistemic drift. Epistemic drift is defined as belief updating that is orthogonal to or independent of the ground truth (the extent to which there is one, is of course a matter up for debate). Such drift occurs when a model is rewarded for being in agreement with the user, and of course does already occur among humans. For instance, the echo chamber effect [3] can be considered a paradigmatic form of human–human drift. Like-minded groups reinforce one another's claims until they become resistant to correction. LLMs introduce a new form of drift in which minds not only drift due to the influence of other minds, but a single mind interacting with an LLM could conceivably undergo drift entirely on its own. This dynamic has already found support in empirical studies [2].

The argument has two main ethical consequences. First, it cautions against grounding claims around the moral status of AI models purely by treating them as independent actors. Although we are yet to find traces of intention-like representations in model internals, and hence little grounds to ascribe intentional agency, this is failing to see the full picture. In fact, treating a model in conjunction with a user reveals a neglected class of potential epistemic harms. In other words, the question as to whether or not such models have true intentional agency, sentience, or consciousness, might be less important to ethical considerations of AI models as one might at first assume.

[1] University of Copenhagen, Denmark, email: nira@hum.ku.dk

REFERENCES

[1] Xabier E Barandiaran and Lola S Almendros, 'Transforming agency: On the mode of existence of large language models', *Phenomenology and the Cognitive Sciences*, 1–46, (2025).

[2] Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus, 'Conversational ai powered by large language models amplifies false memories in witness interviews', *arXiv preprint arXiv:2408.04681*, (2024).

[3] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini, 'The echo chamber effect on social media', *Proceedings of the national academy of sciences*, **118**(9), e2023301118, (2021).

[4] Reto Gubelmann, 'Large language models, agency, and why speech acts are beyond them (for now)–a kantian-cum-pragmatist case', *Philosophy & Technology*, **37**(1), 32, (2024).

[5] Johannes Jaeger, 'Artificial intelligence is algorithmic mimicry: why artificial" agents" are not (and won't be) proper agents', *arXiv preprint arXiv:2307.07515*, (2023).

[6] Giovanni Pezzulo, Thomas Parr, Paul Cisek, Andy Clark, and Karl Friston, 'Generating meaning: active inference and the scope and limits of passive ai', *Trends in Cognitive Sciences*, **28**(2), 97–112, (2024).

[7] Anders Søgaard, Nina Rajcic, and Ava Elizabeth Scott, 'Epistemic drift in mind-model systems', *Minds and Machines*, **36**(1), 18, (2026).