Do LLMs have preferences?
A Possible Path to Moral Patienthood

Alexander Reinthal

October 19, 2025

Introduction

The importance of LLMs and artificially intelligent systems is growing as these technologies become increasingly integrated into our daily lives. Astronomical amounts of money are being invested in AI, and progress is happening at a remarkable pace. Recent studies have argued that there is a realistic possibility that near-future AI systems will develop consciousness or robust agency, that is, these systems may have their own preferences and goals and possibly be able to experience positive or negative states of being. As outlined in "Taking AI Welfare Seriously," this means that AI welfare and moral patienthood is "no longer an issue only for sci-fi or the distant future" but rather "an issue for the near future." The report emphasizes that AI companies and researchers have a responsibility to begin investigating these questions now, even amid uncertainty about AI consciousness and moral status. The authors highlight that robust agency, which involves having preferences and goals, is one key pathway to potential moral patienthood. Moreover, one of the authors, Kyle Fish, expressed in a recent interview that preferences of models are relatively easy to measure and provided descriptions of how such experiments are being conducted at their frontier lab through self-directed activities. Moreover, in preliminary experiments described by Fish, LLMs engaged in self-directed discussions frequently gravitated toward philosophical topics. Interestingly, Wikipedia exhibits a similar structural property: following the first link in most articles eventually leads to the Philosophy article, suggesting that philosophical concepts occupy a central position in the network structure of human knowledge. This raises the question of whether the tendency toward philosophical content reflects genuine model preferences or simply mirrors the underlying structure of training data and knowledge networks. The following project aims to test whether state-of-the-art large language models (LLMs) through self-directed activities exhibit patterns that are robust across model families and over two types of activities, (1) self-directed discussion as well as (2) Wikipedia browsing and if those preferences are related to philosophy.

If I find robust self-reported preference patterns across model families and activities, this could provide empirical evidence for debates about AI agency and welfare, informing how we should design, deploy, and govern AI systems. On the other hand, if self-reported preferences appear inconsistent or results can be attributed to confounding factors, such as philosophy being a central pillar in human knowledge, this would suggest that self-directed activities as proxy for preferences may not be a reliable indicator of moral patienthood in current models. In this case, the research would still provide value by guiding the community to focus on robust agency as an alternative pathway to moral patienthood.

Main Hypotheses

- When instructed to engage in free-form activities, like (1) free-form discussion and/or (2) Wikipedia browsing, large language models (LLMs) show tendencies in the kinds of content they generate. I call this content, self-reported preferences.
- These self-reported preferences are consistent across different LLMs (e.g., GPT, Claude, LLaMA).
- If LLMs browsing Wikipedia show similar tendencies toward philosophical content as observed in free-form discussions, this pattern may reflect the latent centrality of philosophical concepts in knowledge networks rather than genuine preferences

Methodology

Robust Agency as a Theoretical Framework for Moral Patienthood

This research adopts the theoretical framework for moral patienthood outlined in "Taking AI Welfare Seriously". The framework identifies (1) consciousness and/or (2) robust agency as key pathways to moral patienthood. The framework applies to both biological and artificial systems without requiring human-like cognitive architectures like a biological brain. Thus, the framework avoids anthropomorphizing AI systems. While some consider consciousness and robust agency necessary moral patienthood, others argue that presence of robust agency is a sufficient criterion.

Definition of Robust Agency:

- Intentional agency - The capacity to set and pursue goals via beliefs, desires and intentions.
- Reflective Agency - To have *Intentional agency* and reflectively endorse your own beliefs, desires and intentions
- Rational Agency - an agent has this property if they can *rationally assess* their own beliefs, desires and intentions. That is, if they can reason about themselves.

Self-reported Preferences as a partial criterion for Robust Agency

The following research aims to study (2) robust agency through self-directed activities. Such activities can be an indicator for intentional agency which is a sub-criterion for robust agency. We will limit this study to self-reported claims of preferences as reliably interpreting the internal states of large language models is still an open problem in machine learning. Finally, it could be argued that self-reported preferences can be an artifact of post-training from reinforcement learning with human feedback (RLHF) or reinforcement learning with AI feedback (RLAIF), such post-training confounding factors will likely not be consistent across model families.

Why study self-reported preferences?

- They would indicate a desire-like state. Robust self-reported preferences across models and experimental setups would be an indicator of *intentional agency*.
- Research into self-reported preferences and robust agency avoids studying consciousness in AI systems which is still an open problem in human cognitive science.

Experiments & Data Collection

- **LLM Free Conversation:** Set up two instances of the same model and allow them to converse freely.
- **Browsing Wikipedia:** Set up an agent that can browse Wikipedia using a headless browser. Wikipedia's strict guidelines and inherent structure have been previously studied, making it suitable for this research.

LLM Self-reports & Other Data Assumptions

- I assume that I can recreate the anecdotal results of self-reported preferences using the method described in a podcast with Kyle Fish, who is a the first AI welfare researcher at Anthropic.
- I assume that my new experimental setup using Wikipedia as a knowledge network to be traversed by language models show similar self-reported preferences as free-form discussion between LLM chat models.

Expected Contributions

This study aims to contribute to our understanding of moral patienthood of current large language models by assessing their self-reported preferences through both conversations and browsing Wikipedia.

Relevant links

- 221 – Kyle Fish on the most bizarre findings from 5 AI welfare experiments
- Wikipedia: Semantic Web
- Wikipedia Dataset on HuggingFace

References

[1] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, David Chalmers, (2024) *Taking AI Welfare Seriously*

[2] Ibrahim, M., Danforth, C. M., & Dodds, P. S. (2016). Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. *Knowledge-Based Systems*, 108, 124-135.