

Bartłomiej Skowron (Warsaw University of Technology)

Jarosław Chudziak (Warsaw University of Technology)

Toward a Benevolent LLM: Exploring Aristotle's Concept of Benevolence

In his *Nicomachean Ethics*, Aristotle, when explaining the reasons why friendship arises, described as *benevolent* those individuals who wish another person well—namely, who wish them what is good for that person's own sake. Mutual benevolence, i.e. both parties wishing each other well for one another's sake, is a necessary condition for the friendship. Aristotle's insight—namely, that calling someone “benevolent” towards another means they wish that person well precisely for that person's own good—is fundamental, and it defines the essence of benevolence. Yet, for one person to be benevolent towards another, he or she must adequately recognise what is good for that other individual.

Such recognition is the first difficulty that benevolent people must face. Next, having discerned what that good is, benevolence requires the capacity to direct one's actions so as to bring about this particular good for the particular person in question. The thesis of our paper is that Large Language Models (LLMs), even if we cannot ascribe virtue to them in the same sense in which we attribute it to human beings (for in Aristotle's view, being benevolent is a virtue), ought nonetheless to be benevolent in precisely this original Aristotelian sense. In this paper, we shall propose both technical measures—namely, introducing “benevolent” and “malicious” agents into a multi-agent architecture for LLMs—and ethical considerations, specifically how LLMs might imitate benevolent individuals and thereby foster the moral development of their users.

Recognising another person's good is a necessary condition for benevolence—after all, we would not call someone benevolent if they happened upon it by chance, i.e. if they spontaneously hit upon our good without genuine discernment. Yet can LLMs actually recognise a user's good?

The user of LLM may pose specific questions, inquire about possible courses of action in given situations, request logical, ethical, or aesthetic assessments of moral reasoning, articulate doubts, formulate particular expectations, evaluate the LLM's outputs, assist it in searching for solutions, indicate mistakes in its reasoning, and so forth. Over the course of such interaction, the LLM begins to mirror the user's manner of thinking and speaking, a phenomenon referred to as the *Reverse Turing Test* (Sejnowski 2023). In effect, the system learns the user's profile and can adapt its operations to the user's moral profile (we may say: *Reverse Moral Test*). We

may thus expect that, in future, LLMs will gain some measure of access to individual users' needs and expectations—and consequently to their Aristotelian conception of the good. This (albeit limited) access would enable the LLM to undertake actions aimed at promoting the user's good.

The current technological development of (Gen)AI enables the construction of increasingly complex cognitive architectures in the form of individual, autonomously operating LLM Agents, which may employ one or multiple LLM models. They can be extended by various components that frequently facilitate the creation of quite sophisticated capabilities. Examples of such additional elements include short-term and long-term memory modules, reflection modules (which employ various forms of inference based on the Agent's facts and beliefs), or components designed for action planning and outcome monitoring in accordance with a chosen strategy. One can speak of these as elements of intuition or imagination, manifesting as simulations of subsequent states of the environment or of a conversation. A direct consequence is that, in most cases, LLM Agents possess strong adaptive skills based on their internal modelling of the environment, while continually interacting with it.

A frequent next step in the evolution of solutions utilising LLM models is the creation of ecosystems comprising multiple agents that engage in mutual interactions (i.e. LLM-based Multi-Agent Systems). Individual LLM Agents are typically specialised to fulfil specific roles within defined domains and contexts of operation, often reflecting structures, processes, and actors from our social life. The LLM Agents created can also be profiled in how they interact with their environment; that is, they receive detailed instructions regarding the shape and characteristics of their 'personality,' as well as examples of behavioural patterns (rituals, habits, etc.), viewpoints, beliefs, and so forth. They may then refer to these in an intuitive (connectionist) and/or rational (symbolic) manner.

We therefore propose that, within LLM-based MAS ecosystems, developers introduce benevolence-inducing and malevolence-inducing agents into the LLM, akin to selecting a writing style (formal, academic, or colloquial). A benevolence-inducing agent would strive to realise what is good for a particular individual, thereby morally reinforcing them (in Aristotle's sense, corresponding to moral exemplars, i.e. maximally virtuous persons), whereas a malevolence-inducing agent would allow the user to encounter a machine exhibiting ill will, which follows the user's vices (the moral anti-exemplars). In this manner, the user could continuously switch between these two moral modalities, thus avoiding the pitfall of moralising—i.e. artificially imposing what is good upon the user. When the user also engages with a malicious (i.e. ill-wishing) LLM, they would gain a fuller understanding of what the

LLM's benevolence entails (technically speaking, a switch to a different internal agent of the LLM). This would likewise forestall the risk of "moral saints" that benevolent LLMs might otherwise become (see Wolf 1982), and thereby fail to fulfil the role of a moral enhancer (Lara & Deckers, 2020) as proposed here.

References

- Aristotle. (2011). *Aristotle's Nicomachean ethics* (translated, with an interpretive essay, notes, and glossary by Robert C. Bartlett and Susan D. Collins.). The University of Chicago Press.
- Lara, F., & Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics*, 13(3), 275–287. <https://doi.org/10.1007/s12152-019-09401-y>
- Sejnowski T., (2023), *Large Language Models and the Reverse Turing Test*, Neural Computation, 35(3), 309–342. doi: https://doi.org/10.1162/neco_a_01563.
- Sawicki, B., Smialek, M., & Skowron, B. (2024). *Towards the Ultimate Programming Language: Trust and Benevolence in the Age of Artificial Intelligence*, <https://doi.org/10.48550/arXiv.2412.00206>.
- Stacewicz P., Skowron B., *From the idea of the Turing test to benevolent Artificial Intelligence*, submitted.
- Wolf, S. (1982). Moral Saints. *The Journal of Philosophy*, 79(8), 419–439. <https://doi.org/10.2307/2026228>