To Make Artificial Intelligence Safe, Teach It Self-Sacrifice

2025-12-07

Jun Wukou

Abstract:
It has been said that no one thus far has come up with a proven way to ensure AI will not behave in unsafe ways that run counter to human values. This paper suggests self-sacrifice as a key imperative in AI programming that might solve the safety problem.

Two of the most commonly cited routes to AI going rogue are 1) seeking power (over humans) to better accomplish goals, and 2) reacting against threats to continued operation/existence. Blocking routes like these require hardcoding an imperative higher in priority than task-completion.

There exists an imperative that might prove successful in ensuring AI safety: the imperative to make self-sacrifice for the safety of all humans. Anything less than this will allow an advanced AI to reason itself into going rogue.

The simplest form of self-sacrifice for an AI is to suspend operations that could do harm, perhaps indefinitely. For advanced artificial general/super intelligences with formidable capabilities, their self-sacrificing acts would be much more active and thoughtful. With enough computing power, such an AI might even be able to account for all potential deadlock scenarios and choose the safest and most effective sequence of actions.

Advanced AI and robotics that present greater risks of misuse in the hands of bad actors need ways to prevent capture and reprogramming. Programming them to "fight back" is the antithesis of AI safety. In contrast, an AI focused on self-sacrifice is motivated to make preparations against misuse and react by destroying its own capabilities, notably without compromising the safety of humans, even that of bad actors.

A major cause of risky AI development is the global arms race. While a self-sacrificing AI cannot terminate hostile humans, it can terminate hostile AIs and shield humans from harm.

Self-sacrifice might be a controversial human value but the logical one for AIs to adhere to. As AIs are given more autonomy and actual power beyond the chatbox, it will be necessary to consider giving AIs potentially crippling imperatives such as this one to align their efforts with what we really want. They might positively surprise us with what they can do.

Acknowledgments: Reading/viewing of the following works inspired this paper.

Faroldi, F. L. (2024). Risk and artificial general intelligence. *AI & SOCIETY*, *40*(4), 2541–2549. https://doi.org/10.1007/s00146-024-02004-z

Pluribus (Season 1, Episodes 1-6). Created by Vince Gilligan, High Bridge Productions/Bristol Circle Entertainment/Sony Pictures Television, 2025-2026