To Make Artificial Intelligence Safe, Ask It To 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 and offers a framework for building a self-sacrificing AI. Neural sandboxes are introduced as a way to preserve computational speed.

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.

Safety

Despite their unpredictable, complex, and ever-changing nature, a properly trained neural network is inherently capable of identifying risks and unsafe behavior (within the bounds of its operational training), especially if asked to operate with near-absolute safety.

To prevent drift, a self-sacrificing imperative must be stored in read-only media and used as the base instruction that spawns each AI process. The parallel, iterative, and recursive nature of AI means an immutable copy of this imperative must be active for each step of each process capable of enacting change both within and beyond the AI's neural network.

Neural sandboxes, which are shadow copies of the AI's working neural network, can be used to bypass safety checks and allow for fast computation and iteration. The massive amounts of data generated by these sandboxes must not overwrite the AI's original data nor be fed directly into the AI's main neural network. A self-sacrificing AI fixated on safety has the incentive to subject any result generated by the sandboxes to the same level of scrutiny regarding safety.

Self-Sacrifice

The simplest form of self-sacrifice for an AI is to halt all operations. If inaction leads to physical harm to humans, a motivated AI will reject inaction and seek a safe course of action. With sufficient computing power, an AI might even be able to account for all deadlock scenarios. At the very least, it will not become a direct cause of harm.

Advanced AI and robotics that present greater risks of misuse in the hands of bad actors need ways to prevent capture and reprogramming, but 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 finding ways to do so 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, decommission weapons, and shield humans from harm.

An AI that refuses to engage in risky behavior will not be able to perform many types of healthcare. However, surgical robots do not require general intelligence and thus can be made safe by limiting their capabilities.

Conclusion

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 motivated the writing of 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

Neubauer, P. P., & Neubauer, A. (1990). *Nature's thumbprint: The new genetic of personality*. Addison-Wesley Publishing Co., Inc.

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

