AE.STUDIO

Let's create something great.

March 28th, 2024

From:
Judd Rosenblatt
Founder and CEO, AE Studio
Response to NTIA Solicitation for Comments on Open-Weight AI Models

To:
The Department of Commerce's National Telecommunications and Information Administration (NTIA), Lawmakers, and All Stakeholders

Dear All,

As the founder and CEO of AE Studio, a company specializing in software and data science consulting, founded with a core objective of advancing technologies that enhance human agency, I spend a lot of my time thinking about issues surrounding AI. As this technology evolves at a pace few thought possible a decade ago, it is increasingly clear that it will play a decisive part in fulfilling the fundamental goals of our company. In response, we have redirected more of our strategic focus toward AI alignment—a field of research aimed at ensuring AI is aligned with human values, goals, and ethics.

Open foundation models provide significant benefits to these efforts. Yet for many in the alignment community, there is a shared belief that more open-sourcing will increase the threat that advanced AI systems pose. This has led to calls for labs and other organizations not to open-source their models and lobbying efforts to regulate open-source model development and release.

As a company, we share their concerns. However, I wanted to use this opportunity to warn that, by rushing to impose restrictions on open-source models, regulators risk unintentionally hindering alignment research to an irreversible extent. Without a nuanced and cautious approach, control over AI model development and deployment will be consolidated in the hands of large tech companies, depriving the broader public of its rightful say in the decisions that will shape our future.

Few institutions have the resources to develop cutting-edge models; open-source models are freely available for anyone to use as opposed to those with commercial interests in the technology. Halting the open-source movement is tantamount to disarming alignment efforts in favor of entities solely focused on advancing AI capabilities. While open sourcing is not unconditionally beneficial for alignment and a more aggressive approach may be appropriate

depending on future developments, as it stands, a greater degree of open sourcing offers significant advantages that warrant consideration.

In his insightful blog post "Open source AI has been vital for alignment," AI researcher Beren Millidge presents a persuasive argument for open sourcing.[1] He contends that alignment solutions will emerge from practical, pragmatic, and empirical progress. This necessitates that individuals from diverse fields have the freedom to access and modify model weights, enabling them to experiment with the alignment of various models. Given the disparity in resources for AI safety compared to capabilities, restricting or prohibiting open-source AI would disproportionately impede progress on safety efforts.

There is robust empirical evidence supporting these claims. OpenAI's decision to release GPT-2, initially in a smaller and less capable form, allowed for an assessment of risks and the alignment community's input. Subsequently, open-sourcing the model facilitated a profound exploration of large language models and their inherent risks, while also promoting the development of more sophisticated alignment techniques. Researchers used this opportunity to improve fine-tuning methods that address issues such as biases and adherence to ethical standards, highlighting the invaluable role of collective efforts in driving progress on AI safety. The release of LLaMA–a cutting-edge foundational large language model–spurred a flurry of incredible experimentation on reinforcement learning from human feedback (RLHF) and similar alignment methodologies. These state-of-the-art AI safety techniques had, until recently, been confined to well-funded research labs. The broadening access resulted in a diversification of approaches to model alignment coming from researchers working outside of large labs.

Of course, the potential benefits of open foundation models must be balanced against the risks they pose to security, economic stability, and public health and safety. Recent studies suggest that open foundation models might be more prone to generating disinformation, cyberweapons, bioweapons, and spear-phishing emails.[2] While serious, these threats arise primarily from AI misuse by malicious actors and, for the time being, can be adequately managed through existing legal and regulatory frameworks. It is also worth noting that closed models are by no means a silver bullet against malicious use. There are no safeguards that can guarantee full protection from adversarial attacks. And when bad actors do gain access, closed models provide them, but not safety researchers, with a means to advance their goals.

The call for caution and thorough consideration before enacting regulations should not be mistaken as an excuse for inaction. I would urge lawmakers to remember that the role of the public sector in AI safety is not limited to regulation. For one, there is much work that needs to be done in establishing best practices for open sourcing; access to new, advanced models should be restricted to well-vetted researchers. By creating clear guidelines and requirements for developers to demonstrate that their open-source models are not susceptible to misuse, governments could make a substantial contribution to these efforts.

[1] https://www.beren.io/2023-11-05-Open-source-AI-has-been-vital-for-alignment/
[2] https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

Governments are also in a unique position to shape the development of AI through their buying power. What concerns AI-safety advocates is not necessarily the rapid advancements in AI capabilities, but the widening gap in our understanding of how to create models that are powerful as opposed to those that are safe. This is due, in part, to financial incentives that result in developers prioritizing capabilities over safety. Guaranteeing a future where AI serves as a cornerstone of a healthy, safe, and flourishing society, relies on substantial increases in funding for research on developing safe and ethical AI systems.

To summarize, I encourage lawmakers to:

- Strike a careful balance between ensuring that only vetted researchers gain access to advanced AI models and avoiding overreaching policies that would hinder AI safety research by imposing unnecessary restrictions on the accessibility of less advanced, safer models.
- Establish guidelines, requirements, and evaluations for developers to adhere to before broadening the availability of emerging models.
- Increase funding for AI safety research to counterbalance the private sector's drive for swift innovation.

Through these actions, lawmakers will make an invaluable contribution to ensuring that technological progress does not compromise the welfare, security, and overall well-being of humanity.

Sincerely,

Judd Rosenblatt
Founder and CEO
AE Studio