Response to NTIA RFC on 'Dual Use Foundation Models with Widely Available Model Weights'

Aviv Ovadya
Founder and Executive Director
AI & Democracy Foundation
aviv@thoughtfultech.org
March 26, 2024

The AI & Democracy Foundation (AIDF) appreciates the opportunity to provide input on the critical issue of dual-use foundation models with widely available weights. AIDF is a research and DARPA-model grantmaking organization, created in order to provide new approaches for enabling international buy-in to meaningfully address exactly these kinds of challenging questions, where the choices of any state (or corporation) can impact everyone.

About the author: Aviv Ovadya leads AIDF and has extensive experience in this specific area, helped initially raise the alarm in 2017 and 2018 around many of the potential risks of what are now called foundation models, and has written about the benefits and challenges of variably open AI systems, and the need to go beyond red teaming AI systems like GPT-4 in order to ensure resilience to the inevitable impacts of open systems. He has consulted, advised, or otherwise been affiliated with organizations across the ecosystem including the Partnership on AI, the Berkman Klein Center (Harvard), the Belfer Center (Harvard), the Centre for Governance of AI, Cohere, and OpenAI, and is regularly quoted by publications such as the New York Times on topics related to open AI models. The core focus of his current work is on the mission of AIDF, by providing viable 'democratic' alternatives beyond the default paths of either (a) closed corporate/autocratic control of AI systems, and (b) ungovernable and irreversibly open AI systems.

This comment will focus on **question 7, and particularly 7h and 7i**, though we are happy to discuss other questions given our experience pioneering work in this space over the past seven years (under prior affiliations, before this organization was formed).
However, we first review the some of the key challenges relating to widely available model weights below, which relate to a number of the other questions:

Key challenges relating to widely available model weights

1. **Model release is irreversible**: Decisions to make model weights publicly available are highly likely to be irreversible.
2. **Model release can be unilateral**: Any actor, whether a nation-state, a corporation, individual, etc., can unilaterally make model weights widely or publicly available.
3. **Public model release is a global act**: No matter where in the world a model is released and made publicly available, it can have a global impact.
4. **Benefits are significant—and of unknown size**: There are huge benefits to innovation of model weights being widely available, and at least some level of widely availability of model weights may be critical for important safety and ethics research. Wide availability also may significantly decrease the concentration of power, and the unilateral wielding of that power.
5. **Risks are significant—and of unknown size**: Foundation models are already being weaponized. Many further believe that there is some threshold beyond which widely available model weights provide a net-negative impact on society. One argument goes like this:
    a. While guardrails are possible for closed models, it will be impossible to maintain them for models with openly available weights. In other words, **openly available models are " ungovernable".**
    b. AI systems built on such models will be increasingly helpful at all tasks, given the right context, including tasks which are **extremely destructive** (not only for planning, but also orchestrating, executing, troubleshooting, etc.).
    c. There is sufficient information already available to provide the appropriate context for extremely destructive tasks (including e.g. CBRN-related, debilitating cyberattacks, social engineering, mass persuasion, etc.; this assume the sufficient tacit knowledge is broadly available or can be elicited by a sufficiently capable system).
    d. Such AI systems, if they existed, would be used either intentionally by people, unintentionally, or independently, to do or support these extremely destructive tasks.

    Thus, the argument states, there is a maximum capability level for an AI system such that its model weight should not be publicly available. This then begs the question of what that level is and how to evaluate it which is the focus of the RFC. Some believe that that threshold was even before GPT-2, others believe that we are *very far* from that threshold, or that the risks will never outweigh the benefits.

6. **Preventing model release *may* only buy time**: Given 1 and 2 there is an argument that it's not even worth trying to prevent wide availability, since it would require global coordination to prevent that.

7. **Buying time *may* enable resilience measures**: The time being bought by preventing model release may be used to build resilience measures across society, though some argue that that could require enormous resource expenditure and coordination.
8. **Decentralized networks *may* be able to train powerful models:** There is some evidence to believe that individual actors or decentralized networks may be able to train models.

This combination of significant benefits, risks, unknowns, and wildcards, in combination with irreversibility and globally impactful unilateral action, makes this a particularly tricky challenge.

AIDF does not take a position on the risk versus benefit tradeoff, or what threshold is appropriate, but instead focuses on what mechanisms may enable a viable path forward for effectively choosing a threshold with sufficient buy-in such that it is effectively maintained.

Mechanisms for managing the risks and maximizing the benefits

Question 7, reproduced below, asks what mechanisms might we use to navigate this challenging dilemma.

> **7**. What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?
>
> **7h**) What insights from other countries or other societal systems are most useful to consider?
>
> **7i**) Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem?
>
> Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.

There *are* in fact very promising models from other countries and societal systems which can help navigate some of the challenges around trading off the risks and benefits of widely available model weights. AIDF has outlined one such proposal, building on mechanisms from deliberative democracy, in a recent paper for the Journal of Democracy (attached; preprint link, overview).

In particular, one of the key challenges around widely available model weights is that even if a decision was made to prevent release in one region, **it would have very little impact given the unilateral global irreversible nature of model release**. To address this requires broad global buy-in with significant consequences for defection (and potentially strong incentives to discourage investment in decentralized defection). Relatedly, even if there were effective limits on model release, that could create dangerous levels of concentration of power. As many actors would thus lose power, this would make them far more likely to defect.

Thus from a realist perspective, if there is to be any significant proposal to prevent model release over a certain threshold, there also needs to be some mechanism for collective decision-making around what actions unreleased models can be used for.

The paper outlines the potential of using deliberative democratic mechanisms for making such critical decisions—both for global agreements for open "source" AI systems, and for determining alignment targets and guardrails for the resulting closed systems (so that they are not under unilateral monopolistic control).

AIDF was founded as a result of extensive research identifying deliberative democratic processes as one of the few core mechanisms that might enable us to navigate this particularly pernicious set of challenges. These processes have shown effectiveness around the world, at every scale of government, and have now been institutionalized by the EU as part of their legislative process. There have even been global pilots conducted in association with the UN (including Chinese participants), and by private companies (e.g. by Meta). However, these processes are not yet sufficiently mature in order to handle the level of power that might be brought to bear around the most challenging questions of AI, and there is thus significant need to rapidly invest in increasing the quality and robustness of these processes particularly at a global level.

In an ideal world, new mechanisms might not be necessary and the existing levers of state and the international order would be sufficient to handle these challenges. However, in consultation with others who have worked across other challenging global coordination issues such as climate, there has been significant skepticism that any of the existing mechanisms will be sufficient given the even more difficult challenge posed by AI coordination. Thus we believe that we need rapid investment in new mechanisms, such as those discussed above and in the attached paper.

Recommendations

In light of that, we provide a few initial recommendations, though we are still early in exploring concrete next steps and we expect this to evolve:

1. **National Science Foundation (NSF) and National Institute of Standards and Technology (NIST): Fund targeted research and pilot projects**
   - NSF should prioritize funding for research on AI-augmented deliberation techniques, platforms, and governance models through programs like the Future of Work at the Human-Technology Frontier and the Fairness in Artificial Intelligence program.
   - NIST should support pilot projects applying deliberative governance approaches to AI systems development and deployment, building on its existing work on AI standards and best practices.
2. **Office of Science and Technology Policy (OSTP): Establish an interagency task force on deliberative democratic AI governance**
   - OSTP should convene a task force comprising representatives from key agencies involved in AI policy and governance, such as NTIA, NSF, NIST, and the Department of Energy.
   - The task force should develop a coordinated strategy for advancing deliberative AI governance, identify opportunities for collaboration and resource sharing, and provide guidance to individual agencies.
3. **Department of State and White House Office of Science and Technology Policy (OSTP): Collaborate with international partners**
   - The Department of State should prioritize deliberative democratic AI governance as a key issue in its international engagements, working with partners to develop shared principles and frameworks.
   - OSTP should lead efforts to exchange knowledge and best practices with international organizations focused on AI policy, such as the OECD and GPAI, leveraging its role as the co-chair of the National Science and Technology Council.