The Computing Research Association (CRA)'s Computing Community Consortium (CCC) Response to the National Telecommunications and Information Administration (NTIA) Request for Information: Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights

This response is from Computing Research Association (CRA)'s Computing Community Consortium (CCC). CRA is an association of nearly 250 North American computing research organizations, both academic and industrial, and partners from six professional computing societies.

The mission of the CCC, a subcommittee of CRA, is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges. Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations, or of the National Science Foundation, which funds the CCC.

March 27, 2024

Written by: Markus Buehler (Massachusetts Institute of Technology), David Danks (University of California, San Diego), Casey Fiesler (University of Colorado, Boulder), Catherine Gill (Computing Community Consortium), Daniel Lopresti (Lehigh University), and Duncan Watson-Parris (University of California, San Diego).

Question 2: How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?

We believe that most of the risks associated with generative models are minimally exacerbated by making model weights widely available. Generative models are inherently risky, due to their ability to quickly generate enormous amounts of believable content based on user inputs and their almost limitless application areas. We grant that the problems most people envision from generative models being publicly available, such as the public having easy access to tools to help create bioweapons or deepfakes to spread misinformation, could arise from these models being completely open-source. However, similar risks have also been shown to potentially arise from proprietary models, even those that have so-called guardrails or other protections. At the current time, there is little evidence that making the weights widely available creates significant *additional* risk beyond what could already be done with proprietary or closed systems.

That being said, the situation could certainly change in the coming years, so we suggest that NTIA should revisit this question regularly.

Moreover, one could make public only a portion of a model's weights, which might further reduce any increases in marginal risk. Making a portion of the model weights or making the APIs publicly available would not give users full capabilities, but would nonetheless support individuals and organizations who desire access to study how these models work and apply them to new use cases, but without being able to change the underlying model for their own purposes (i.e. removing safeguards). If this were the approach to be taken, we believe additional research would be required to confirm that the release of such open models would in fact be useful to those who wish to study them, while at the same time lowering the potential risks.

One risk that could potentially be aggravated if model weights are made publicly available is data exposure. Open-source models have the potential to be reverse engineered to expose training data. While it might be unlikely that training data would be revealed from reverse engineering publicly available model weights, we are not comfortable saying that it is impossible without formal (mathematical) assurances. NTIA should consider different types of gating mechanisms and safeguards to make reverse engineering models with open weights as difficult as possible.

Given how widely available these foundation models are likely to continue to be for general use, the biggest risks, in our opinion, come from not making weights to representative foundation models openly available.

Question 3: What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

Economic innovation is often flagged as the paramount motivator for increasing the openness of foundation models. However, we believe that democratization of these models is an equally strong, if not more substantial, argument in favor of open models. Closed models, such as most of the proprietary foundational models today, exclude the majority of people beyond large, well-resourced technology companies from building upon or researching these models, keeping the decision-making power in the hands of these few large companies. Given the persistent lack of diversity in tech, this has the potential to drastically limit the types of communities that might be supported and the types of harms that might be uncovered.

For example, closed models might exclude the majority of the research community from studying these technologies, and profit-driven companies may not be incentivized to

conduct certain kinds of research, such as bias audits. Nor are they typically interested in exploring, developing, or supporting use cases for certain users or groups of users (e.g., those concerned about the social impacts of AI). However, in order to create models which are fair and equitable for all users, this kind of research needs to be performed.

Increasing access to these models also allows for more research on increasing representation across all facets of society in these models. Government, civil society, academia, and non-profit organizations all have different motivations and diverse research questions they wish to pursue, and denying these groups access will lead to these research questions going unanswered. As a result, many opportunities, including advances in education, commerce, health, and safety and security will go unexplored. For instance, we can imagine how people in a rural community might adapt an AI system to their own needs in ways that are very different from what might be built for them by a large tech company.

Education of the future workforce is another incredibly important consideration. If students and educators are denied access to these models, then the United States will not have an established, well-educated workforce that is able to operate on existing foundational models and create new models in transparent, privacy preserving ways. It is important that students can explore these models during their education to understand their basic functionality, and to learn how to incorporate ethical considerations in developing new models. If these models are not made open, then the only individuals who will have the expertise to maintain and develop foundational models will be those employed at large tech companies. This means those companies will also be the only organizations with the expertise to train future developers on foundational models. Learning about these models solely from an industry standpoint can result in siloed thinking, and these organizations may overlook a holistic education that access to these models can provide in favor of a more efficient learn-as-needed framework. Industry may also neglect to develop frameworks that incorporate needs or applications that are not necessarily foreseen by large for-profit entities. This could result in future foundation models which lack transparency, privacy-preserving, and ethical frameworks. In order to maintain US leadership in AI and all areas that AI could benefit, and to create foundation models which best serve the nation and its citizens, it is imperative to give the next generation a proper education on dual use foundation models so that informed citizens can make their own decisions.

Establishing a culture of openness can also be as important as regulating these technologies. Creating the expectation that models ought to be open will incentivize large companies to be more transparent in their development, and to consider input

from government and research community members in developing these models. It is naive to expect that we can keep the broad community in the dark as to these models' development in perpetuity. In the same way that the effort to control the spread of printing presses in the 1400s failed, which was fortunate for society as a whole, attempting to hide the functionality and capability of generative models from the public today, by security through obscurity, is myopic[1]. Generative models will continue to be developed, and the community will continue to be affected, in ways that are both increasingly beneficial and detrimental. Increasing access to these models will help the community understand how they function, but more importantly encourage critical thinking and allow members to develop the skills and mindsets necessary for recognizing AI generated content.

Question 5: What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?

We believe that the greatest issue is our current lack of knowledge and need for additional research around foundation models. We thus provide a list of questions and topics below, but we emphasize that many of these questions are unlikely to be asked or answered within industry settings. Rather, they will require the ability of researchers in academia, government, and civil society to be able to investigate and experiment on foundation models.

- To what extent can current or future models be utilized to extrapolate beyond training data and generate new knowledge and/or develop autonomous agency?
- What are the impacts of such intelligent systems on the ecosystem of discovery, creativity, innovation, and so on? E.g., imagine a world in which innovation is possible at the push of a button that may have taken years of effort in the past. A wide swath of academic, industrial and office jobs may be made obsolete, or at least transformed in enormous ways. Will this force a portion of the workforce toward more manual labor for tasks that can not yet, or in the foreseeable future, be automated?
- How can we manage the risk of AI technology being used by our adversaries for innovation or warfare, in ways that challenge established rules of war or commerce?

[1] It is a generally held belief within the cybersecurity research community that the strongest forms of security do not depend on obscurity, but rather assume that nearly every aspect of a system's implementation is known and available to potential adversaries.

- What are societal implications of broad access to superintelligence, for all aspects of life and work? E.g., how does it match with the human pace of thinking? Likely, humans will adapt as we adapted to access to fast travel or the internet, but it seems likely to bring major changes to the way we live.
- How can we manage compute resource needs (training and inference) of very large models, especially considering the importance of broad access to this technology for education, entrepreneurship, and beyond?

A large number of researchers, scientists, scholars, and experts in social issues are poised to start answering such questions, provided they receive the open access they need to the kinds of large foundation models that industry is now exploiting. Our continued success as a society depends on it.