WIKIMEDIA
FOUNDATION

Comments of the Wikimedia Foundation
In response to the
National Telecommunications and Information Administration's
Request for Comments on
Dual Use Foundation Artificial Intelligence Models
With Widely Available Model Weights
27 March 2024

The Wikimedia Foundation appreciates the opportunity to comment on the National Telecommunication and Information Administration's (NTIA) Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights.[1] The Wikimedia Foundation is the nonprofit organization that hosts and supports a number of free, online, collectively-produced resources, including Wikipedia. The Foundation's objective is to help create a world where every human being can share freely in the sum of all knowledge.

Widely available access to information has always been particularly important to that mission; restrictions or limitations on access to information impair the ability to freely share, benefit from, and improve knowledge. The Foundation has used free culture and free and open source software copyright licenses since its inception and is currently the world's largest repository of Creative Commons licensed material. This also places the Wikimedia Foundation and the projects it hosts in an important role with relation to artificial intelligence (AI) systems based on large language models: freely licensed technology, including various machine learning (ML) tools, help to support the quality of the Wikimedia projects and aid volunteer editors in working more effectively and efficiently.[2] For example, virtually all of the content on Wikipedia is created, edited, and verified by volunteers. Volunteers, in conjunction with Foundation staff, also build and maintain automated tools to help them maintain the projects' high quality, such as tools to detect vandalism or assess the relative "health" of articles so that volunteer editors can prioritize their efforts. At the same time, the freely licensed high quality information and media on

[1] Request for Comments, NTIA, 89 FR 14059 (26 Feb 2024) https://www.federalregister.gov/documents/2024/02/26/2024-03763/dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights.

[2] See, e.g., Jonathan T. Morgan, *Designing ethically with AI: How Wikimedia can harness machine learning in a responsible and human-centered way*, Wikimedia Foundation, (18 Jul 2019), https://wikimediafoundation.org/news/2019/07/18/designing-ethically-with-ai-how-wikimedia-can-harness-machine-lear ning-in-a-responsible-and-human-centered-way/; Miriam Redi et al., *Can machine learning uncover Wikipedia's missing 'citation needed' tags?*, Wikimedia Foundation, (3 Apr 2019), https://wikimediafoundation.org/news/2019/04/03/can-machine-learning-uncover-wikipedias-missing-citation-needed-t ags/; Miriam Redi, *How we're using machine learning to visually enrich Wikidata*, Wikimedia Foundation, (14 Mar 2018) https://wikimediafoundation.org/news/2018/03/14/machine-learning-visually-enriching-wikidata/.

Wikipedia and the other Wikimedia projects forms one of the most important bases for training generative AI programs.[3]

We embrace these and other forms of openness as fundamental aspects of open knowledge.[4] In the context of ML and AI, we believe that openness presents many benefits. For example, the transparency inherent to open source ML projects means that the shortcomings of a dataset—i.e., biases, incompleteness, and errors—can be identified and can be addressed. More broadly, open and widely available AI models, along with the necessary infrastructure to deploy them, could be an equalizing force for many jurisdictions around the world by mitigating historical disadvantages in the ability to access, learn from, and use knowledge.

When considering the benefits of openness, we believe that the greatest benefits come when source materials are sufficiently open to enable others to study, research, and modify technology. This level of openness allows people to understand the tools that impact their lives, correct problems caused by the tools in question, and to improve the world around them. These high level principles inform our assessment of the risks and benefits of open source AI, discussed in greater detail in our responses to the NTIA's inquiry, below:

1. How should NTIA define "open" or "widely available" when thinking about foundation models and model weights?

We urge the NTIA to carefully consider the potential consequences of creating an official government definition of the term "open" in the context of AI models. We are concerned that, unless the definition of "open" is phrased to require maximum transparency about AI models and to allow maximum reuse and study of AI models, the definition would set a standard that could negatively impact the free and open source software (FOSS) movement.[5]

Regardless of the wording of these definitions, we observe that the risks of the distribution of models or model weights beyond the original developer are not dependent on either openness or wide availability. Conversely, the benefits of such models lean strongly towards greater openness in our view. A relatively closed model could be leaked. An open model need not be widely available to fall into the hands of bad actors. Even if the likelihood of misuse may increase as tools are made more open or more widely available, closed systems are an incomplete defense.

[3] See Selena Deckelmann, *Wikipedia's value in the age of generative AI*, Wikimedia Foundation, (12 Jul 2023), https://wikimediafoundation.org/news/2023/07/12/wikipedias-value-in-the-age-of-generative-ai/.
[4] Open knowledge, Wikipedia (last accessed 27 Mar 2024) https://en.wikipedia.org/wiki/Open_knowledge.
[5] See *Free and open-source software*, Wikipedia, last accessed 20 March, 2024, https://en.wikipedia.org/wiki/Free_and_open-source_software.

In addition to providing imperfect protection, closed systems can limit the distribution of the benefits of dual use technologies.[6] In comparison, open technologies allow a broader range of people to better understand them and create beneficial uses and modifications to them. Therefore, we suggest that regulatory approaches should support and encourage the development of beneficial uses of open technologies rather than depending on more closed systems to mitigate risks.

a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?

Within the nascent AI field, there are already examples of attempts to develop open models as well as leaks to circumvent corporations' efforts to provide limited access to models.[7] Based on our historical experience hosting platforms dedicated to freely and openly licensed knowledge and software for over two decades, we anticipate that new actors who focus on these kinds of efforts will continue to appear globally: both to deliberately develop and share open models and to find ways to make less open models more widely available.[8]

As both stewards of large, open datasets and developers of ML tools, we offer our own practices as examples. Managed correctly, data sources that are publicly available via open data licenses can maximize transparency and reusability. The transparency inherent to open source projects means that the shortcomings of a dataset—i.e., biases, incompleteness, and errors—are known and can be addressed. On the Wikimedia projects, this means that anyone can participate in identifying and addressing knowledge gaps related to gender, ethnicity, and/or other backgrounds that exist on the platforms. Similarly, anyone can leverage this repository of culture and heritage

[6] Compare the impacts of both patents and trade secrets on public health and human rights: Olga Gurgula, *Strategic Patenting by Pharmaceutical Companies–Should Competition Law Intervene?* IIC Int Rev Ind Prop Copr Law (2020) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7592140/; Allison Durkin & Patricia Ann Sta Maria, et al., *Addressing the Risks That Trade Secrets Pose for Health and Rights*, Health Hum Rights, (2021) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8233014/; Julie E Zink, *When Trade Secrecy Goes Too Far: Public Health and Safety Should Trump Corporate Profits*, Vanderbilt Journal of Entertainment and Technology Law, (2020) https://scholarship.law.vanderbilt.edu/jetlaw/vol20/iss4/4/.

[7] Companies like EleutherAI are building products that attempt to match the capabilities of the most powerful closed systems. See https://www.eleuther.ai/. Meta Platforms Inc. attempted to provide limited access to its LLaMA model, but the model was leaked online within a week. See James Vincent, *Meta's powerful AI language model has leaked online — what happens now?*, The Verge (8 Mar 2023), https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

[8] The model weights for two LLMs have been made available recently: xAI released the model weights and network architecture for its Grok-1 large language model, Benj Edwards, *Elon Musk's xAI releases Grok source and weights, taunting OpenAI*, Ars Technica (18 Mar 2024) https://arstechnica.com/information-technology/2024/03/elon-musks-xai-releases-grok-source-and-weights-taunting-openai/, and Databricks released both the base weights and the fine-tuned weights for its new LLM, DBRX, under an open license, The Mosaic Research Team, *Introducing DBRX, A New State-of-the-Art Open LLM*, Databricks (27 Mar 2024) https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm.

for their own educational purposes like accessing oral histories or teaching endangered languages so that these might live on.[9]

Transparency around datasets also improves understanding about how that data has been tagged and modified to develop AI and ML tools. To support transparency in our role as open source developers of human-centered ML tools, the Foundation creates model cards for ML models hosted on our servers.[10] Model cards are primary documentation about a model, reporting on the reasons why it was made, proper and improper use cases, and model evaluation scores across a variety of conditions, including across different cultural, demographic, and intersectional groups.[11] A similar concept exists for datasets, which would provide an additional layer of transparency when evaluating a model.[12]

c. Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (*e.g.,* 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"

We would agree with an assumption that, once there has been a publicly accessible publication of model weights, they are considered widely available. We would also suggest that the same assumption holds true for making weights available through licensing options: if a distribution model makes it possible for actors to engage in redistribution of model weights without needing to seek permission, the weights could be considered widely available. We note that the nature of the entities or institutions who are granted access to model weights may be an additional factor to consider: a distribution to two entities representing access by thousands of people might be considered differently than a distribution to two entities representing access by ten people.

We offer two other observations: First, accurate distribution metrics may be difficult or impossible to obtain. Even if the original developer can count the number of times model weights were accessed or downloaded, measuring secondary distribution becomes significantly more difficult. This is true even where developers attempt to control access in some way—in the

[9] Amrit Sufi, *How to help save endangered languages in India – a project on oral culture digitization*, Wikimedia Diff (8 Aug 2022), https://diff.wikimedia.org/2022/08/08/how-to-help-save-endangered-languages-in-india-a-project-on-oral-culture-digitization/; Gareth Morlais, *Using technology to promote Welsh Language: Wikipedia*, (7 Aug 2017), https://digitalanddata.blog.gov.wales/2017/08/07/using-technology-to-promote-welsh-language-wikipedia/.

[10] Chris Albon, *apply() Conference 2022 | More ethical machine learning using model cards at Wikimedia*, (31 May 2022), https://www.youtube.com/watch?v=t4GMq7MC7Js.

[11] Margaret Mitchell et al., *Model cards for model reporting*, (14 Jan 2019), https://arxiv.org/abs/1810.03993.

[12] Pushkarna, Zaldivar, and Kjartansson, *Data Cards: Purposeful and Transparent Data Set Documentation for Responsible AI*, (3 Apr 2022), https://arxiv.org/abs/2204.01075.

case of a breach or leak of model data, the original developer will not be able to measure the distribution of the model by third parties.

Second, we note that the reasons for defining "widely available"—beyond the basic obligation to follow instructions set out in the Executive Order—may be more important factors in this determination than any particular threshold. For example, if a purpose of defining "widely available" is to assign different obligations to developers or users of widely available model weights to mitigate the risk of misuse, then NTIA would either need to: a) assess how risk of misuse scales with availability and determine an appropriate distribution threshold; b) assume that the potential harms of misuse are great enough to warrant a very low threshold of availability for the application of such obligations; or, c) assume that obligations aimed at risk mitigation will not deter bad actors intent on misuse, and should focus instead on different approaches to dealing with potential harms. We suggest that the last option—i.e., c)—provides the most benefits overall, and that attempts to limit access to dual use models will not be an effective deterrent against misuses of AI systems.

The NTIA should also consider what factors other than availability may need to be defined or assessed in order to better approach risk mitigation.[13] For example, it may be true that the risk of misuse correlates more strongly with the capabilities of the model than with its availability. If that is the case, then approaches to risk mitigation may be better tied to capability rather than availability. Additionally, we observe that making information about AI systems available for independent research and testing is one of the most reliable ways to determine those systems' capabilities.

d. Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?

In general, we believe that the most benefits flow from the most open modes of access to data, models, and computing power. However, we acknowledge that mitigating risks and preventing harms may become more difficult as AI technology becomes more affordable and readily available to more people. To address this tension, we recommend considering the types of risks associated with access to the basic inputs of most AI systems (i.e., data, models, and compute) as well as the likelihood that restricting access to these inputs will effectively mitigate these risks. For example, as we will discuss below, access to some types of data can increase the risk of privacy harms because of the direct causal link between having information about a person and

[13] See, generally, Kapoor and Bommasani, et al., *On The Societal Impacts of Open Foundation Models*, (27 Feb 2024) https://arxiv.org/pdf/2403.07918.pdf.

being able to cause harm to that person—in many cases, the exposure of private information is what causes the harm.[14]

It is less clear, however, what risks may be associated with access to model weights, so we encourage policymakers to articulate their theories of risk and causation. As an analogy, anyone with access to the internet, or even a local library, can easily learn how to construct any number of harmful devices or compounds. With the help of a computer-assisted design program, they could even produce schematics for producing these harmful materials. However, this knowledge and these capabilities are not self-executing. Typically, other tools and resources are required to implement a harmful act. If minimizing the risks of harm while amplifying the benefits of an emerging technology is a primary goal, then policy interventions must be focused appropriately: Limiting access to information about a technology is rarely the best point of intervention to achieve this goal. Instead, we believe that access to information about technology is the best way to support accountability for that technology, and that this information empowers the researchers, developers, and policymakers who can help to mitigate potential harms.

When considering the benefits of openness, we believe that the greatest benefits come when source materials are sufficiently open to enable others to study, research, and modify technology. This level of openness allows people to understand the tools that impact their lives, correct problems caused by the tools in question, and to improve the world around them. Modes of access that limit a person to simply using a "black box" that hopefully does what they want present a risk of exploitation, discrimination, and stagnation, and likely should not be classified as open at all. A tool's widespread availability or utility to the public should not factor into its designation as open, a category which implies specific benefits related to transparency and reuse.

In some cases, openness may increase the risk for certain types of harm. As we mentioned above and now will explain in more detail, releasing training data that contains sensitive personal information, for instance, would likely present significant privacy and safety risks to the subjects of that training data. In such cases, risk mitigation may require curating datasets or applying other privacy preserving techniques prior to training and publication of models.

Finally, we do not suggest that the government should, in all circumstances, require private AI developers to publish model weights or other AI system documentation. Instead, we encourage the NTIA to refrain from recommending restrictions on the availability of this information and to support regulatory policies that promote openness, especially for AI research and AI systems funded or adopted by the government. Relying on access limitations to mitigate risks may forfeit many of the potential public benefits of dual use AI models. As an alternative, we encourage the

[14] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, (2021) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782222.

NTIA to consider not just the modes of access, but also the types of content being made available as well as the different risks and mitigation methods available for each.[15]

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

With widely available model weights, there is at least the possibility of determining the capabilities and potential performance specifications of an application using that model, as well as of identifying vulnerabilities and designing countermeasures to mitigate harmful uses. With nonpublic model weights, it takes a lot more work to get that kind of valuable information, if you can get it at all, and there is still risk of misuse whether by the developer or a third party who acquires the model weights in one way or another. Closed systems prevent good-faith actors from helping to improve AI systems and mitigate harms, but do not prevent bad actors from finding and exploiting flaws and vulnerabilities.[16]

a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

In general, we believe that the risks associated with widely available model weights are similar to the risks associated with the availability of information enabling the reconstruction of other dual use technologies: the risks of misuse. As with other nascent technologies, the full range of potential uses for foundational models is not yet known. This could make it very difficult to anticipate the function or capabilities of a new AI system that is built from the same inputs as an existing system, but with a different objective (i.e., to circumvent moderation filters).

However, we remain firm in our assertion that open ecosystems provide more public benefits and help to mitigate some risks. We believe this is true because expanding access to information also expands the number of potential good-faith contributors working to identify flaws and improve systems.

[15] For example, as the Executive Order requires in some cases, requiring Infrastructure as a Service (IaaS) providers to verify the identity of and maintain usage records for certain clients who contract for computing services could help to deter bad actors from using commercially available resources to deploy harmful models. White House, *Executive Order 14110*, Sec. 4.2(c) (30 Oct 2023) https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[16] Joseph Lorenzo Hall & Stan Adams, *Taking the Pulse of Hacking: A Risk Basis for Security Research*, Center for Democracy & Technology (10 Apr 2018), https://cdt.org/insights/report-taking-the-pulse-of-hacking-a-risk-basis-for-security-research/; Lillian Ablon, Martin C. Libicki, Andrea M. Abler, *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*, RAND (25 Mar 2014), https://www.rand.org/pubs/research_reports/RR610.html.

b. Could open foundation models reduce equity in rights and safety-impacting AI systems (*e.g.,* healthcare, education, criminal justice, housing, online platforms, etc.)?

Use of ML in safety-impacting systems does present significant risks, especially in situations where existing training data is subject to historical biases or where algorithmic prediction impacts a person's rights or opportunities. However, making the operational information available to more people generally helps to reduce the harms associated with biased models or datasets.

Based on many years of experience with diverse populations contributing to freely and openly licensed knowledge, we believe that following open source guidelines is one of the best ways to address this kind of risk. Allowing people to study how models function means that they can identify potential errors and biases that would impact lives. Allowing people to modify and redistribute such models may also allow them to identify and implement improvements, correct biases and errors, and potentially propose improvements that governments and private actors deploying the models can adopt. At the same time, we recommend that the deployment of safety-impacting ML systems undergo appropriate review processes, which would help prevent the deployment of harmful models.

c. What, if any, risks related to privacy could result from the wide availability of model weights?

It remains uncertain to what extent it is possible to reverse engineer or extract training data details by starting with model weights. Based on recent research, it may be possible for someone to use an open model trained on private data to extract identifiable data about individuals included in a training data set.[17] We note, however, that this kind of manipulation appears to be available in a variety of contexts, even in relatively non-open models, and likely would only be prevented in the limited circumstance in which the deployer of the model tightly controls who can make use of it.[18]

An additional risk is the use of widely available model weights to support a generative AI system capable of reproducing the likenesses of existing individuals. Without appropriate guardrails, these systems could be used to impersonate, blackmail, or cause reputational harm to people. We observe that technology capable of supporting such acts has been available for decades, but note that the sophistication of AI tools and the ability to generate content, quickly, at low cost, and on a large scale may require novel approaches to mitigate the potential harms. However, we suggest

[17] Jeffery Wang and Jason Wang, et al., *Pandora's White-Box: Increased Training Data Leakage in Open LLMs* (2024) https://arxiv.org/abs/2402.17012.

[18] Tiernan Ray, *ChatGPT can leak training data, violate privacy, says Google's DeepMind,* ZDNet (4 Dec 2024) https://www.zdnet.com/article/chatgpt-can-leak-source-data-violate-privacy-says-googles-deepmind/.

that attempting to prevent harmful uses of technology by regulating its availability is unlikely to be an effective method when compared to focusing on deterring and defending against specific harmful uses of the technology.

d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?

i. How do these risks compare to those associated with closed models?

The risk of state or non-state actors using AI to threaten security, public health, rights, democratic processes, and more is perhaps more strongly correlated with factors other than the relative availability of the model: for instance, the model's capabilities or the threat actor's available resources. Consider the following contrast: A closed model that offers sufficient cloud infrastructure resources to generate custom content at scale could empower disinformation campaigns, while a malicious actor with access to an open model but without sufficient cloud infrastructure resources of their own for deployment might struggle to leverage the model's capabilities effectively. We further observe that closed models are also subject to attacks that exploit flaws or vulnerabilities in the system's defenses.[19]

ii. How do these risks compare to those associated with other types of software systems and information resources?

Perhaps the biggest difference between the models underpinning emerging AI systems and other types of software resources is that the future capabilities of and use cases for these new AI tools are relatively unknown. This makes it more difficult to anticipate and plan for potentially harmful uses.

Very large AI models may compound this phenomenon by bringing two potent capabilities together in ways that may present new risks: recognizing incredibly complex patterns, and making predictions based on those patterns, both of them at scale. For example, a bad actor could deploy a system to generate large quantities of disinformation, spam, or fraudulent materials engineered to be especially convincing and compelling for a variety of niche audiences. This is not entirely new—automation of many kinds has existed for years, such as that used maliciously in distributed denial-of-service (DDoS) attacks—nor is the mere generation of manipulative content harmful without an effective distribution mechanism. Nonetheless, the combination of scale and the ability to use generative AI to produce material that appears relatively high-quality

[19] Nicole Clark, *'Grandma exploit' tricks Discord's AI chatbot into breaking its own ethical rules*, The Verge (19 Apr 2023), https://www.polygon.com/23690187/discord-ai-chatbot-clyde-grandma-exploit-chatgpt.

all together presents new challenges. Because the range of applications of these new capabilities is unknown, it is difficult to predict where and how they might be exploited to cause harm.

e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?

One of the benefits of making models widely available is that, in theory, more people across all jurisdictions can access and use them. However, where access to tools like predictive models or the necessary infrastructure to deploy them is limited or unequal, those without the ability to use these powerful tools could face increasingly disadvantaged positions compared to others who do have it. Access to technology has long been a deciding factor in the distribution of wealth and competition for resources. Widely available models, along with the necessary infrastructure to deploy them, could be an equalizing force for many jurisdictions or, at least, assist them in keeping pace.

In addition to the equity considerations of unequal access to AI technology, we note that some governments may be less assertive in regulating to mitigate harms. Some may even encourage and support the use of AI models by non-state actors to harm the governments' rivals, whether through AI-supported disinformation campaigns or other attacks. Despite these risks, controlling model weight availability is not the answer. To reiterate, we believe that making AI technology more open will provide the greatest benefits for humanity, and caution that attempts to mitigate harms through the use of closed systems is unlikely to deter those who wish to use AI to cause harm.

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

In general, we believe that open systems are more likely to deliver benefits to society than closed systems. Although the question above relates specifically to model weights—only one component of a functional AI system—we believe this generalization holds true. These systems are: a) powerful tools, which b) use incredibly complex internal processes, and c) could be applied in an unknown number of use cases. Each of these factors weighs in favor of allowing the public to study, understand, improve, and adapt such tools.

The study of open models helps us to understand their strengths and weaknesses, the extent of their abilities, and the best ways to refine and deploy them. Public understanding of models helps people use them more effectively and appropriately, and to make more informed decisions about the benefits, risks, and regulation of AI systems. The ability to improve models and their applications helps those who wish to use existing systems built on those models as well as those who wish to build new systems. Finally, the ability to adapt and repurpose models to perform different functions or to amplify certain features of a model's performance increases the diversity of uses to which models might be applied, creating new tools beyond those envisioned by the original developers.

a. What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/training in computer science and related fields?

We anticipate significant benefits to researchers, who can help improve our understanding of how deployed tools affect people, institutions, and public interests. We also anticipate significant benefits to smaller communities, who can use open models to research tools, identify likely biases and errors, and modify them for their communities' unique needs where appropriate.

For example, the Foundation uses open source software to develop ML tools in collaboration with volunteer contributors. Such ML tools, in turn, assist volunteer editors in their work to build and maintain Wikipedia and other Wikimedia projects. But it is the volunteers, the humans of Wikipedia, who ensure that ML tools are beneficial to the Wikimedia projects and people. This is in part due to the human-centric design and development process the Foundation uses, which accounts for the needs of the volunteers who will use these tools through a collaborative design process.[20] Furthermore, the benefits flow primarily from volunteers' responsible use of ML tools to augment, not replace, their own human capabilities. Although most of these systems are designed to serve specific purposes in the Wikimedia project environment, they are also transparent and open source, available for anyone to inspect, test, improve, or modify for a new purpose.[21]

In the future, open source dual use AI models could further enhance Wikipedia volunteers' capacity to verify, document, and preserve the world's knowledge, and to make that knowledge available across the spectrum of human languages as well as accessible for persons with disabilities. Although the ML tools that volunteer editors presently use now are simpler than today's most sophisticated AI models, open source technology combined with human-centric practices will continue to bring benefits and stability to this oftentimes volatile technological and regulatory space. Even if the future uses of open source AI systems are hard to predict, empowering people to use these systems for good serves the public interest.

[20] Mediawiki, *Wikimedia Product/Inclusive Product Development* (last visited 25 Mar 2024), https://www.mediawiki.org/wiki/Wikimedia_Product/Inclusive_Product_Development.

[21] Wikimedia Meta-Wiki, *Machine learning models*, (last visited 27 Mar 2024), https://meta.wikimedia.org/wiki/Machine_learning_models#; Mediawiki, https://www.mediawiki.org/wiki/MediaWiki.

Finally, we observe that open models offer more opportunities than just recreating or reusing particular model weights. Publishing model documentation might also reveal things like novel approaches to neural network architecture or optimization techniques that could be applied in many additional contexts.

We encourage the NTIA to recommend policies that support open technology and the public benefits it can enable, and to suggest appropriate limitations and exceptions to policy recommendations that would otherwise inhibit openness in the design, development, and deployment of technology.

b. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?

In addition to the public interest benefits flowing from open source technologies, the technologies themselves benefit from being open to research, testing, and review. We call the NTIA's attention to the demonstrated benefits of independent computer security research and note that, regardless of whether a model's weights are widely available, vulnerabilities in the model might still be identified and exploited to cause harm.[22] To quote renowned software engineer Linus Torvald: "Given enough eyeballs, all bugs are shallow."[23] We believe this axiom holds true for large dual use AI models as well. Researchers trying to help make AI systems more secure, resilient, and resistant to abuse or harmful uses can do some work, even with closed systems, but their contributions grow with increased access to source materials like training data, model weights, source code, and training methodology.

This kind of access and testing is important even for more closed systems: if the system's security is to be an effective barrier to accessing model weights or other system specifications, then researchers also need to be able to test that security, in good faith, without fear of legal retribution.

In either context, whether the functioning of the system or the security that protects it, obscurity is not an adequate defense. More eyeballs, more researchers, more people using and studying the way technology works can improve that technology. Attempting to keep systems closed, especially those with many potentially beneficial applications, will produce fewer benefits than making those systems open, even if there are inherent risks that the same technology might be used to cause harm.

[22] *Nathan Van Buren v. United States*, Brief of Amici Curiae Computer Security Researchers, Electronic Frontier Foundation, Center for Democracy & Technology, Bugcrowd, Rapid7, Scythe, and Tenable in Support of Petitioners, 19-783 (US 2021) https://www.supremecourt.gov/DocketPDF/19/19-783/147214/20200708124706913_19-783%20Amici%20Brief.pdf; Matt Burgess, *The Security Hole at the Heart of ChatGPT and Bing*, Wired, (25 May 2023) https://www.wired.com/story/chatgpt-prompt-injection-attack-security/.

[23] See *Linus's law*, Wikipedia, (last visited 20 Mar 2024), https://en.wikipedia.org/wiki/Linus%27s_law.

Regarding trustworthiness in particular, making technical specifications available may be especially important. Although we believe that people are best served by sources of information created and verified by humans, we understand that the ways humans retrieve and digest information are changing. From that perspective, we think it is especially important to be able to understand, test, and improve the AI systems that may become critical intermediaries between people and knowledge. Or, if those systems prove incapable of consistently providing high-quality information in response to human inquiries, then we suggest that making more information about predictive models available will help people understand the limitations of such technologies, and aid the development of new approaches to accessing and distributing knowledge.

d. How can the diffusion of AI models with widely available weights support the United States' national security interests? How could it interfere with, or further the enjoyment and protection of human rights within and outside of the United States?

In the case of Wikipedia, the deployment of transparent AI models and tools can support human beings in curating and expanding reliable knowledge online and combating disinformation at scale. This supports the general public's right to freedom of expression by keeping reliable information widely accessible, while defending national security by serving as an effective deterrent, watchdog, and mitigation against coordinated disinformation campaigns by malicious actors.

e. How do these benefits change, if at all, when the training data or the associated source code of the model is simultaneously widely available?

We propose that the benefits of openness increase as more information is made available. This may be especially true in the context of AI system model weights, training data, and source code—each is a key component of the composite system, with its own unique impacts on the capabilities, functionality, flaws, and vulnerabilities of the system. Making one of these components accessible is helpful for research and understanding, but making all three available, in addition to the training methodology, greatly improves the depth, nuance, and detail possible for researchers and developers who wish to study an AI system.

We identify four ways that fully open models can improve the benefits of AI systems. First, access to training data, source code, and model weights means that students, researchers, companies, and governments can learn any new or interesting techniques that were used to train the model. This broad awareness can inspire significant and beneficial cross-pollination of ideas. Second, access to datasets means that content producers—like Wikipedia editors—can see how their data is being used in these models. Third, access to datasets and source code allows researchers to identify and communicate biases found in the system. Fourth, access to datasets and the source code used in automated dataset curation makes it easier to determine whether private data was included in the model's training so that privacy remediations can be implemented. As explained in response to Question 2(d), however, where training datasets are known to contain large proportions of sensitive personal data, developers must take care when releasing dataset documentation.

4. Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.

There are four components of a foundation model that allow for complete open source rights (unrestricted access, use, and modification). These are:

1) The source code to run the model,
2) The model weights,
3) The training dataset, and
4) The training methodology, including any source code used to automate aspects of the training.

As noted in response to Question 2, it may not always be appropriate to fully release all four of these components. Making all four types of information available maximizes the benefit in terms of study, testing, and modification. However, fully open systems can also increase privacy risks and risks of other harms. Even so, we note that these risks are not eliminated in fully closed systems and that making system information more available is, on balance, the more socially beneficial option.

Additional components, such as high-capacity computing hardware and reliable energy supplies, are necessary to enable current state-of-the-art systems to serve millions of users or to operate at maximum capacity over a period of time. Lacking this infrastructure will limit a dual use model's potential impacts.

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?

b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?

We support the approach that the Executive Order takes toward addressing these risks, that is to say, by requiring Infrastructure as a Service (IaaS) providers to keep records in order to identify their customers and report acquisitions of computing services above a certain capacity threshold.[24] We believe that this approach strikes the right balance to allow good faith research and development of dual use models while discouraging bad actors and helping enforcers take action against entities seeking to develop or deploy harmful applications.[25] We note, however, that nation-states and very wealthy individuals could acquire computing resources through other channels, without triggering the reporting requirement.

f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

We call the NTIA's attention to the four basic components we identify in our response to Question 4: model weights, training data, source code to run the model, and any source code used to automate aspects of the training.

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?

d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?

We would strongly encourage governments to require open source documentation for any AI research they fund and any AI tool they adopt, leaning toward as much disclosure as is safe to do, so that people can understand how government decisions that affect them are being made by these systems and identify potential risks and problems. Open source acquisition policies also help to ensure that that the government is not entirely beholden to proprietary owners, which

[24] White House, *Executive Order 14110*, Sec. 4.2(c) (30 Oct 2023) https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[25] As a matter of law, it is unclear whether the compelled disclosure of such customer records would be limited by statutes such as 18 U.S.C. 2702.

presents its own kind of risk. Consistent with our response to Question 1, we encourage the NTIA to maintain its high standards for what qualifies as "open" in the context of AI models.[26]

e. What should the role of model hosting services (*e.g.*, HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?

The roles and responsibilities of hosting services should not be defined by government entities. These services fit the definition of "interactive computer services" under Section 230 of the 1996 Communications Decency Act (CDA), and should remain free to determine their own policies, standards, and practices for moderating the content that they host.[27] We understand that the leading hosting services already employ many methods to assess the safety and security of the models uploaded to their platforms, and note that even an array of robust screening measures provides an incomplete defense against potentially malicious material.[28] Although there are risks related to the hosting of third-party content, we contend that these risks are outweighed by the benefits of facilitating the sharing of models, datasets, source code, and other components of AI systems.

f. Should there be different standards for government as opposed to private industry when it comes to sharing model weights of open foundation models or contracting with companies who use them?

Yes, the government should insist on maximum disclosure of information about models and applications that are provided to the government by private contractors or developed using public funding. Models developed or funded by the government should be completely open by default. Private industry should remain free to make information about their products available as they see fit. However, there may be some circumstances in which public interest dictates the mandatory disclosure of enough information about privately developed technologies to allow public research for accountability purposes.

[26] The definition used by the US Department of Commerce includes the correct elements, but may benefit from additional clarity regarding the non-software components of an AI system. United States Department of Commerce, *Open Source Code*, https://www.commerce.gov/about/policies/source-code.

[27] 47 U.S.C. §230.

[28] Bill Toulas, *Malicious AI models on Hugging Face backdoor users' machines*, Bleeping Computer (28 Feb 2024), https://www.bleepingcomputer.com/news/security/malicious-ai-models-on-hugging-face-backdoor-users-machines/.