GitHub

88 Colin P Kelly Jr Street
San Francisco, CA 94107
United States of America

March 27, 2024
Docket: NTIA–2023–0009

GitHub Response to NTIA Request for Comment on "Dual-Use Foundation Artificial Intelligence Models With Widely Available Model Weights"

GitHub welcomes the National Telecommunications and Information Administration (NTIA) consultation on widely available model weights as an important step to gather diverse perspectives and empirical evidence to inform the implementation of Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. This consultation can support the collection of information today, and the creation of a framework for future assessment, to determine whether policy measures may be needed.

GitHub is the home of open source collaboration globally, with more than 100 million developers on our platform building and sharing components at every level of the AI stack. Below GitHub contributes expertise and open source community perspective in responses to specific RFC questions. At the outset, we offer several principles to inform NTIA's evaluation of the risks, benefits, and policy measures for widely available model weights.

Open source is a public good. Open source software is a non-rivalrous and non-excludable knowledge base, enabling use and contribution by professional developers, hobbyists, companies, non-profits, and governments alike. This public good has created immense economic value[1] and has been supported by policymakers around the world for its benefits to digital modernization, local industry, and cost savings.[2] The U.S. government and other stakeholders increasingly recognize the importance of open source as public infrastructure and the need for public support.[3] Open source and open

[1] Manuel Hoffmann, et al., "The Value of Open Source Software," *Social Science Research Network*, January 1, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4693148; Gizem Korkmaz et al., "From GitHub to GDP: A Framework For Measuring Open Source Software Innovation," *Research Policy* 53, no. 3 (April 1, 2024): 104954, https://www.sciencedirect.com/science/article/pii/S0048733324000039.

[2] CSIS analysis of 669 policies globally between 1999 and 2022 finds that stated objectives of policies mentioning open source sought to modernize IT (43%), support local industry (20%), and decrease costs (18%). Georgia Wood and Eugenia Lostri, "Government's Role in Promoting Open Source Software." *CSIS,* January 9, 2023, https://www.csis.org/analysis/governments-role-promoting-open-source-software.

[3] The Digital Public Goods Alliance, the G20 Digital Public Infrastructure framework, U.S. Open Technology Fund, and the German Sovereign Tech Fund are relevant examples. In a recent request for information, five U.S. government departments, led by the Office of the National Cyber

science have been essential to AI development to date.[4] Widely available model weights, particularly those made available under open source licenses, provide similar benefits.

AI models are a general purpose technology. Like software, AI and specifically foundation models are general purpose, with "dual-use" deriving from application-specific properties. Executive Order 14110 defines as "dual-use" an AI model with "at least tens of billions of parameters, is applicable across a wide range of contexts, and that exhibits or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters." Clarity on which tasks and levels of performance are to be invoked under the "dual-use" definition will be essential for open source developers and the ecosystem at large.

Evaluation and regulation are better focused on the AI system, rather than the model. Models alone do not determine AI system performance on tasks,[5] including tasks with national security implications. Context of deployment, system affordances including tool-use and safety modifications, and user intent all warrant consideration. To adequately evaluate the benefits and risks of widely available model weights, a broader focus is needed to holistically consider systems using AI models, including proprietary provision, and their integration with other software innovations.

Widely available model weights support research and safety. AI researchers have credited widely available model weights with advancing the

Director, stated "The federal government recognizes the immense benefits of open-source software, which enables software development at an incredible pace and fosters significant innovation and collaboration. In light of these factors, as well as the status of open-source software as a free public good, it may be appropriate to make open-source software a national public priority to help ensure the security, sustainability, and health of the open-source software ecosystem."

[4] Mike Linksvayer, "OSI's Deep Dive Is an Essential Discussion on the Future of AI and Open Source," *Open Source Initiative,* September 29, 2022, https://opensource.org/blog/osi-leading-an-essential-discussion-on-the-future-of-ai-and-open-source; Nathan Benaich and Alex Chalmers, "The Case for Open Source AI," *Air Street* Press, February 8, 2024, https://press.airstreet.com/p/the-case-for-open-source-ai; Bureau of Competition and Office of Technology, "Generative AI raises competition concerns," *Federal Trade Commission,* June 29, 2023, https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.

[5] Matei Zaharia, et al., "The Shift from Models to Compound AI Systems." *Berkeley Artificial Intelligence Research*, February 18, 2024, https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/; Tom Davidson, et al., "AI Capabilities Can Be Significantly Improved Without Extensive Retraining," *arXiv*, https://arxiv.org/pdf/2312.07413.pdf; Megan Kinniment, et al., "Evaluating Large-Model Agents on Realistic Autonomous Tasks," *arXiv*, Janurary 4, 2024, https://arxiv.org/pdf/2312.11671.pdf.

interpretability, safety, and security of AI models.[6] As AI models prove useful to research across academic disciplines, direct access to model weights permits peer review and reproducibility. More investment is needed in AI interpretability, evaluation, and safety research. Across these research directions, norms of open science and open source can accelerate needed discoveries.

Wide availability of model weights is a function of discovery, governed by online platforms. Even for content posted publicly on the internet, the default state is obscurity. Whether content is widely available will depend on ecosystem activity, distribution channels, and, particularly, sharing on platforms that enable virality. Ecosystem monitoring and governance can help inform and implement risk-based mitigations for widely available model weights.

Regulatory risk assessment should weigh empirical evidence of possible harm against the benefits of widely available model weights. Evidence of harmful capabilities in widely available model weights and their use should consider baselines of closed, proprietary AI capabilities and the availability of potentially dangerous information in books and via internet search.[7] The US AI Safety Institute (AISI), companies undertaking large-scale AI research governed by Executive Order 14110 (4.2), and research community efforts may yield such evidence in the future. Today, available evidence of the marginal risks of open release does not substantiate government restrictions.

Government should invest in societal resilience amid growing risks and benefits of AI. The development and deployment of AI systems—regardless of whether their model weights are made widely available—will have profound effects on society. Societal resilience to new developments and use of AI systems for unforeseen purposes warrants attention. The diffusion and diversity of widely available models across society supports public education and incentives for protective measures, ultimately increasing resilience to risks posed by malicious use of AI systems. Policymakers should support societal resilience, including funding research in AI evaluation and measurement science, consolidating best practices via the AISI Consortium, and demonstrating leadership in defensive use of AI, including as directed in EO

[6] Andrew Critch, "My followers might hate this idea, but I have to say it: There's a bunch of excellent LLM interpretability work coming out from AI safety folks (links below...," *X*, October 4, 2023, https://twitter.com/AndrewCritchPhD/status/1709690861003694418; Beren Millidge, "Open Source AI Has Been Vital For Alignment," *Beren's Blog*, November, 5, 2023, https://www.beren.io/2023-11-05-Open-source-AI-has-been-vital-for-alignment/.

[7] Sayash Kapoor, et al., "On the Societal Impact of Open Foundation Models," *CRFM Stanford*, February 27, 2024, https://crfm.stanford.edu/open-fms/paper.pdf.

14110. Policymakers should prioritize AI regulation against reckless use that causes harm today and evaluate criminal justice policies and emergency plans for malicious use.

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

Summary and recommendations:

1. *Government should support community-led definitions of openly available AI models.*
2. *Policy should acknowledge that availability of model weights is not binary and that discoverability plays an important role in model availability.*

We define "available model weights" to be when an AI model has been shared publicly such that developers can have direct access to its trained parameters. With this direct access, developers have the capacity to run or modify the model to suit their purposes. The "wide availability" of AI models is not a foregone conclusion of public sharing; rather, models must be discovered.[8] Discovery depends on ecosystem activity, distribution channels, and, particularly, viral sharing.[9]

In contrast to available models, open source models should reflect the open source definition maintained by the Open Source Initiative. To be "open source," a model must be released under licensing terms that permit anyone to read, modify, (re)distribute, and use the model for any purpose. The Open Source Initiative–the organization that has stewarded the definition of open source for twenty-five years–has ongoing definitional work in adapting these principles to AI models specifically,[10] and other community stakeholders have published perspectives on defining model openness.[11] Government should support community-led definitions, and for the purposes of this RFC, acknowledge that, although widely available model weights pose similar risks, the subset of such models that are available under open source terms that

[8] Although available, the default state of content posted to the public internet is obscurity.

[9] What ought to constitute "widely available" cannot be easily boiled down to a single metric. Rather, it must reflect the risks and benefits with context for use, measured by developer dependencies via online platforms like GitHub, discussion and links on social media, and integration into popular applications. See Question 7 for recommendations on ecosystem monitoring.

[10] Open Source Initiative, "Join the Discussion on Open Source AI," https://opensource.org/deepdive; Open Source Initiative, "The Open Source AI Definition - draft v. 0.0.6," https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-6.

[11] Matt White, et al., "The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI," *arXiv,* March 21, 2024, https://arxiv.org/abs/2403.13784; Heather Meeker, "Toward an Open Weights Definition," *Copyleft Currents*, June 8, 2023, https://heathermeeker.com/2023/06/08/toward-an-open-weights-definition/.

permit lawful use, modification, and redistribution provide particular benefits. Below we refer to such models as “openly available.”

To further define terms, in this submission we use “model developer” to mean those who train an AI model and decide how to make it available. More broadly, we use the term “developer” to refer to those who write software more generally and decide how to make it available, including GitHub users. Developers may be model developers, or may integrate proprietary models-as-a-service or available models into AI systems. Developers may be a company or non-profit, a loose collection of individuals, or an individual. Developers may be, but are not necessarily, the user of an AI system, subject to AI system outputs, or an AI provider who runs system inference for users. The model developer may build in the open, with public access to training and intermittently posting model check-points,[12] or build privately to later share the model publicly. Models are publicly shared today via online platforms including GitHub and via decentralized file-sharing protocols. Models are often discovered by downstream developers via online platforms, particularly those that enable viral sharing.

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

Summary and recommendations:

1. *Reckless-use risks that see harm caused today warrant priority from policymakers.*
2. *Evidence-based, harm-specific analyses of malicious-use scenarios are warranted.*
3. *Societal resilience against malicious-use risk requires a harm-reduction perspective, not an attempt at security through obscurity.*

AI systems of all kinds pose risks that can be categorized in one of two ways: reckless use or malicious use.[13] In contrast to use-based risks from AI systems, risks posed by models reflect developer decisions along the value chain, including model developers’ trusted builds and application developers’

[12] For example, EleutherAI’s use of Weights and Biases for GPT NeoX 20B and the TinyLlama project, respectively.

[13]

Malicious Use	Reckless Use
Deception (fraud, misinformation, persuasion) Hacking Terrorism (designing weapons) Harassment (deepfakes, spam)	Exploited vulnerabilities (prompt injection, data leakage) Bias (flawed decisions, discrimination, representational harm) Accidents (inappropriate or dangerous deployment)

responsible integrations.[14] In evaluating both development and system-integrated use, the risks posed by widely available model weights should be considered marginally, with respect to risks posed by other AI models, and not counterfactually as if no AI capability exists.[15]

We should prioritize focus on reckless-use risks that cause harm today. Policymaking globally has focused on this challenge, particularly in high-risk settings and irrespective of open or proprietary provision.[16] U.S. policymakers should take note of this global trend. Ultimately, societal resilience against malicious-use risk requires a harm-reduction perspective, not an attempt at security through obscurity. The diffusion and diversity of models across society supports public education and incentives for protective measures, ultimately increasing resilience to risks posed by malicious use of AI systems.

In reckless-use scenarios, widely available model weights pose no marginal risk of additional harm relative to closed models, and may instead provide benefits. Reckless deployment or use of a model-integrated AI system may harm the user or those subject to the outputs of the system. In such cases, the system was trusted when it should not have been. Such misplaced trust may include model vulnerabilities due to model developer's lack of awareness or malice, or include poorly governed deployment that sees a system misused, for example, in cases that do not adequately reflect the training data.[17] Widely available models with greater openness, particularly by including open source code, model and data documentation, and/or open data, can reduce risk of reckless use, as downstream developers and users have better information to build applications and select (or contest) use cases (See Question 4). In building and using AI systems, developers must trust or have other assurance in the model developer (or in the proprietary models-as-a-service provider). Trust in the value chain is a common problem in software generally, where developers write software that makes calls to software packages produced by others. Solutions include verified software builds,[18] and early work is ongoing for such approaches specific to AI models.[19] Regardless of whether the model

[14] In some cases, malicious actors may develop applications for their own malicious use.
[15] Sayash Kapoor, et al., "On the Societal Impact of Open Foundation Models," CRFM Stanford, February 27, 2024, https://crfm.stanford.edu/open-fms/paper.pdf.
[16] Peter Cihon, "How to Get AI Regulation Right for Open Source," *GitHub Blog*, July 26, 2023, https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/.
[17] I.e., out-of-distribution use: Jingkang Yang, et al., "Generalized Out-of-Distribution Detection: A Survey," *arXiv*, Januray 23, 2024, https://arxiv.org/abs/2110.11334.
[18] Notable projects include SigStore, Reproducible Builds, and Boostrappable Builds. See also Brian Dehamer and Philip Harrison, "Introducing npm Package Provenance," *GitHub Blog*, April 19, 2023, https://github.blog/2023-04-19-introducing-npm-package-provenance/.
[19] Mithril Security, "AI Cert: Open-source tool to trace AI model's provenance," https://www.mithrilsecurity.io/aicert; Tobin South, et al., "Verifiable Evaluations of Machine Learning Models Using zkSNARKS," *arXiv*, February 5, 2024, https://arxiv.org/pdf/2402.02675.pdf.

is widely available or proprietary, if it contains vulnerabilities, a downstream user may experience security risks, with implications for privacy and other harms.

Malicious use of widely available models poses marginal risks of additional harm. Once a model developer releases the model, they cannot fully determine how it will be used downstream. Developers may integrate widely available models into AI systems without features commonly (but not necessarily) found in proprietary systems, including prompt filters and monitoring mechanisms. Malicious users, in some cases possibly stymied and/or detected by these features, may gravitate towards lax AI systems providers or, if they have the capability, to develop applications with widely available models. In some cases, however, features common in proprietary AI systems face challenges in detecting or preventing malicious use, including in the generation of misinformation and software code intended for malicious ends. Thus, evidence-based, harm-specific analyses of malicious-use scenarios are warranted.

Consider malicious actors in three categories: state actors, non-state actors, and individuals. State actors have the capability to recreate powerful foundation models from scratch today.[20] The demonstration of closed capabilities, absent any architectural detail, may well be sufficient to stimulate free discussion of possible methods for achieving said capabilities and enable well-resourced actors including state actors to create similar models.[21]

Non-state actors and individuals may not be able to recreate such models directly, and thus widely available models may be counterfactually useful. Evidence-based, harm-specific marginal risk assessments should consider the full chain of malicious actions required for these actors to do harm.[22] AI systems that complement models with other components can demonstrate greater capabilities along the malicious activity chain and warrant evaluation in their own right (See Question 5).

In practice, individuals using AI for their own ends, including possible malicious use enabled by widely available models, supports societal adaptation and resilience. The history of open source software suggests that

[20] For example, government-supported efforts in the UAE and China have trained and publicly shared 100+ billion parameter models.

[21] On idea hazards, see Nick Bostrom, “Information Hazards: A Typology of Potential Harms from Knowledge,” *Review of Contemporary Philosophy*, Vol. 10, 2011, https://nickbostrom.com/information-hazards.pdf.

[22] Information provision or knowledge creation that may be facilitated by AI models does not cause harm directly. The malicious activity chain may be shorter or longer by particular harm, e.g., cybersecurity does not face a cyber/physical barrier in the way that weapons manufacture does.

individual (malicious) use will be larger in number and of high variance in effort and effectiveness, in contrast to those from highly motivated and resourced actors. State actors have a harder time hoarding vulnerabilities to create targeted software attacks, for example, when numerous security researchers identify vulnerabilities and lead maintainers to issue patches. Similar arguments extend to epistemic security[23] and other malicious-use scenarios. Societal resilience requires a harm-reduction perspective, not an attempt at security through obscurity.

Imposing restrictions on widely available model weights in one jurisdiction may yield unintended or counterproductive effects. Restrictions on release will not prevent malicious use of models developed or released elsewhere. Cybercrime and cyber-enabled fraud, for example, are widely recognized as transnational, with the malicious actors located in different jurisdictions than their victims.[24] Restrictions will, however, limit the lawful use of such models, harming national economic competitiveness and additional benefits outlined below in Question 3. Furthermore, restrictions may undermine societal resilience over time: diffusion of models across society supports public education and incentives for protective measures, ultimately increasing resilience to risks posed by malicious use of AI systems.

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

Summary and recommendations:

1. *Although widely available model weights pose similar risks, the subset of models that are available open source bring additional benefits.*
2. *These benefits include innovation, market competition, and diffusion of AI across the economy; support for AI development and safety; use of AI in research across disciplines; developer education; and government use.*

Widely available AI models, specifically those that are available under open source licenses, present notable benefits for U.S. economic dynamism, AI safety, and human rights. We use “openly available” to refer to these models in particular.

Openly available AI models support innovation, market competition, and diffusion of AI across the economy. The wide availability of models may

[23] Elizabeth Seger, et al., “Tackling Threats to Informed Decision-Making in Democratic Societies,” *The Alan Turing Institute*, October, 2020, https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf.

[24] Isabella Wilkinson, “What Is the UN Cybercrime Treaty and Why Does it Matter?” *Chatham House*, August 2, 2023, https://www.chathamhouse.org/2023/08/what-un-cybercrime-treaty-and-why-does-it-matter.

commoditize particular AI capabilities, driving down the cost associated with running AI models within an application as proprietary models-as-a-service face increased competition. Openly available models present additional options to organizations of all kinds as they evaluate a build-buy decision, enabling choices that separate model provider from infrastructure host, permit the fine-tuning or other direct modification of the model, and, if the training code and sufficient detail of the training data is provided, re-train a model from scratch. Open source and widely available models have enabled extensibility innovations that reduce the hardware required to run inference, and enable further training on private or otherwise sensitive data that may not be shared with third-parties. The result means more competition in the AI market.[25] It also supports the diffusion of AI into all sectors, including regulated industries, government use, and niche cases for which markets may not adequately provide.

Open source and widely available AI models support research on AI development and safety, as well as the use of AI tools in research across disciplines. To-date, researchers have credited these models with supporting work to advance the interpretability, safety, and security of AI models[26]; to advance the efficiency of AI models enabling them to use less resources and run on more accessible hardware[27]; and to advance participatory, community-based ways of building and governing AI.[28] Various kinds of AI models have been identified as holding promise to advance scientific research as well as academic scholarship broadly.[29] In order for such research to be reproducible, models and software used must be accessible to scholars and access must be assured over time.[30]

[25] Bureau of Competition and Office of Technology, “Generative AI raises competition concerns,” *Federal Trade Commission,* June 29, 2023, https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.

[26] Andrew Critch, “My followers might hate this idea, but I have to say it: There's a bunch of excellent LLM interpretability work coming out from AI safety folks (links below...,” *X*, October 4, 2023, https://twitter.com/AndrewCritchPhD/status/1709690861003694418; Beren Millidge, “Open Source AI Has Been Vital For Alignment,” *Beren’s Blog*, November, 5, 2023, https://www.beren.io/2023-11-05-Open-source-AI-has-been-vital-for-alignment/.

[27] E.g., Tim Dettmers, et al., “QLoRA: Efficient Finetuning of Quantized LLMs,” *arXiv*, May 223, 2023, https://arxiv.org/abs/2305.14314 and its associated GitHub repository.

[28] E.g., the BigScience Project.

[29] OECD, “Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research,” 2023, https://www.oecd.org/publications/artificial-intelligence-in-science-a8d820bd-en.htm; Anton Korinek, “Language Models and Cognitive Automation for Economic Research,” *NBER*, February, 2023, https://www.nber.org/papers/w30957.

[30] Sayash Kapoor and Arvind Narayanan, “OpenAI’s Policies Hinder Reproducible Research on Language Models,” *AI Snake Oil*, March 22, 2023, https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible.

Openly available AI models and the open source communities that build, maintain, and extend them support developer education. Open source supports self-learning, lowers intellectual property barriers to education of all kinds, and enables learning-by-doing as developers make direct contributions to projects and communities. An expanded developer base, particularly outside of a small set of companies located in a few major tech hubs, supports diversity of identity and perspective in the ecosystem. The expanded developer base also means that departments at all levels of government can have an easier time locating talent to build regulatory capacity.

Openly available AI models can support government use of the technology. AI systems hold promise for innumerable public-interest applications, and some scholars have called for government investment in a public, open option.[31] The Federal government has an open source code policy dating back to 2016 that seeks to increase the use of open source software in custom-developed software projects for the government.[32] Openly available models can support this policy and its objectives of enabling software reuse and in doing so reducing costs to the American taxpayer. Additionally, openly available models can be further modified on sensitive data that may not be able to be provided outside of government.

Openly available AI models can support U.S. foreign policy goals, building on the track record of open source software.[33] Openly available models present challenges for the censorship activities of foreign adversaries, as they can be shared outside the context of web domains that may be effectively restricted. Such models could provide consolidated knowledge repositories to further people's right to information around the world under Article 19 of the Universal Declaration of Human Rights.[34] The innovation benefits of open models noted above also support U.S. national competitiveness.

[31] Bruce Schneier, "Build AI by the People, for the People," *Foreign Policy*, June 12, 2023, https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance/.
[32] Department of Commerce, *Source Code Policy*, https://www.commerce.gov/about/policies/source-code.
[33] Consider, for example, the Open Technology Fund and its history.
[34] United Nations, *Universal Declaration of Human Rights*, https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Question 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.***

Summary and recommendations:

1. *Two categories of additional components warrant attention: (1) components used to create and document model weights and (2) components of AI systems, of which models are but one.*
2. *Greater openness in model components supports benefits and reduces risks.*

Additional components are useful to understand and recreate widely available model weights. Training an AI model relies on data, software code, and compute, and may be described in complementary components including shared notebooks or an academic paper. Models and their training data may be documented in model cards and datasheets for datasets, among other methods. As outlined in Question 1, there are multiple community perspectives on levels of openness.[35] Providing additional components, including the training code as open source software, documentation, and/or open data, encourages the benefits of widely available model weights. With greater openness, developers can more readily scrutinize, modify, or retrain the model, supporting economic dynamism and research. Greater openness reduces risks from reckless use and promotes human rights, as AI system providers can make better informed decisions on use cases, users make better-informed adoption decisions, and advocates or those subject to system outputs can better contest use cases.

Models are but one component of an AI system. To realize any benefit or risk, models must be put into service via an AI system. Additional software components in an AI system impact both risks and benefits posed by its use, regardless of whether the model is openly available or proprietary. Safety filters on prompts and outputs are common practice today. Orchestration or scaffolding frameworks are an increasing focus of research and development. Many of these safety and capability features are shared as open source software**.**[36] Evaluating models, or indeed other components including software

[35] Open Source Initiative, "The Open Source AI Definition - draft v. 0.0.6," https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-6; Matt White, et al., "The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI," *arXiv,* March 21, 2024, https://arxiv.org/abs/2403.13784; Heather Meeker, "Toward an Open Weights Definition," *Copyleft Currents*, June 8, 2023, https://heathermeeker.com/2023/06/08/toward-an-open-weights-definition/.

[36] Task Force for a Trustworthy Future Web, "Annex 2: Scaling Trust on the Web: Building Open Trust and Safety Tools," *Atlantic Council*, June, 2023, https://www.atlanticcouncil.org/in-depth-research-reports/report/scaling-trust_annex2/; see GitHub projects under agents and autonomous-agents topics.

code, in isolation risks missing risks and benefits posed by new categories of AI systems.

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

Summary and recommendations:

1. *AISI and other stakeholders should support openly available evaluation suites to enable wider testing of dual-use risks.*
2. *Evaluations should assess models not in isolation, but as integrated into AI systems.*
3. *Stakeholders including NSF, NIST, AISI, and OSTP should support needed research directions.*

To manage the risks and amplify benefits of AI models, including those with widely available model weights, we need better evaluations. Leading AI labs are investing in evaluation science and proprietary evaluations. Although open evaluation suites are increasingly available for some benchmarks, these are primarily for capability measures. Methods of evaluating dual-use risks that motivate government concern are not openly available today. To support broader use of evaluations in the community, we need openly available evaluation suites.

The performance of AI models can be altered after training, via direct modification, including fine-tuning, and by integration into AI systems, including with orchestration software that supports tool use.[37] While direct modification requires access to model weights, closed AI models-as-a-service as well as widely available models can be integrated into broader systems to serve specific ends. In practice, this raises the need for system evaluations—not simply model-level evaluations—and offers a warning against focusing too narrowly on fine-tuning away safeguards as the risk to prevent.

More research is needed on numerous fronts. Below we outline several important directions. Across these research directions, norms of open science and open source can accelerate needed discoveries.

[37] Matei Zaharia, et al., “The Shift from Models to Compound AI Systems.” *Berkeley Artificial Intelligence Research*, February 18, 2024, https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/; Tom Davidson, et al., “AI Capabilities Can Be Significantly Improved Without Extensive Retraining,” *arXiv*, https://arxiv.org/pdf/2312.07413.pdf; Megan Kinniment, et al., “Evaluating Large-Model Agents on Realistic Autonomous Tasks,” *arXiv*, Janurary 4, 2024, https://arxiv.org/pdf/2312.11671.pdf.

- Interpretability science is nascent and cannot explain how large foundation models produce specific outputs.[38]
- There is some indication that so-called emergent capabilities are a function of discontinuous evaluation benchmarks, pointing to the need for better, continuous capability benchmarks.[39]
- Evaluations of models and systems for malicious use risks are understudied relative to capability benchmarks and specific risks including loss of control and persuasion.[40] Evaluations of autonomous capabilities of systems is even more nascent.[41]
- More research for data governance to reduce malicious use risks is warranted; at least one prominent paper points to the promise of restricting data leading to safer models.[42]
- Methods of restricting downstream modification of model weights to remove protections should be further explored.[43]
- Methods of assurance for model builds, their provenance, and evaluation outcomes should be further developed.[44]

Question 6: *What are the legal or business issues or effects related to open foundation models?*

Summary and recommendations:

1. *Open source software provides a useful analogy for the ecosystem that may emerge with widely available models.*
2. *License terms that reduce friction to sharing have enabled wide reach and societal benefit from open source software.*

Widely available AI models, specifically those that are available open source ("openly available"), present opportunities for integration and use across the

[38] Neel Nanda, et al., "Progress Measures for Grokking via Mechanistic Interpretability," *arXiv*, October 19, 2023, https://arxiv.org/abs/2301.05217; Steven Bills, et al. "Language Models Can Explain Neurons in Language Models," *OpenAI*, May 9, 2023, https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html; Anthropic, "Transformer Circuits Thread," March 2024, https://transformer-circuits.pub/.

[39] Rylan Schaeffer, et al., "Are Emergent Abilities of Large Language Models a Mirage?" *arXiv*, May 22, 2023, https://arxiv.org/abs/2304.15004.

[40] Megan Kinniment, et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," *arXiv*, January 4, 2024, https://arxiv.org/abs/2312.11671; Mary Phuong, et al., "Evaluating Frontier Models for Dangerous Capabilities," *arXiv*, March 20, 2023, https://arxiv.org/abs/2403.13793.

[41] See METR's Autonomy Evaluation Resources.

[42] Lukas Berglund, et al., "The Reversal Curse: LLMs trained on 'A is B' fail to learn 'B is A'," *arXiv*, September 22, 2023, https://arxiv.org/abs/2309.12288.

[43] Peter Henderson, et al., "Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models," *arXiv*, August 9, 2023, https://arxiv.org/abs/2211.14946.

[44] Mithril Security, "AI Cert: Open-source tool to trace AI model's provenance," https://www.mithrilsecurity.io/aicert; Tobin South, et al., "Verifiable Evaluations of Machine Learning Models Using zkSNARKS," *arXiv*, February 5, 2024, https://arxiv.org/pdf/2402.02675.pdf.

economy. Open source software provides a useful analogy to understand how the ecosystem may evolve and how industries and governments alike may adopt openly available AI components. Licensing restrictions that contravene the open source definition may limit these benefits, and restrictive licensing terms present enforcement challenges in practice. Ultimately, public policy rather than copyright licenses will govern the responsible use of AI systems.

Open source software provides analogies and lessons for the emerging ecosystem of openly available AI models. Once software is written, it can be copied at zero marginal cost, as can open source AI systems. The point of open source software, as well as openly available AI models, is to remove barriers to sharing this zero-marginal-cost good, empowering developers to improve and build upon the software. When software code is compiled into a binary or otherwise packaged for distribution to end users in a form that can be executed on appropriate hardware, non-developers gain access to the software's functionality. Likewise, once a model is trained, developers can use pre-trained model weights to easily gain access to the model's capabilities by running inference on appropriate hardware. In this way, model weights are akin to a compiled software library.

Open source is wildly successful. Today, 96% of software contains open source components, and a given software stack is 77% open source software.[45] It is widely used across government and industries,[46] with 99% of the Fortune 500 using open source software,[47] and open source adoption is a key innovation differentiator between firms.[48] However, the broader societal benefits of this ecosystem are challenging to measure.[49] Researchers from the Bureau of Economic Analysis, National Science Foundation (NSF), and elsewhere have estimated that investment in open source development contributes roughly $38 billion to U.S. GDP.[50] However, GDP measures expenditures, and thus does not account for the unique benefits from open

[45] Synopsys, "2024 Open Source Security and Risk Analysis Report," February, 2024, https://www.synopsys.com/software-integrity/engage/ossra/ossra-report, p.4.

[46] Ibid., p.5; General Services Administration, "Open Source," *Digital.gov*, https://digital.gov/topics/open-source/.

[47] Pranay Ahlawat, et al., "Why You Need an Open Source Software Strategy," *BCG*, April 16, 2021, https://www.bcg.com/publications/2021/open-source-software-strategy-benefits.

[48] Shivam Srivastava, et al., "Developer Velocity: How Software Excellence Fuels Business Performance," *McKinsey & Company*, April 20, 2020, https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/developer-velocity-how-software-excellence-fuels-business-performance.

[49] Peter Cihon, "Open Source Creates Value, But How Do You Measure It?" *GitHub Blog*, January 20, 2022, https://github.blog/2022-01-20-open-source-creates-value-but-how-do-you-measure-it/.

[50] Gizem Korkmaz et al., "From GitHub to GDP: A Framework For Measuring Open Source Software Innovation," *Research Policy* 53, no. 3 (April 1, 2024): 104954, https://www.sciencedirect.com/science/article/pii/S0048733324000039.

source software, namely its frictionless, zero-marginal-cost reuse. One recent study attempted to fill this gap by measuring the demand-side value of open source, and estimated it to be $8.8 trillion.[51]

The open source ecosystem, its wide reach, and large societal benefit, has been enabled by clear licensing that permits anyone to read, modify, (re)distribute, and use the software for any purpose. To enable frictionless sharing, these licenses disclaim liability and warranty for the freely offered software code. Since its founding in 1998, the Open Source Initiative has maintained a list of approved licenses, supporting their widespread understanding and adoption.

Recognizing that a public good should be used for good, some developers have experimented with not-open source "ethical use" restrictions in software licenses. Because these terms are often ambiguous, cause friction, and have conflicts between different "ethical use" terms, software under these terms has not gained widespread adoption.[52] AI researchers have rediscovered this type of ethical licensing with the RAIL family of licenses. However, they face the same challenges: third parties are not well suited and may not be able to directly enforce license terms that restrict use, the terms can conflict with other "ethical use" terms, and developers may not have the resources or motivation to enforce such terms.[53] The aim of these sorts of terms is admirable: they try to mitigate harmful use of technology. However, because of conflicts between terms and potentially differing developer interpretations, it is hard to build a frictionless open innovation ecosystem on these terms. Clarity in law, not ambiguity in license terms, supports a vibrant innovation ecosystem.

Another trend in public licensing is to try to "capture" open source innovation by allowing customers and developers to use your software under open source-like terms, but to forbid the use of your software by competitors.[54] In this way, companies try to "own" the open source ecosystem by capturing the economic upside of the open source innovation cycle. AI developers have followed this "own the ecosystem" suggestion,[55] with licenses on some

[51] Manuel Hoffmann, et al., "The Value of Open Source Software," *Social Science Research Network*, January 1, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4693148.
[52] E.g., Stephen Shankland, "'Don't-be-evil' Google Spurns No-Evil Software," *CNET*, December 28, 2009, https://www.cnet.com/culture/dont-be-evil-google-spurns-no-evil-software/.
[53] E.g., Robert Gorwa and Michael Veale, "Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries," *arXiv*, February 15, 2024, https://arxiv.org/abs/2311.12573.
[54] E.g., Armon Dadgar, "HashiCorp Adopts Business Source License," *HashiCorp*, August 10, 2023, https://www.hashicorp.com/blog/hashicorp-adopts-business-source-license.
[55] Dylan Patel and Afzal Ahmad, "Google 'We Have No Moat, and Neither Does OpenAI'," *Semianalysis*, May 4, 2023, https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

popular models designed to permit community modification while inhibiting certain types of commercial use or use by competitors.[56]

Such ethical and economic upside restrictions do not affect malicious-use proclivities, but do limit the competitive and other benefits of frictionless zero-marginal-cost copying of software and models.[57] For this reason among others, regulation has singled out open source software, in contrast to other software licenses, for special consideration. In recent EU policymaking, this has been seen across the AI Act, Cyber Resilience Act, and Product Liability Directive. With these files, EU policymakers distinguished between the development and supply phases, and focus regulatory scrutiny on supply, particularly monetized or other commercial provision and the high-risk use of open source software and AI systems.

Question 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?***

Summary and recommendations:

1. *Existing best practices for documentation mitigate reckless-use risks and have been codified into EU law.*
2. *An emerging transnational testing regime can support evidence-gathering for malicious-use risks. This can be complemented by ecosystem monitoring.*
3. *Malicious-use risks are a function of the number of possible users. Applications that enable easy inference warrant attention proportionate with their lowering of expertise barriers to misuse.*
4. *GitHub enables developers to share AI models and other software components for further development, in accordance with law and our Acceptable Use Policies.*
5. *Specific government bodies and other stakeholders can take important steps today to provide clarity on openly available AI models, improve evaluation science, and support a trusted AI value chain.*

Developers around the world collaborate on modifying, integrating, and using widely available AI models, particularly those available under open source licenses, in an emerging, decentralized innovation ecosystem that will help maximize the benefits of these technologies. Established best practices for

[56] Technology Innovation Institute, “Terms and Conditions: Falcon 180B TII License Version 1.0,” September 2023, https://falconllm.tii.ae/terms-and-conditions.html; Meta, “Llama 2 Community License,” July 18, 2023, https://ai.meta.com/llama/license/.

[57] Bureau of Competition and Office of Technology, “Generative AI raises competition concerns,” *Federal Trade Commission,* June 29, 2023, https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.

documentation, namely model cards and thorough dataset reporting, reduce risks of reckless use by informing downstream providers and deployers of the capabilities of the model and out-of-scope use cases. Thus informed, downstream providers and deployers can integrate widely available models, particularly those available open source, into AI systems, products, and workflows as they do other open source software components. The forthcoming EU AI Act encourages or requires such documentation, dependent on how the model is deployed and its size.

Today, malicious-use risks are informed by an emerging transnational testing regime, which will prove instrumental in establishing whether there is credible evidence to warrant further policy measures on the open release of AI models, their downstream integration into applications, and criminal use. Given the benefits of open release, absent specific evidence of model-level risks, risks are best addressed through policy focused on integrated AI systems and their use. Reflecting practices of open source innovation and regulation, upstream developers providing software components as a public good should be supported to adopt best practices, while downstream integrators seeking to adapt or otherwise privatize public goods for specific ends should face heightened expectations in safely and responsibly providing products that are directly usable.

The emerging transnational testing regime reflects voluntary commitments from model developers, forthcoming regulatory requirements in the EU AI Act, and community best practices. All three involve pre-release evaluations to detect capabilities of concern, including support for malicious use, which may inform voluntary decisions to restrict the public release of a particular AI model.[58] The U.S. should prioritize coordination with EU and UK policymakers as well as other AI Safety Institutes on information sharing, advancing evaluation infrastructure, and consolidating best practices.

Model evaluations alone are insufficient for malicious-use analysis; as described in Question 5, system-level evaluations are a necessary complement. Ecosystem monitoring is another important complement, to understand discovery mechanisms by which models become widely available. These range from social media platforms with viral sharing, integrated use in applications that are widely adopted (perhaps via app stores), and use of models as dependencies of software projects developed on platforms like

[58] Voluntary commitments from model developers and forthcoming EU AI Act model evaluation requirements focus narrowly on the largest, most capable AI models or those that may otherwise be designated as posing systemic risks. Both Executive Order 14110 (4.2(C)) and the forthcoming EU AI Act require advanced notice of training for such models (Art 52(1)), as well as the documentation of any evaluations on such models and express sharing in the case of the EO. Voluntary commitments support similar evaluation-sharing with the UK government.

GitHub. Additional data from real-world use will also be useful, particularly to understand how closed AI may already be misused. Monitoring AI developments and societal impacts over time can improve risk assessments and policymaking.

While widely available AI model weights may frustrate government restrictions on the availability of AI capabilities to specific entities,[59] it does not prevent restrictions on harmful use of downstream applications and services. Malicious-use risk is, in part, a function of the number of people able to develop and deploy applications that use a model. Current scarcity of machine learning expertise moderates the risk, by narrowing the population of developers who have the knowledge and resources to fine-tune or otherwise modify an openly available AI model towards malicious ends. A larger population of developers have the basic programming skills and resources required to run inference with openly available models. In both cases, however, these populations pale in comparison to the broader population using end-user applications. Model inference applications and services, particularly those that enable deepfake generation—regardless of the type of model they may use—warrant close policy scrutiny proportionate with risks posed by their lowering barriers to (mis)use.

GitHub hosts content that meets standards set forth by law and our Acceptable Use Policies (AUPs). GitHub does not make openly available models directly usable for inference. 100+ million developers around the world use GitHub to collaborate on and share software, including software at every layer of the AI stack and AI models in particular. Content posted to GitHub must be lawful, and is governed by our AUPs.[60] Our AUPs permit dual-use content, supporting its use for research and education. However, in cases of abuse of or malicious intent in such dual-use content, GitHub uses a range of tools to restrict access to the specific content on the platform.[61] GitHub periodically reassesses our policies, and offers developers the opportunity to

[59] Given that publicly available open source software is not subject to export controls. Steve Winslow, et al., “Understanding US Export Controls With Open Source Projects,” *Linux Foundation*, July 2021, https://www.linuxfoundation.org/resources/publications/understanding-us-export-controls-with-open-source-projects.

[60] GitHub’s AUPs prohibit activity and content that is sexually obscene; libelous, defamatory, or fraudulent; discriminatory or abusive; false, inaccurate, or intentionally deceptive and is likely to harm the public interest (including election integrity); harasses or abuses; or threatens violence or glorifies violence. GitHub, “GitHub Acceptable Use Policies,” *GitHub Docs*, https://docs.github.com/en/site-policy/acceptable-use-policies/github-acceptable-use-policies#2-user-safety.

[61] GitHub, “GitHub Active Malware or Exploits,” *GitHub Docs*, https://docs.github.com/en/site-policy/acceptable-use-policies/github-active-malware-or-exploits.

comment on any proposed changes.[62] Our AUPs and dual-use policies would not permit developers to host an openly available AI model that had been fine-tuned or otherwise modified for malicious ends.

Government can take important steps today to provide clarity on openly available AI models, improve evaluation science, and support a trusted AI value chain. To provide clarity, NIST and the AISI should publish and iteratively update guidance on metrics for risk assessment and provide clarity on what categories and performance thresholds may weigh against open release.[63] NTIA and Bureau of Industry and Security (BIS) should align on definitions to ensure that openly available models meet the "published" definition under the Export Administration Regulations to avoid unintended export control impacts. To advance the state of the art and use of AI safety research, NSF should prioritize funding research for AI interpretability, evaluations, and durable model-level safety interventions. NIST and AISI should support open source software evaluation suites to enable all model developers to evaluate models for capabilities of concern prior to their release. AISI and the National Artificial Intelligence Research Resource (NAIRR) should support less-resourced actors to perform evaluations. More broadly, government can support the refinement and incentivize adoption of responsible best practices in the AI value chain. These include model documentation and trusted model builds.[64]

Model developers, academia, civil society, and other stakeholders should lead further refinement of best practices. The Partnership on AI is working to update its Deployment Guidance for Foundation Model Safety[65] to support model developers in making open release decisions, and could similarly provide metrics and clarity on risks. Once released, a model cannot be fully recalled. However, harm can be reduced, by developing procedures to notify the recall of models. Model developers, academia, and philanthropies can accelerate efforts by open sourcing existing and new evaluation suites for

[62] E.g., Mike Hanley, "Updates to Our Policies Regarding Exploits, Malware, and Vulnerability Research," *GitHub Blog*, June 4, 2021, https://github.blog/2021-06-04-updates-to-our-policies-regarding-exploits-malware-and-vulnerability-research/.

[63] As part of ongoing activities directed by EO 14110 (4.1)(i)(C) "launching an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities through which AI could cause harm, such as in the areas of cybersecurity and biosecurity."

[64] See footnotes 18 and 44.

[65] Partnership on AI, "PAI's Guidance for Safe Foundation Model Deployment," https://partnershiponai.org/modeldeployment/.

capabilities of concern.[66] Further steps still may be needed to address costs of evaluation for less resourced actors.[67]

Question 8: ***In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?***

Summary and recommendations:

1. *Government and other stakeholders should make plans today to strengthen societal resilience to increasing AI capabilities.*
2. *Ecosystem monitoring can support detection of novel uses of AI that pose unforeseen risks and benefits.*
3. *Evaluations of new paradigm-shifting AI systems warrant attention prior to wide release.*

Governments, companies, and individuals can make plans today with an understanding that increasing AI capabilities, whether made available in closed or available forms, will present increasing risks and benefits, including ones that may not be foreseen. As such, societal resilience and specifically cybersecurity should be invested in, even more than might be dictated by risks that are anticipated today. Securing the open source ecosystem and promoting secure by design[68] as a principle for all software, including AI, should be accelerated.[69] Government can demonstrate leadership in the defensive use of AI and take further non-AI defensive measures, including those directed in EO 14110.[70]

Open source has promoted competition across the IT industry, giving developers more opportunity to transfer skills and experience across employers, lowering the barriers for entrepreneurs to compete at every layer of the stack, and facilitating national competitiveness. Openly available AI

[66] To-date, leading model developers have open sourced some evaluation suites, but not the very evaluations that are used to justify possible need for restricting openly available AI models.

[67] See, e.g., Nathan Lambert, “Evaluations: Trust, Performance, and Price,” *Interconnects*, March 20, 2024, https://www.interconnects.ai/i/142801100/the-rising-price-of-evaluation.

[68] Mike Linksvayer, “GitHub Response to the Office of the National Cyber Director Request for Information on Open-Source Software Security: Areas of Long-Term Focus and Prioritization,” *Regulations.gov*, November 8, 2023, https://www.regulations.gov/comment/ONCD-2023-0002-0084.

[69] Mike Linksvayer, “GitHub Response to CISA Request for Information on Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software,” *Regulations.gov*, February, 20, 2024, https://www.regulations.gov/comment/CISA-2023-0027-0080.

[70] Including building upon the DHS and DoD pilot projects for detecting and remediating vulnerabilities at scale in critical government software and screening release of sensitive data and customers of nucleic acid synthesis services.

models and other open source elements of the AI stack are continuing and accelerating each of these factors over time.

As innovation in AI models, systems, and use cases continues, stakeholders should invest in better understanding the risks and impacts. Government has a leading role to resource and mature interpretability research, evaluation science, and monitoring capacity, and should support stakeholders from companies, academia, and civil society, to accelerate work in these areas.

Multiple stakeholders should prioritize monitoring the innovation ecosystem for downstream uses of widely available models and uses of models generally that enhance their capabilities following pre-training. GitHub today contributes data on AI-related development activity to the OECD and Stanford AI Index, among other entities, and could endeavor to support AISI, other government, or key multi-stakeholder initiatives. Multiple efforts are underway to document "AI incidents," which ought to record, when known, the specific model involved.[71] Governments should monitor malicious use, namely by specifying and collecting crime statistics, to inform the extent to which models vs. systems, widely available weights vs. closed models warrant particular concern. Among other activities, considering and observing leading indicators of use may be useful. Market incentives drive individuals and small groups to use AI systems to drastically scale up their impact. Thus, market monitoring of AI use, including via existing regulatory mechanisms, may provide leading indicators of malicious-use risks.

Evaluation of the benefits and risks of new categories of models warrants specific consideration. Text-to-image models are now ubiquitous as proprietary solutions, widely available model weights, and applications using both. In hindsight, it is unclear if a rubicon was crossed with the development of diffusion-based models, given prior generative-adversarial-network techniques published openly for years. Better understanding is needed to evaluate the extent to which the public release of possibly paradigm-shifting models presents unacceptable risks relative to their benefits, considered marginally to similar closed capabilities. Stakeholders may consider using staged-release[72] methods prior to releasing model weights publicly, in order to gather more information about new paradigm-shifting models.

[71] Namely, the AI Incident Database and OECD AI Incidents Monitor.
[72] Staged release is not a unified concept: it can take many forms depending on the development context, intermediate deployment scenario, and specific goals. For example, one form may see use restricted to model-as-a-service provision only whereas another may share full access to model weights with a set of individuals.

Today, available evidence of the marginal risks of public release does not substantiate restrictions on current AI model paradigms. Instead, government should prioritize AI regulation against reckless use and consider criminal justice policies and emergency national security plans for malicious use.

. . .

Thank you for the opportunity to share GitHub's perspective on widely available model weights. We appreciate NTIA's commitment to a thorough consultation to surface evidence and diverse perspectives on the benefits of, risks from, and policy mechanisms for widely available AI models. As you analyze responses and evaluate next steps pursuant to EO 14110, we stand ready to work with you and to answer any further questions that may arise.

Respectfully submitted,

Mike Linksvayer
VP of Developer Policy, GitHub
mlinksva@github.com

Peter Cihon
Senior Policy Manager, GitHub
pcihon@github.com