Center for Research on Foundation Models

Stanford University Human-Centered Artificial Intelligence

CENTER FOR INFORMATION TECHNOLOGY POLICY
CITP.PRINCETON.EDU

Department of Commerce
National Telecommunications and Information Administration
Docket No. 230407-0093
RIN 0660-XC057
AI Accountability Policy Request for Comment

June 12, 2023

To Whom It May Concern:

We, researchers from the Stanford Center for Research on Foundation Models (CRFM), part of the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and Princeton University's Center for Information Technology Policy (CITP), offer the following submission in response to the Request for Comment (RFC) by the National Telecommunications and Information Administration on AI accountability policy.[1] We center our response on *foundation models* (FMs), which constitute a broad paradigm shift in AI. Foundation models require substantial data and compute to provide striking capabilities that power countless downstream products and services.[2] While many prominent uses and abuses of FMs have been highlighted, we focus on consequential aspects that, if addressed effectively, will significantly improve the state of the FM ecosystem.

Given the significance of foundation models, we argue that *pervasive opacity* compromises accountability for foundation models. Foundation models and the surrounding ecosystem are insufficiently transparent, with recent evidence showing this transparency is deteriorating further.[3] Without sufficient transparency, the federal government and industry cannot implement meaningful accountability mechanisms as we cannot govern what we cannot see. The history of regulating online platforms and social media foretells the story for foundation models: If we fail to act now to ensure foundation models are sufficiently transparent, we are destined to repeat the avoidable errors of the past.

Our submission recommends the following in response to questions posed in the RFC:

- **Invest in digital supply chain monitoring for foundation models** (Section 2; Questions 5, 11, 15, 20).
- **Invest in public evaluations of foundation models** (Section 3; Questions 3, 21, 23, 29, 30b).
- **Incentivize research on guardrails for open-source models** (Section 4; Question 7, 32).

[1] This response reflects the independent views of the undersigned scholars.
[2] Rishi Bommasani, …, and Percy Liang. *On the Opportunities and Risks of Foundation Models*. 2021. https://crfm.stanford.edu/report.html.
[3] Rishi Bommasani et al. *Ecosystem Graphs: The Social Footprint of Foundation Models*. 2023. https://crfm.stanford.edu/2023/03/29/ecosystem-graphs.html.

Foundation models are an immature technology advancing at an unprecedented clip. To successfully implement our recommendations, collaboration between the federal government, academia, and industry will be necessary. As academic researchers, we highlight academia's unique strengths: interdisciplinary scholarship and neutral science divorced from commercial interests. We highlight the importance of additional investment in resources and infrastructure to advance these efforts on AI accountability: academic compute infrastructure[4] and federal funding for AI research in the public interest.[5]

1. Background on foundation models

Foundation models are general-purpose technologies that function as platforms for a wave of AI applications, including generative AI: AI systems that can generate compelling text, images, videos, speech, music, and more. Well-known examples include OpenAI's ChatGPT, which is a language model that can converse with users and perform complex tasks as instructed through its language interface, and Stability AI's Stable Diffusion, which is a text-to-image model that can generate photorealistic images from text-based prompts.

Foundation models constitute a paradigm shift in AI development and deployment. Rather than developing a single bespoke model for each application, foundation models require tremendous upfront resource investment (e.g., tens or hundreds of millions of dollars and trillions of bytes of data for the most capable systems like OpenAI's GPT-4). These high upfront costs are justified by the significant new capabilities of these models, which can be reused across many downstream use cases.

Why are foundation models a critical priority? Foundation models underpin many of the recent advances in AI: We highlight five properties that indicate why they merit significant priority.

1. **Nascent.** Given their recent development, there is no well-developed understanding of how their risks will be addressed by any combination of self-regulation and regulation.
2. **Prominent.** Foundation models are the center of the public awareness on AI, mediated by daily global media attention from many of the world's largest news outlets.
3. **Burgeoning.** Foundation models are the fastest-growing consumer technology in U.S. history[6] with tremendous commercial investment in startups[7] and established companies.[8]

---

[4] White House. *National Artificial Intelligence Research Resource Task Force Releases Final Report*. 2023. https://www.whitehouse.gov/ostp/news-updates/2023/01/24/national-artificial-intelligence-research-resource-task-force-releases-final-report/.

[5] White House. *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety*. 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

[6] Krystal Hu. *ChatGPT Sets Record for Fastest-Growing User Base*. Reuters. 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[7] More than $11 billion in the first fiscal quarter of 2023 was directed toward foundation model startups. See: Ian Hogarth. *We Must Slow Down the Race to God-like AI*. Financial Times. 2023. https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2

[8] According to an Accenture market survey, "98% of global executives agree AI foundation models will play an important role in their organizations' strategies in the next 3 to 5 years." See Accenture. *Technology Vision 2023: When Atoms meet Bits*. 2023. https://www.accenture.com/content/dam/accenture/final/accenture-com/a-com-custom-component/iconic/document/Accenture-Technology-Vision-2023-Full-Report.pdf.

There has also been significant investment in building the technology by a number of countries around the world.[9]

4. **Pervasive.** Foundation models already power products spanning dozens of market sectors[10] and will only continue to spread, akin to other defining technologies like the internet, computers, mobile phones, and semiconductors.
5. **Consequential.** Foundation models are expected to influence multiple dimensions of the lives of almost every American. Their importance has already been recognized by policymakers around the world: They are the centerpiece of most recent revisions to the EU AI Act, the subject of a new task force reporting to the U.K. prime minister, and the topic of interest across multiple U.S. federal entities beyond the NTIA, including the White House Office of Science and Technology Policy (OSTP), the National AI Advisory Committee (NAIAC), the Federal Trade Commission (FTC), and the National Institute of Standards and Technology (NIST).[11]

2. Invest in digital supply chain monitoring for foundation models (Questions 5, 11, 15, 20)

Foundation models function as the understructure for a diverse range of products. Consequently, they play a central role in the broader AI ecosystem and digital supply chain.[12] The foundation model ecosystem can be summarized by three categories of assets:

1. **Resources**, meaning the data (e.g., the billions of words and images on the Internet) and computation (e.g., through cloud providers like Amazon, Google, and Microsoft) necessary for training foundation models;
2. **Foundation models**, such as ChatGPT and Stable Diffusion; and
3. **Products and services** built atop these models (e.g., Bing Search, Khan Academy's Khanmigo AI-powered tutor).

These assets determine much of the digital supply chain, thereby intermediating dependences between organization (e.g., Khan Academy depends on OpenAI because GPT-4 powers Khanmigo[13]) and, in turn, the sectors affected by foundation models. The ecosystem view makes clear where existing sector-level regulatory authority can be used to hold foundation models, the

---

[9] U.K. Department for Science, Innovation and Technology. *A Pro-Innovation Approach to AI Regulation.* 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper; Jeffrey Ding and Jenny Xiao. *Recent Trends in China's Large Language Model Landscape*. 2023. https://www.governance.ai/research-paper/recent-trends-chinas-llm-landscape; Pat Brans. *Sweden is Developing its Own Big Language Model*. 2023. https://www.computerweekly.com/news/366538232/Sweden-is-developing-its-own-big-language-model.

[10] Rishi Bommasani et al. *Ecosystem Graphs.*

[11] See White House. *FACT SHEET*; National AI Advisory Committee. *Year 1 Report*. 2023. https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf; Federal Trade Commission. *Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale*. 2023. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale; National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework*. 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[12] Aleksander Mądry. *Written Statement for the Hearing, Advances in AI: Are We Ready for a Tech Revolution? In Front of the House Cybersecurity, Information Technology, and Government Innovation Subcommittee*. 2023. https://oversight.house.gov/wp-content/uploads/2023/03/madry_written_statement100.pdf; Hopkins et al. *The Diverse Landscape of AI Supply Chains: The AIaaS Supply Chain Dataset*. Thoughts on AI Policy. 2023. https://aipolicy.substack.com/p/supply-chains-3-5.

[13] OpenAI. *Khan Academy*. 2023. https://openai.com/customer-stories/khan-academy.

companies that provide them, and their downstream products and services (e.g., in medicine or law) to account.

This digital supply chain is critical to many dimensions of commerce in the United States, including well-functioning, competitive markets. To date, the supply chain has been recognized as a key focus area for regulatory efforts per the EU AI Act,[14] the foundation model market review of the U.K.'s Competition and Markets Authority,[15] and the recent testimony of MIT Professor Aleksander Madry before the House Oversight Subcommittee on Cybersecurity, Information Technology, and Government Innovation.[16] To parallel the market surveillance initiatives in both the EU and the U.K., we recommend federal investment into digital supply chain monitoring for foundation models. We review precedent for such monitoring in other settings for digital technology as well as motivations for why such an initiative would improve AI accountability.

The practice of digital supply chain monitoring. While the foundation model ecosystem is complex and evolving, it is hardly unique. Almost every consumer product is the composite of many materials or ingredients. In the setting of digital technologies, we can look to the Software Bill of Materials (SBOM) as an example of effective dependency tracking mediated by government intervention.[17] As described by the Cybersecurity and Infrastructure Security Agency (CISA), an SBOM is "a list of ingredients that make up software components [that] has emerged as a key building block in software security and software supply chain risk management."[18] In particular, "SBOM work has advanced since 2018 as a collaborative community effort, driven by National Telecommunications and Information Administration's (NTIA) multistakeholder process."[19] As a direct analogy, the federal government should track the assets and supply chain in the foundation model ecosystem to understand market structure, address supply chain risk, and promote resiliency. As an example implementation, Stanford's Ecosystem Graphs currently documents the foundation model ecosystem, supporting a variety of downstream policy use cases and scientific analyses.[20]

How digital supply chain monitoring improves accountability. Supply chain monitoring is highly multifunctional—we describe three clear benefits. First, supply chain monitoring enables *recourse* to stop further harm. For example, the National Highway Traffic Safety Administration (NHTSA) monitors the automobile supply chain, conducting recalls when a part (e.g., a batch of

[14] For example, see Article 28 entitled "Responsibilities Along the AI Value Chain of Providers, Distributors, Importers, Deployers or Other Third Party" in the current European Parliament version: European Parliament. *DRAFT Compromise Amendments on the Draft Report, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. 2023. https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf.
[15] U.K. Competition and Markets Authority. *AI Foundation Models: Initial Review*. 2023. https://www.gov.uk/cma-cases/ai-foundation-models-initial-review.
[16] Aleksander Mądry. *Written Statement for the Hearing.*
[17] Executive Order 14028. *Improving the Nation's Cybersecurity*. 2021. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/.
[18] Cybersecurity & Infrastructure Security Agency. Software Bill of Materials (SBOM). https://www.cisa.gov/sbom.
[19] National Telecommunications and Information Administration. *Software Bill of Materials*. https://ntia.gov/page/software-bill-materials.
[20] Rishi Bommasani et al. *Ecosystem Graphs.*

car brakes) has been identified to be faulty.[21] Second, supply chain monitoring identifies *algorithmic monoculture*[22]: the pervasive dependence on a single asset (e.g., foundation model). As SEC Chair Gary Gensler has noted, such dependence can yield "concentrated risk."[23] Finally, supply chain monitoring for foundation models, which cuts across sectoral boundaries, exposes *regulatory opportunities and weaknesses*. Fundamentally, supply chain monitoring identifies which sectors are impacted by foundation models (e.g., products in the sector depend on FMs; data from the sector powers FMs). This provides a principled means for identifying regions where sector-specific regulatory authority will not suffice to hold AI to account and chokepoints for precision regulation in the future.

3. Invest in public evaluations of foundation models (Questions 3, 21, 23, 29, 30b)

Evaluation is the standard methodology for quantifying the capabilities, limitations, and risks of AI systems. Standardized evaluations simultaneously clarify the current status and orient future progress. At present, some researcher evaluations exist for foundation models (especially language models), but none have risen to the status of bona fide standards. Effective evaluations are complex to design: How should the endless use cases for a general-purpose technology be evaluated?[24] The most recent Parliament draft version of the EU AI Act, for example, requires that foundation model providers evaluate their models on public or industry standard benchmarks.[25] We highlight three specific considerations for effective evaluation of foundation models that the federal government should implement. The requirement for the National Institute of Standards and Technology (NIST) to develop AI testbeds as directed in the CHIPS and Science Act provides a direct opportunity to implement these recommendations.[26]

Public. To ensure foundation models are transparent, evaluations should be public and follow clear, openly disclosed practices.[27] Such evaluations will improve the scientific discourse surrounding these models, combatting various forms of hype and advancing our collective understanding of this emerging technology. Public evaluations set the baseline for how we should reason about this technology. To evaluate models publicly and to engender trust, we reference efforts like Stanford's Holistic Evaluation of Language Models (HELM).[28] These

---

[21] National Highway Traffic Safety Administration. *Motor Vehicle Safety Defects and Recalls*. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/14218-mvsdefectsandrecalls_041619-v2-tag.pdf.
[22] Jon Kleinberg and Manish Raghavan. *Algorithmic Monoculture and Social Welfare*. Proceedings of the National Academy of Sciences. 2021. https://www.pnas.org/doi/10.1073/pnas.2018340118; Rishi Bommansani, et al. *Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?* Advances in Neural Information Processing Systems. 2022. https://arxiv.org/abs/2211.13972.
[23] Betsy Vereckey. *SEC's Gary Gensler on How Artificial Intelligence Is Changing Finance.* 2022. https://mitsloan.mit.edu/ideas-made-to-matter/secs-gary-gensler-how-artificial-intelligence-changing-finance.
[24] Inioluwa Deborah Raji et al. *AI and the Everything in the Whole Wide World Benchmark*. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021. https://arxiv.org/pdf/2111.15366.pdf
[25] Annex VIII, Section C on the registration of foundation models in a free and public EU database https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf
[26] U.S. Congress. H.R.4346: CHIPS and Science Act (Section 10232). 2022. https://www.congress.gov/bill/117th-congress/house-bill/4346/text.
[27] Rishi Bommasani et al. *Improving Transparency in AI Language Models: A Holistic Evaluation*. Stanford Institute for Human-Centered Artificial Intelligence. 2023. https://hai.stanford.edu/foundation-model-issue-brief-series.
[28] Rishi Bommasani, Percy Liang, Tony Lee. *Language Models Are Changing AI: The Need for Holistic Evaluation*. 2023. https://crfm.stanford.edu/2022/11/17/helm.html.

efforts demonstrate the components required for useful public evaluations: (i) clear evaluation methodology and definition of the relevant metrics, (ii) easily inspected results for individual models, and (iii) the underlying predictions or specific model behaviors that get aggregated to yield the results.

Holistic. To ensure public evaluations surface the relevant dimensions of foundation models, these evaluations should be holistic. Since foundation models are broad-reaching, general-purpose technologies, they can be used across a range of use cases and should satisfy a range of objectives (e.g., be accurate, robust, trustworthy, fair, efficient). Evaluation should address these many dimensions: The NIST AI Risk Management Framework recognizes the importance of such multidimensional evaluations.[29] Of specific importance are the inevitable trade-offs: For example, one foundation model may be more accurate but also more discriminatory than another in a given context. Or, in other cases, different metrics may be highly correlated: Multiple works establish that more accurate models are more robust or reliable.[30]As an example implementation, we point to HELM once again: language models are evaluated across a range of use cases (e.g., question answering, document summarization), desiderata (e.g., fairness, uncertainty), and capabilities/risks (e.g., world knowledge, disinformation generation).

All-encompassing. To ensure public and holistic evaluations hold all foundation models to account, evaluations should encompass models that are restricted or closed, meaning models that are not openly accessible to researchers and the public. At present, foundation model providers adopt a variety of policies to release models.[31] [32] Some models like EleutherAI's GPT-NeoX are open-sourced, whereas others, like Google's Flamingo, are entirely closed to the public. The inability of the public, including researchers and civil society, to investigate and interrogate these models inhibits external scrutiny. Given these models provide the core capabilities that power products, including technologies like Google Search that affect hundreds of millions of users, the models themselves should be evaluated.[33] To improve the status quo, the government should require all model developers to create programs for external researcher access.[34] Given companies may be disincentivized to provide meaningful access (e.g., due to concerns of intellectual property or competitive pressure), we encourage investigation into innovative

[29] National Institute of Standards and Technology. *AI Risk Management Framework*. 2023. https://www.nist.gov/itl/ai-risk-management-framework.
[30] John Miller et al. *Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization*. Proceedings of the International Conference on Machine Learning. 2021. https://arxiv.org/pdf/2107.04649.pdf; Percy Liang et al. *Holistic Evaluation of Language Models*. 2023. https://arxiv.org/pdf/2211.09110.pdf.
[31] Percy Liang et al. *The Time Is Now to Develop Community Norms for the Release of Foundation Models*. 2022. https://crfm.stanford.edu/2022/05/17/community-norms.html.
[32] Irene Solaiman. *The Gradient of Generative AI Release: Methods and Considerations.* 2023. https://arxiv.org/pdf/2302.04844.pdf.
[33] Rishi Bommasani et al. *Improving Transparency in AI Language Models.*
[34] Microsoft. *Microsoft Turing Academic Program (MS-TAP)*. https://www.microsoft.com/en-us/research/collaboration/microsoft-turing-academic-program/.

strategies such as *sandboxes*[35] to allow for testing in contained environments or *developer-mediated access*[36] to allow evaluations to be gated by provider consent.

4. Incentivize research on guardrails for open-source models (Questions 7, 32)

For foundation models to advance the public interest, their development and deployment should ensure transparency, support innovation, distribute power, and minimize harm. The *release policies* for foundation models,[37] in particular, directly influence these four goals. Currently, foundation model providers adopt different release policies, with the most capable foundation models often restricted in terms of the access available to the public.[38] In other words, while some of these models (e.g., OpenAI's ChatGPT, Anthropic's Claude) are available for the public to interact with, the internals (i.e., model weights) and training data are not publicly available.

Open-source efforts for digital technologies like operating systems (e.g., Linux) and browsers (e.g., Mozilla Firefox) establish the precedent that open-source can simultaneously achieve these four goals. Therefore, we consider *open-source foundation models*: foundation models where the internals (i.e., model weights) are available (without major use restrictions) and the training is reproducible by researchers.[39] We argue open-source foundation models can achieve all four of these objectives, in part due to inherent merits of open-source (pro-transparency, pro-innovation, anti-concentration), if the federal government incentivizes a responsible open-source ecosystem.

Transparency. Transparency is a hallmark virtue of open-source: We should expect open-source approaches to perform especially well in terms of their transparency. Open-source foundation models guarantee a certain degree of transparency: If the model is open-sourced, other entities can access and scrutinize the model, improving public understanding and trust. In practice, open-source models are often released alongside the entire training data and codebase to reproduce them.[40] This improves their auditability, since independent researchers can examine how well they work. In particular, the transparency interventions recommended in earlier sections could be easier to implement for open-source models.

Innovation. With the release of open-source models, many researchers and technologists can experiment with new directions for developing FMs. As one example of how open-source models enable innovation, consider LLaMA. The release of the LLaMA language model by Meta spawned a large number of research projects, including advances in miniaturization[41],

---

[35] Title V, Article 53, entitled "AI Regulatory Sandboxes." https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf.
[36] Toby Shevlane. *Structured Access: An Emerging Paradigm for Safe AI Deployment*. 2022. https://arxiv.org/ftp/arxiv/papers/2201/2201.05159.pdf.
[37] Percy Liang et al. *The Time Is Now to Develop Community Norms for the Release of Foundation Models*.
[38] Solaiman. *The Gradient of Generative AI Release.*
[39] For a deeper discussion into the various release strategies for foundation models, see: Solaiman. *The Gradient of Generative AI Release.*
[40] The MosaicML NLP Team. *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*. 2023. https://www.mosaicml.com/blog/mpt-7b.
[41] Georgi Gerganov's ggml library allows users to run capable open-source models like Meta's LLaMA and OpenAI's Whisper on their local computers. See GGML - AI at the edge. https://ggml.ai/.

instruction-following[42], and fine-tuning models efficiently[43]. These innovations would not be possible without access to a model's weights, so in the absence of open-source models, these innovations would either be restricted to AI companies, or they wouldn't take place at all. There is precedent for such innovation—most notably under the umbrella of Free and Open Source Software (FOSS), like Linux and Mozilla Firefox.

Distribution of power and expertise. Concentration of AI resources in corporations would make it harder for people in the public sector to develop expertise in FMs and to develop public interest technology atop FMs. A robust open-source ecosystem would allow developers and researchers from a diversity of backgrounds to build expertise in and contribute to the development of FMs. We have seen such efforts before; notably, the BLOOM language model was built by an open-source collaboration of over a thousand researchers.[44]

Security and risk mitigation. Open-source generally involves contributions from many individuals, amounting to a more chaotic ecosystem that might be less secure. Namely, several organizations have argued AI systems are more secure if restrictions are enforced on who can create state-of-the-art FMs, similar to nuclear weapons.[45] However, such efforts could have undesirable effects and hamper our ability to deal with AI risks.[46]

If closed-source models cannot be examined by researchers and technologists, security vulnerabilities might not be identified before they cause harm. (One example of such a vulnerability is memorization: language models' tendency to memorize data, including sensitive information like credit card numbers, which can later be extracted by users.[47] Another example is prompt injection, where a malicious instruction can trick a language model into performing unintended tasks, such as leaking private information when using personal assistants.[48]) On the other hand, experts across domains can examine and analyze open-source models, which makes security vulnerabilities easier to find and address.

In addition, restricting who can create FMs would reduce the diversity of capable FMs and may result in single points of failure in complex systems. If the same FM powers many different products and services, a security vulnerability in the FM would affect all of them.[49] [50] A diverse

---

[42] Rohan Taori et al. *Alpaca: A Strong, Replicable Instruction-Following Model.* 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html.
[43] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs.* 2023. https://arxiv.org/abs/2305.14314.
[44] Teven Le Scao et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* 2023. https://arxiv.org/abs/2211.05100.
[45] Sam Altman, Greg Brockman, and Ilya Sutskever. *Governance of Superintelligence.* OpenAI, 2023. https://openai.com/blog/governance-of-superintelligence.
[46] For an overview of why such restrictions are unlikely to be effective, see: Sayash Kapoor and Arvind Narayanan. *Licensing Is Neither Feasible Nor Effective for Addressing AI Risks.* 2023. https://aisnakeoil.substack.com/p/licensing-is-neither-feasible-nor.
[47] Nicholas Carlini et al. *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.* https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.
[48] Simon Willison. *Prompt Injection: What's the Worst That Can Happen?* 2023. https://simonwillison.net/2023/Apr/14/worst-that-can-happen/.
[49] Monoculture has also affected security issues in the past. For examples of security risks due to lack of diversity in computational infrastructure, see Peter Eder-Neuhauser, Tanja Zseby, and Joachim Fabini. *Malware Propagation in Smart Grid Monocultures*. 2018. https://link.springer.com/article/10.1007/s00502-018-0616-5.
[50] Rishi Bommasani, …, and Percy Liang. *On the Opportunities and Risks of Foundation Models*.

selection of capable open-source models could avoid single points of failure that arise from restrictions on developing state-of-the-art models.

Guardrails for open-source FMs: a four-pronged approach. Open-source FMs do bear risks. Users could harm themselves or others using these models.[51] For the federal government to incentivize a flourishing and responsible open-source ecosystem as we have seen for other technologies, the regulatory approach for open-source must differ from proprietary FMs. For example, open-source developers are often ill-equipped to meet requirements on downstream uses of these models, whereas providers who bring models to market are better targets for such requirements. We suggest a more nuanced approach that addresses different steps in the foundation model life cycle such as development, deployment, and use.

1. **Transparency of model developers.** The federal government should require developers to perform transparent evaluations of open-source foundation models prior to release, so that stakeholders can understand the capabilities and risks. Previous sections illustrate how these requirements could be shaped.

2. **Compliance of downstream providers.** A foundation model is not a product by itself. Consumers won't use FMs directly, but rather products and services that incorporate them. These products and services are subject to sectoral consumer protections and product safety restrictions. Regulatory agencies should enforce these requirements on providers of the consumer-facing products or services built using open-source FMs.

3. **Resilience of attack surfaces.** Bad actors may use FMs to generate disinformation, find security vulnerabilities in software, or cause other forms of harm.[52] [53] Each of these malicious uses involves an *attack surface*. For example, the attack surface for disinformation is typically a social media platform—that is, where influence operators seek to disseminate disinformation and persuade people. For security vulnerabilities, the attack surface may be software codebases. While efforts may be taken to prevent the adaptation of foundation models for malicious purposes,[54] policy should focus on attack surfaces that come under greater pressure due to the proliferation of FMs. Such policies could include incentivizing social media platforms to strengthen their information integrity efforts, and increasing funding for cybersecurity and infrastructure defense efforts such as via CISA.

4. **Research on open-source FMs.** Open-source foundation models are a nascent methodology where our understanding of the risks of FMs and ways to address them is rapidly evolving. To realize a thriving and responsible open-source foundation model

[51] Laura Weidinger et al. *Taxonomy of Risks Posed by Language Models.* ACM FAccT, 2022. https://dl.acm.org/doi/10.1145/3531146.3533088.
[52] For instance, the June 6, 2023, letter from Sen. Richard Blumenthal and Sen. Josh Hawley to Meta outlines various risks of releasing open-source models. See: https://www.blumenthal.senate.gov/imo/media/doc/06062023metallamamodelleakletter.pdf.
[53] Josh A. Goldstein et al. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* 2023. https://arxiv.org/abs/2301.04246.
[54] Eric Mitchell et al. *Self-Destructing Models: Increasing the Costs of Harmful Dual Uses in Foundation Models*. 2022. https://arxiv.org/pdf/2211.14946.pdf.

ecosystem, we must resolve fundamental research problems in transparency, compliance, and malicious use. Federal funding for research on the risks and mitigations of open-source FMs would ensure that our understanding of the policy options can keep pace with technology.[55]

Sincerely,

Rishi Bommasani
Researcher & Society Lead, Stanford Center for Research on Foundation Models
Ph.D. Candidate, Stanford University

Sayash Kapoor
Researcher, Princeton Center for Information Technology Policy
Ph.D. Candidate, Princeton University

Daniel Zhang
Senior Manager for Policy Initiatives, Stanford Institute for Human-Centered Artificial Intelligence

Dr. Arvind Narayanan
Incoming Director, Princeton Center for Information Technology Policy
Professor of Computer Science, Princeton University

Dr. Percy Liang
Director, Stanford Center for Research on Foundation Models
Associate Professor of Computer Science and (By Courtesy) of Statistics, Stanford University

[55] As an example of research on the data governance practices of open-source models, see Jernite et al. on the release of the BLOOM model: Yacine Jernite et al. *Data Governance in the Age of Large-Scale Data-Driven Language Technology.* ACM FAccT, 2022. https://dl.acm.org/doi/abs/10.1145/3531146.3534637.