Issue Brief
HAI Policy & Society
December 2023

Considerations for Governing Open Foundation Models

Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, Percy Liang

Introduction

FOUNDATION MODELS (E.G., GPT-4, LLAMA 2) ARE AT THE EPICENTER OF AI, driving technological innovation and billions in investment. This paradigm shift has sparked widespread demands for regulation. Animated by factors as diverse as declining transparency and unsafe labor practices, limited protections for copyright and creative work, as well as market concentration and productivity gains, many have called for policymakers to take action.

Central to the debate about how to regulate foundation models is the process by which foundation models are released. Some foundation models like Google DeepMind's Flamingo are fully closed, meaning they are available only to the model developer; others, such as OpenAI's GPT-4, are limited access, available to the public but only as a black box; and still others, such as Meta's Llama 2, are more open, with widely available model weights enabling downstream modification and scrutiny. As of August 2023, the U.K.'s Competition and Markets Authority documents the most common release approach for publicly-disclosed models is open release based on data from Stanford's Ecosystem Graphs. Developers like Meta, Stability AI, Hugging Face, Mistral, Together AI, and EleutherAI frequently release models openly.

Key Takeaways

Open foundation models, meaning models with widely available weights, provide significant benefits by combatting market concentration, catalyzing innovation, and improving transparency.

Some policy proposals have focused on restricting open foundation models. The critical question is the *marginal risk* of open foundation models relative to (a) closed models or (b) pre-existing technologies, but current evidence of this marginal risk remains quite limited.

Some interventions are better targeted at choke points downstream of the foundation model layer.

Several current policy proposals (e.g., liability for downstream harm, licensing) are likely to disproportionately damage open foundation model developers.

Policymakers should explicitly consider potential unintended consequences of AI regulation on the vibrant innovation ecosystem around open foundation models.

Governments around the world are issuing policy related to foundation models. As part of these efforts, open foundation models have garnered significant attention: The recent U.S. Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence tasks the National Telecommunications and Information Administration with preparing a report on open foundation models for the president. In the EU, open foundation models trained with fewer than 10^{25} floating point operations (a measure of the amount of compute expended) appear to be exempted under the recently negotiated AI Act. The U.K.'s AI Safety Institute will "consider open-source systems as well as those deployed with various forms of access controls" as part of its initial priorities. Beyond governments, the Partnership on AI has introduced guidelines for the safe deployment of foundation models, recommending against open release for the most capable foundation models.

Policy on foundation models should support the open foundation model ecosystem, while providing resources to monitor risks and create safeguards to address harms. Open foundation models provide significant benefits to society by promoting competition, accelerating innovation, and distributing power. For example, small businesses hoping to build generative AI applications could choose among a variety of open foundation models that offer different capabilities and are often less expensive than closed alternatives. Further, open models are marked by greater transparency and, thereby, accountability. When a model is released with its training data, independent third parties can better assess the model's capabilities and risks.

However, an emerging concern is whether open foundation models pose distinct risks to society. Unlike closed foundation model developers, open developers have limited ability to restrict the use of their models by malicious actors that can easily remove safety guardrails. Recent studies claim that open foundation models are more likely to generate disinformation, cyberweapons, bioweapons, and spear-phishing emails.

Open foundation models provide significant benefits to society by promoting competition, accelerating innovation, and distributing power.

Correctly characterizing these distinct risks requires centering the *marginal* risk: To what extent do open foundation models increase risk relative to (a) closed foundation models or (b) pre-existing technologies like search engines? We find that for many dimensions, the existing evidence about the marginal risk of open foundation models remains quite limited. In some instances, such as the case of AI-generated child sexual abuse material (CSAM) and nonconsensual intimate imagery (NCII), harms stemming from open foundation models have been better documented. For these demonstrated harms, proposals to restrict the release of foundation models via licensing of compute-intensive models are mismatched, because the text-to-image models used to cause these harms require vastly lower amounts of resources to train.

More broadly, several regulatory approaches under consideration are likely to have a disproportionate impact on open foundation models and their

developers, without meaningfully reducing risk. Even though these approaches do not differentiate between open and closed foundation model developers, they yield asymmetric compliance burdens. For example, legislation that holds developers liable for content generated using their models or their derivatives would harm open developers as users can modify their models to generate illicit content. Policymakers should exercise caution to avoid unintended consequences and ensure adequate consultation with open foundation model developers before taking action.

Developers have many intermediary options between the fully closed setting (nothing is released to anyone) and the fully open setting (every asset is released to everyone).

Background

Central to understanding foundation models is the topic of *release*: To what extent and through what mechanisms are foundation models made available to entities beyond the foundation model developer? The landscape of release is multidimensional: Different assets (e.g., training data, code, model weights) can be released to chosen entities or to the public at large. Developers have many intermediary options between the fully closed setting (nothing is released to anyone) and the fully open setting (every asset is released to everyone).

The release of foundation models is a gradient: Models can be *fully closed* (not available to anyone outside the developer organization, like Google DeepMind's Flamingo); *hosted access* (available via a web interface, like Inflection's Pi); *cloud-based access* (available via an API, like OpenAI's GPT-4); *cloud-based fine-tuning access* (both the model and the ability to fine-tune it are available via an API, like OpenAI's GPT-3.5); *widely available weights* (like Stability AI's Stable Diffusion and Meta's Llama 2); and available with the *weights, code, and data*—either with use restrictions, like BigScience's BLOOM, or without, like EleutherAI's GPT-NeoX.

We use the notion of *open foundation models*: models released with widely available weights, which corresponds to the three rightmost categories in the figure (modified with permission from Solaiman (2023)). This aligns with the distinction drawn in the

Level of Access	Fully closed	Hosted access	API access to model	API access to fine tuning	Weights available	Weights, data, and code available with use restrictions	Weights, data, and code available without use restrictions
Example	Flamingo (Google)	Pi (As of 2023; Inflection)	GPT-4 (As of 2023; OpenAI)	GPT-3.5 (OpenAI)	Llama 2 (Meta)	BLOOM (BigScience)	GPT-NeoX (EleutherAI)

Foundation models with widely available weights

U.S. Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Many of the concerns surrounding open foundation models arise from the fact that once model weights are released, developers relinquish control over their downstream use. Even if developers impose restrictions on downstream use and who can download the model, such restrictions can be ignored by downstream users, especially malicious users. In contrast, in the face of malicious use or other important risks, closed foundation model developers can restrict access to the models (i.e., reduce access by shifting to a more restrictive point on the gradient of model release). We stress, however, that this categorical distinction may oversimplify the gradient of model release: Closed models may also be susceptible to malicious use, given that current safeguards are not robust enough to withstand adversarial attacks.

Closed models may also be susceptible to malicious use, given that current safeguards are not robust enough to withstand adversarial attacks.

Open foundation models are reminiscent of, but not the same as, open-source software: Machine learning models are built using datasets as well as code, making them fundamentally different from many kinds of software. The standard definition of open-source software prohibits restrictions on specific users or use cases, while open foundation models often include these restrictions; Meta restricts the use of its Llama 2 model by entities with more than 700 million monthly active users, and other organizations use Open & Responsible AI licenses with use restrictions.

Nevertheless, the history of open-source software provides insight on how to govern open foundation models. Open-source software validates the tremendous societal benefits of open technologies: The European Commision reports that an investment of "around €1 billion in [open-source software] ... resulted in an impact on the European economy of between €65 and €95 billion." Open-source software powers critical infrastructure: The U.S. Digital Services Playbook encourages U.S. government digital services to "default to open" due to the benefits of reusability, robustness, transparency, and collaboration. The Department of Defense's website on open-source software states, "Continuous and broad peer-review, enabled by publicly available source code, improves software reliability and security through the identification and elimination of defects that might otherwise go unrecognized by the core development team." There is no empirical evidence that open-source software is more vulnerable or insecure than closed-source software.

Benefits of open foundation models

We highlight three fundamental societal objectives where open foundation models provide clear benefits by distributing power, catalyzing innovation, and ensuring transparency.

Distributing power. Foundation models are emerging technologies: Given their influence, these models create new forms of socioeconomic power, which demands an assessment of how that power is distributed. In the words of MIT economists Daron Acemoglu and Simon Johnson: "The consequences of any technology depend on who gets to make pivotal decisions about how the technology develops. This is doubly true for AI, because these new tools can be developed for many different types of activities, with the potential to spread rapidly in every sector of the economy and in every aspect of our lives." Closed model developers exert greater power in defining and restricting use cases they deem unacceptable, whereas downstream consumers of foundation models can better make these decisions for themselves with open models. Further, closed model developers may more directly shape downstream markets through vertical integration, potentially leading to problematic monocultures where many downstream products/services depend on the same foundation model. Overall, closed foundation models may contribute to more concentrated power in the hands of developers, which we should scrutinize given the well-established risks of market concentration for digital technologies.

Catalyzing innovation. Foundation models are general-purpose technologies that can produce sharp increases in innovation. Notably, foundation models bolster economic and scientific productivity, with Bloomberg Intelligence projecting that generative AI will become a $1.3 trillion market by 2032. Open foundation models are necessary for several forms of research (e.g., interpretability work, public development of watermark techniques, forms of security research, and model training and inference efficiency techniques). Overall, open foundation models are generally more customizable and provide deeper access, which are key ingredients for greater innovation.

Ensuring transparency. Digital technologies such as foundation models are plagued by opacity, from dark patterns on social media to ghost work as invisible labor. Adequate transparency from foundation model developers is instrumental for many objectives: civil society, governments, industry, and academia have all called for transparency. UN Secretary-General António Guterres proposed to "make transparency, fairness and accountability the core of AI governance … [and] consider the adoption of a declaration on data rights that enshrines transparency." The 2023 Foundation Model Transparency Index demonstrates that major open foundation model developers are consistently and considerably more transparent on average than their closed counterparts. Open foundation model developers score on average 20 percentage points higher on the index, outperforming closed developers in terms of transparency with respect to each part of the supply chain. Such transparency may be important to avoid reproducing the harms facilitated by opaque digital technologies in the past, but the current lack of transparency about downstream impacts on the economy and society remains a concern.

Risks of open foundation models

In spite of the significant benefits of open foundation models, current policy attention on the risks of open foundation models is largely motivated by their potential for malicious use. Here, we consider a range of misuse threat vectors to better characterize the state of evidence today for each. Rigorous evidence of marginal risk remains quite limited. This does not mean that open foundation models pose no risk along these vectors but, instead, that more rigorous analysis will be required to ground policy interventions. In particular, critical

research is needed on *marginal risk*: To what extent do open foundation models increase risk either relative to closed foundation models or relative to pre-existing technologies (e.g., search engines)?

Disinformation. Foundation models may reduce the cost of generating persuasive disinformation. While closed foundation model providers may be better positioned to reject requests to generate disinformation, the ambiguity of what constitutes disinformation calls into question the technical feasibility of such refusals. More fundamentally, the key bottleneck for effective influence operations is not disinformation generation but disinformation dissemination: Online platforms that control the reach of content are better targets for policy intervention. To date, we are unaware of empirical evidence that open foundation models increase societal susceptibility to disinformation campaigns.

Biorisk. Several studies have claimed that open foundation models can instruct users on how to construct bioweapons. But the evidence behind these studies remains weak. In particular, studies indicating that today's language models provide "dangerous" information related to bioweapons do not acknowledge that the same information is available via Wikipedia and the National Academies of Sciences, Engineering, and Medicine. In addition, even if foundation models can be used to provide sensitive information, pathogens would still need to be developed in labs and deployed in the real world. Each of these steps requires significant expertise, equipment, and real-world lab experience. As with many other threat vectors, the best policy choke points may hence lie downstream. For example, the U.S. AI Executive Order aims to strengthen customer screening for purchasers of biological sequences. Still, there are efforts underway to measure the marginal risk of textual foundation models relative to information on the internet that will provide useful evidence for future policy proposals.

Critical research is needed on marginal risk: *To what extent do open foundation models increase risk either relative to closed foundation models or relative to pre-existing technologies (e.g., search engines)?*

Cybersecurity. Though open code models could improve the speed and quality of offensive cyberattacks, it appears that cyber defenses will also improve. For example, Google recently demonstrated that code models vastly improve the detection of vulnerabilities in open-source software. As with previous automated vulnerability-detection tools, widespread access to open models for defenders, supplemented by investment in tools for finding security vulnerabilities by companies and governments, could strengthen cybersecurity.

Spear-phishing scams. Foundation models are capable of generating high-quality spear-phishing emails. Both open and closed models could be used to produce spear-phishing emails, because the key factor that makes spear-phishing emails dangerous is the malware that accompanies the email; the text itself is usually

benign. As with disinformation, the key bottleneck for spear-phishing is not always the text of emails but downstream safeguards: Modern operating systems and browsers implement several layers of protection against such malware. Moreover, as a result of existing protections, phishing emails might not reach the recipient in the first place.

Voice-cloning scams. The last few months have seen a number of examples of real-world voice-cloning scams where malicious users impersonate a person's friends or family and successfully get them to transfer money. These impersonations rely on AI tools that can clone someone's voice based on a few seconds of audio—for instance, from their social media account. As of now, it is unclear if voice-cloning scams are more effective or scalable compared to traditional scams, especially since tens of thousands of traditional scams are already being reported to the FTC each year. Though it is yet to be determined if closed model developers can successfully prevent such scams, they do offer a measure of deterrence by, for instance, requiring users to sign up using credit cards and being able to trace any audio back to the specific user who created it.

Nonconsensual intimate imagery (NCII) and child sexual abuse materials (CSAM). Open text-to-image models appear to present unique risks related to NCII and CSAM as they substantially lower the barrier to generating such content. Safeguards for closed models are relatively more effective in this area, and monitoring closed models can deter the generating of such imagery, especially of real people. We have already seen open text-to-image models being used for creating nonconsensual deepfakes and CSAM. There remains an open question about whether policy interventions are more effective with downstream platforms, such as CivitAI and social media platforms. Organizations that are tasked with combating NCII and CSAM such as the National Center for Missing & Exploited Children may benefit from additional resources and support to address AI-generated CSAM.

Although many policy proposals and regulations do not mention open foundation models by name, they may have uneven distributive impact on open foundation models.

Governance

With policy efforts across the United States, China, European Union, U.K., and G7 focusing on foundation models, we now consider how these efforts affect open foundation models.

Although many policy proposals and regulations do not mention open foundation models by name, they may have uneven distributive impact on open foundation models. Specifically, they may impose greater compliance burdens on open foundation model developers than their closed counterparts, even when open developers are more likely to be resource-poor compared to the largest AI companies, which are disproportionately closed developers. In particular, downstream use compliance may be challenging for open foundation model developers, since they exert far less control over downstream use. We illustrate these

tensions of how proposals may disproportionately damage open foundation models.

Liability for downstream use. Since the distinction between open and closed foundation models is predicated on release, policies governing the usage of foundation models are likely to have differential impacts. Therefore, liability for harms arising from downstream usage could chill the open foundation model ecosystem by exposing open foundation model developers to severe liability risk. For example, the U.S. AI Act, introduced by Senators Richard Blumenthal and Josh Hawley, suggests potential liability for developers. In contrast, because closed foundation model developers exercise greater control over downstream use, some developers already provide liability protections for copyright to downstream users of their models.

Content provenance for downstream use. Akin to liability, if foundation model developers are required to ensure content provenance for downstream use, then these requirements may be technically infeasible for open foundation models. Given that the most salient applications of foundation models are generative AI, there is pervasive emphasis on content provenance techniques like watermarking to detect machine-generated content: The U.S. Executive Order, White House Voluntary Commitments, Canadian Voluntary Code of Conduct, Chinese generative AI regulations, and G7 Voluntary Code of Conduct all highlight content provenance. However, today's watermarking methods for language models do not persist if models are modified (e.g., fine-tuned) and require that users of a model follow certain protocols for the watermarking guarantee to hold. Fundamentally, open foundation model developers do not control how their models are modified or used to generate content. By contrast, efforts to track the provenance of trustworthy content may be more fruitful as such initiatives rely only on the participation of good-faith actors.

Liability for open data. While foundation models can be released openly without the release of the underlying data used to build the model, some developers choose to release both the model weights and the training data. Of the 10 major foundation developers assessed by the 2023 Foundation Model Transparency Index, the two developers that released data openly also released their foundation models openly. In addition, several other open foundation model developers, such as EleutherAI, the Allen Institute for Artificial Intelligence, and Together AI, tend to release data openly. However, open release of data exposes these entities to greater liability risk as exemplified by lawsuits against Stability AI on the basis of its use of LAION datasets that allegedly included plaintiffs' work. While the legality of training foundation models on copyrighted data remains unclear across many jurisdictions, the status quo poses perverse incentives. Namely, model developers that transparently disclose and openly provide data are subject to greater risk than developers that obfuscate the data they use, even if the underlying facts are identical. In light of this perverse incentive, mandated disclosure of training data may be beneficial in some cases.

Conclusion

Governments around the world are crafting different policies on foundation models: The design and implementation of these policies should consider both open and closed foundation model developers. In particular, open foundation models provide significant societal benefits in terms of the distribution of power, innovation, and transparency. While open foundation

models are conjectured to contribute to malicious uses of AI, the weakness of evidence is striking. More research is necessary to assess the marginal risk of open foundation models.

Policymakers should also consider the potential for AI regulation to have unintended consequences on the vibrant innovation ecosystem around open foundation models. When regulations directly address open foundation models, the precise definition used to identify these models and developers should be duly considered. Hinging regulation exclusively on open weights may not be appropriate given the gradient of release. Hostile actors, for instance, could leverage open data and source code—without model weights—to retrain models and generate comparable harms. And even when regulations do not directly address open foundation models, they may have an adverse impact: Liability for downstream harms and strict content provenance requirements may suppress the open foundation model ecosystem. Consequently, if policymakers are to implement such interventions, direct consultation with the open foundation model community should take place, with due consideration given to their interests.

Rishi Bommasani is the society lead at the Stanford Center for Research on Foundation Models (CRFM); a graduate student fellow at the Stanford Regulation, Evaluation, and Governance Lab (RegLab); and a PhD candidate in computer science at Stanford University.

Sayash Kapoor is a researcher at the Princeton Center for Information Technology Policy (CITP) and a PhD candidate in computer science at Princeton University.

Kevin Klyman is a researcher at Stanford CRFM and an MA candidate in international policy at Stanford University.

Shayne Longpre is a PhD candidate at the MIT Media Lab.

Ashwin Ramaswami is a JD candidate at the Georgetown University Law Center.

Daniel Zhang is the senior manager for policy initiatives at the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Marietje Schaake is the international policy director at the Stanford Cyber Policy Center and an international policy fellow at Stanford HAI.

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, professor of political science, professor of computer science (by courtesy), senior fellow at HAI and the Institute for Economic Policy Research, and director of the RegLab at Stanford University.

Arvind Narayanan is the director of Princeton CITP and a professor of computer science at Princeton University.

Percy Liang is the director of Stanford CRFM, senior fellow at HAI, and an associate professor of computer science at Stanford University.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.

The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Stanford University Human-Centered Artificial Intelligence

Center for Research on Foundation Models

CITP Princeton University

RegLab

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**