Convening stakeholders across industries to craft principles and concrete codes of practice for the development and use of artificial intelligence.

March 26, 2024

RE: NTIA, Request for Comments (RFC) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

These comments are submitted on behalf of the Alliance for Trust in AI (the Alliance), a nonprofit association of companies using artificial intelligence (AI) representing diverse sectors. Members of the Alliance seek to ensure that AI can be a trusted tool by promoting effective policy and clear codes of practice for AI. We appreciate the opportunity to respond to the National Telecommunications and Information Administration's (NTIA) Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights.

The Alliance is appreciative of the work that NTIA, and the rest of the U.S. government, is engaging in to make access to and development of AI systems safe, equitable, and innovative. In these comments, we will discuss different models for delivery of AI components, systems, tools, and products, including aspects of openness; the benefits and risks associated with this kind of access; and the role that the U.S. government should play. We hope that NTIA's report will recognize the transformative importance of open source and openness, the differences between different kinds of openness, and find more effective ways to evaluate risk for dual-use foundation models with widely available model weights.

Part of our mission is to make AI widely, and safely, available to organizations of all sizes. Considering and supporting different release models and operating models will enable more organizations to adopt and deploy AI, and it is important to ensure trust in AI across the many different technologies and delivery methods available. We encourage NTIA to keep in mind the broad diversity of technologies, contexts, and uses of AI, and to take a risk-based approach to developing guidelines and regulations for AI, including dual-use AI models with publicly available model weights. Finding nuanced, context-driven approaches to defining risk will be critical to these efforts since the risks cannot be captured by simply describing the size of the model or what elements are widely available, usable, or understood.

About the Alliance

The Alliance for Trust in AI brings together companies using advanced AI in many sectors to advocate for ways that we can build trust in all the kinds of AI that empower companies across the country and world. The Alliance works with companies developing foundational AI models, creating AI systems, and implementing these systems and models in their own work across industries.

We aim to give organizations concrete guidance around how to build AI responsibly, implement AI principles, support learning and information sharing across sectors, and establish a shared voice for the many users of AI now and in the future. The Alliance is building on work done by technologists, policymakers, and academics to create a shared understanding of how to develop and use AI responsibly. Through multi-stakeholder partnership with members across industries and sectors, the Alliance is developing definitions, principles, and codes of practice that ensure that AI is available, and trusted, to everyone.

Open Source, Public Access, and AI

Open Source AI, and AI openness, can deliver transformative benefits

Technology and the digital sector have been driven forward by a long history of openness and collaboration, and openness has the potential to underpin a similar set of advances – both in AI technologies and how they are used. The incredible acceleration in work around AI across sectors, building on decades of more retrained advances, has been spurred by the democratization of AI: making advanced generative language and image models accessible to the public has spurred acceleration in AI adoption, investment, and advancement.

Openness in AI can be as transformative as openness has been for other emerging technologies. Online and digital tools are traditionally created using many components, including open-source libraries and software (OSS). This open-source model – where anyone can inspect, modify, and reuse the code and other elements according to certain terms and conditions – have made creation and deployment of software much more accessible across all sectors, especially to small companies.[1] The vast majority of programming languages in use today have a freely available implementation available, including compilers and interpreters, and there are many freely and openly available libraries that developers can use to kickstart their own development. There are many different varieties of OSS, including open-source platforms, open-core, and licenses of all kinds. The key is that the code, libraries, systems, and software

[1] "The Value of Open Source Software - Working Paper," Harvard Business School, https://www.hbs.edu/faculty/Pages/item.aspx?num=65230

have been available as components for anyone to take and use in their own development, potentially making their own changes or commercializing them.

Releasing open source software is not the same as making model weights available, and conflating the two is counterproductive. Policies around OSS should not be directly applied or translated to openness in AI given these differences. More clearly defining these terms will be useful, and thoughtfully translating recommendations, such as those in NIST's "Software Security in Supply Chains: Open Source Software Controls" under Executive Order 14028, Improving the Nation's Cybersecurity.[2] The underlying themes of this guidance – including ensuring trustworthy sources, guardrails to mitigate vulnerabilities, due-diligence, and secure deployment – should help inform guidance on open foundation models. However, while these themes serve as a good basis for commonality, direct translation of technical specifics should be avoided given the differences in how open source is demonstrated in OSS and AI.

NTIA should use established definitions of "open source" in relation to AI, while acknowledging the benefits of openness

It is important to note that the term "open source" as it is applied to AI models is not the same as "open source" as has been applied to code and OSS. Open source in the context of AI is used colloquially as a much broader term than it has been traditionally in OSS, and applied to many situations where people can customize or build on the core models. We will discuss the benefits of other kinds of AI openness in these comments, but we urge NTIA to use the term "open source" precisely, referring to established definitions around the availability of code that can be used, modified, and redistributed. This generally should include being able to reproduce the weights and contribute back, as well as other considerations. There are a number of constructive definitions of open source AI, including from the Open Source Initiative[3] and Stanford University's Institute on Human-Centered Artificial Intelligence.[4]

Foundation models, or elements such as model weights, training data, and other assets that can be used to create or modify foundation models, can be made available in similar ways to OSS. And similarly, the code and methods used to train and test AI models can also be released for broad or public use. Creating sophisticated or foundation models with these tools is still quite challenging and resource intensive, beyond the reach of many individuals and organizations – and likely will be for some time.

---

[2] "Software Security in Supply Chains: Open Source Software Controls," NIST, https://www.nist.gov/itl/executive-order-14028-improving-nations-cybersecurity/software-security-supply-chains-open

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

[4] "Considerations for Governing Open Foundation Models," Stanford University Institute on Human-Centered Artificial Intelligence, https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

There are also other approaches to making AI development more accessible, which often take the form of delivery methods such as web applications, edge deployment, and others. Each of these approaches will have their own advantages and challenges, and the risk associated with each must be evaluated in context.

The Benefits and Risks of AI Delivery and Distribution Models

Regulation should protect the many benefits of increasing access to AI models and tools

Industries across the country have already benefited from access to advanced AI. These include healthcare, marketing, education, manufacturing, agriculture – it's difficult to think of a sector that isn't using advanced AI. Additional access to AI, especially open models that they may not be able to create themselves, will improve efficiency and competition in these sectors. Openness in foundation models, specifically, can promote competition, improve cooperation, and accelerate innovation by allowing small companies and researchers greater access to AI tools, and the elements needed to build, modify, and commercialize their own AI systems.

Self-governance and diligence should always be a part of the designing, developing, deploying, and using AI. To help users of open-source foundation models, the NIST U.S. Artificial Intelligence Safety Institute should develop resources, guidelines, and tools should that allow users to assess and address potential risks before using a specific model. Such resources can help users engage in due-diligence before working with either intentional or unintentional faulty models. This allows deployers and end users to benefit from expert guidance, and to use that guidance within context of their specific use case – which they know best.

Guidance and regulation should not artificially restrict open source AI, public access to AI, and foundational AI models

Openness should not be artificially restricted based on a misplaced belief that this will decrease risk. Each form of access to open foundation models carries different risks and has different benefits. The focus should be on ensuring that comprehensive risk management is built in by developers and deployers. Taking a more technology-agnostic approach, including relying on existing technology agnostic standards, will allow greater applicability to different instances without creating overly prescriptive standards.

Risk of AI models can be effectively managed when developers and deployers engage in due diligence and testing. This is true for all attributes of trustworthy AI and are not unique to open source or dual-use foundation models. The kinds of diligence that are necessary will change within the AI development lifecycle, and many kinds of risk mitigation will depend on the deployment use case and implementation. In many cases, only the deployers will be able to fully assess this impact.

Mitigating malicious and harmful use of available model weights are also context dependent

Availability of model weights are one element to consider in assessing risk by availability of AI - but it is not determinative. A model with widely available model weights determines the kinds of fine-tuning that can be done with it. In discussing these risks, it is important to distinguish between the risk of creation of new, malicious foundation models; the use of existing, open foundation models towards malicious or harmful ends; and the use of open foundation models in other contexts. Each of these will require different kinds of efforts to address and are context dependent.

Open (or widely available) model weights may be adapted by bad actors in ways that remove safeguards built in by the original developer. This is a risk for any model where the developer is not in control of the software, model, or product. While available model weights may make it easier to develop advanced AI, there are still significant barriers to run and modify large or advanced models. It is not clear whether the model weights themselves provide enough information to end users to significantly change what they can do or develop themselves.

Any restrictions on openness, or on delivery models, must recognize their own risks

Calls for increased regulation and restrictions around AI development are understandable given concerns around risks stemming from advanced AI. However, limitations on openness and transparency create their own risks in turn, and may not appropriately address the concerns at hand, as discussed above. Openness, and broad access to AI components, models, and weights, facilitates the ability of regulators, academics, and researchers to understand the risks and operations of AI models. A lack of access to model architectures, training data, and other core components obscures how these models truly operate, hampering efforts by researchers and watchdogs to analyze their safety and identify potential risks or biases.

Onerous restrictions also raise barriers to entry, putting cutting-edge AI tools out of reach for smaller organizations, stifling innovation. An open ecosystem allows scrutiny of powerful AI and fosters a diversity of approaches – critical for developing AI that is robust, accountable, and aligned with human values. While ensuring responsible governance is wise, diminishing openness could make AI less transparent and its societal impact more fragmented and unpredictable.

Guidelines for Open Foundation Models Must Be Risk-Based, Like Other AI Standards

Deploying AI, regardless of the deployment, carries risks. It is therefore important to have shared high-level principles for deployers to use as they evaluate, mitigate, and manage risks. This applies to AI of all kinds, not just open foundation models. Organizations across sectors

have established their own AI governance with shared goals and common approaches, and the Alliance's members have brought together a set of high-level principles around development, implementation, and use of AI based on these experiences. The principles are centered around governance, data, society, safety, and implementation. The NIST AI Risk Management Framework[5] and the work of the U.S. AI Safety Institute Consortium,[6] are also examples of constructive risk-based efforts to address risk from AI models.

We do have existing tools to regulate and govern AI, including advanced AI and foundation models. Technology-neutral laws and regulations apply, and many sectors (e.g., financial, health, and employment) have advanced programs to ensure compliance and manage potential unintended bias. As NTIA works to develop this report around the risks and benefits of open source AI and foundation AI models, the Alliance hopes that you will look to the sectors that have already developed similar practices for risk governance by necessity. While each sector approaches these questions from different perspectives, there are many commonalities that can help inform NTIA's work, and the broader work around risk management across the U.S. government.

Guidance should not be based on delivery mechanisms; instead, risk-based standards and regulations can help ensure that undue burdens are not placed on developers, stifling the open source foundation model ecosystem. Finding nuanced, context-driven approaches to defining risk will be critical to efforts to manage risks going forward. The risks cannot be captured by simply describing the size of the model or what elements are widely available, usable, or understood. Instead, risk must be understood as the likelihood and impact of the use (or non-use) of a given kind of AI, within a given context.

Policy, guidance, and standards for AI should not use simple proxies for risk

Any risk management for AI will necessarily include open foundation models. When possible, the guidelines, risk measurement, and standards for AI should be applicable to all types of AI. Assessing the risk of open foundation models, dual-use foundation models, and publicly available model weights must go beyond simply assessing their openness, and instead address risk in the context of the use of the AI model. Policy, guidance, and standards should not demonize particular development, distribution, or business strategies. Instead, guidance should be broadly applicable and achievable by organizations of many sizes and with diverse approaches to AI to support existing diversities in business models.

Rules for dual-use foundation models, including open source models or models with widely available model weights, should be carefully scoped and targeted to address concrete risks,

[5] "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," U.S. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[6] U.S. AI Safety Institute, U.S. National Institute of Standards and Technology, https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

rather than hypothetical ones. Standards should not create undue or asymmetric burdens on the developers of these open foundation models. Most risks from open foundation models occur through downstream uses by third-parties, which developers cannot control; therefore, heavily regulating the developers and their models will increase regulatory burdens without meaningfully reducing risk. This in turn will likely reduce the number of open-source foundation models available to companies and researchers, slowing the pace of innovation and development of AI for good.

In assessing concrete risks, neither the size of a foundation model nor the number or people or organizations with access to a particular kind of AI is a good proxy for risk. While we appreciate that the AI Executive Order put a threshold for dual-use foundation models in place, it will be important to find risk measurements that are more meaningful than the size of a model. These factors may contribute to the scope of the risk but are not inherent risks themselves. A small model accessed by ten people could carry more risk than a large model accessed by ten thousand people depending on how it is implemented, for what purpose, and under what context. The Alliance urges NTIA to not exclusively rely on these factors as a determinant of risk.

Additionally, we caution against any overly simplistic or prescriptive definitions as we differentiate between risks and benefits from different approaches. Terms like “open” and “openness” have been the topic of significant, animated discussion within the OSS community for decades. Where fine distinctions in how AI is created or delivered are not meaningful to the kinds of risk that should be considered, avoiding these will be helpful to achieving the goals within the AI Executive Order.

International Collaboration

In developing best practices and guidelines for foundation model safety and security, the U.S. should collaborate broadly with national and international standards organizations. International alignment on risk management will help ensure that all jurisdictions take the same kind of care and create guidance that works together, instead of fragmenting the ecosystem. This in turn will ensure a greater secure openness of AI. Given broad multistakeholder support for NIST’s AI RMF, we suggest that this forms the basis of these international collaborations. The work of the U.S. AI Safety Institute will be to provide international policymakers with information and technical tools for regulations.

In international collaborations, it will be important to recognize the global and decentralized nature of open source. This makes it difficult to enforce requirements on open source projects. Instead, it will be more effective to ensure that open source projects have the necessary guidance and use rules on the deployment of AI systems to create the necessary incentives and pressures within the ecosystem to encourage developers and deployers to adopt that guidance.

Conclusion

Democratizing AI and ensuring wider availability is beneficial for everyone. When researchers and organizations have greater access to different AI models, they are more likely to adopt and innovate - and researchers can test and evaluate these same models. Making sure that AI is trusted is a core element of getting more organizations comfortable with using powerful AI.

Thank you for the opportunity to comment on these questions. If you have questions, or believe that we can be helpful to your work in any way, please contact the Alliance's coordinator Heather West at hewest@venable.com.