moz://a

Mozilla's Comments on the NIST AI 800-1 IPD

About moz://a

Mozilla's mission is to ensure the internet is a global public resource, open and accessible to all. An internet that truly puts people first, where individuals can shape their own experience and are empowered, safe, and independent.

Founded as a community open source project in 1998, Mozilla consists of several organizations, most notably the non-profit Mozilla Foundation, which leads our movement-building work, and its wholly owned subsidiary, the Mozilla Corporation, which leads our market-based work, including the development of the Firefox web browser. They work in close concert with each other and a global community of tens of thousands of volunteers under the single banner: Mozilla.

For the past five years, Mozilla has been committed to advancing trustworthy AI. Mozilla published a paper in early 2024, Accelerating Progress Toward Trustworthy AI, that outlines how Mozilla and its allies are advancing openness, competition, and accountability in AI. Mozilla is putting its resources behind these priorities as well: The Mozilla Foundation has been dedicating 100% of its $30M a year budget to philanthropic activities, advocacy, and programmatic work on this topic. Mozilla is also investing another $30M in research and development on trustworthy AI via Mozilla.ai, as well as $35M in responsible tech startups — including startups with a focus on trustworthy AI — through Mozilla Ventures and the Mozilla Builders accelerator program. On the frontlines of modern AI practices, Mozilla freely provides an open-source, large-language model (LLM) AI model deployment system for local use and empowers more people to enhance the safety of models online through the 0Din bug bounty program.

As an independent and mission-driven organization, Mozilla is committed to working with regulators to develop effective policies that ensure that innovation and growth in AI serve the public interest. We put people above profit.

Executive Summary

Mozilla has been on the frontline of defending the open internet for 25 years. Our history is deeply intertwined with that of the open source movement, and we believe that this same openness is increasingly important for AI as its adoption and development increase exponentially. Increased concentration and centralization in the AI market, favoring the construction of walled gardens, threaten the widespread benefit that AI can provide to the world.

Yes, openly available models come with risks and vulnerabilities — AI models can be abused by malicious actors or deployed by ill-equipped developers. However, we have seen time and time again that the same holds true for proprietary technologies — and that increasing public access and scrutiny makes technology safer, not more dangerous. The idea that tight and proprietary control of foundational AI models is the only path to protecting us from society-scale harm is naive at best, dangerous at worst.

The following list highlights key feedback from this document:

1. The current draft focuses on AI services deployed on the internet and accessed through some interface or API. The reality is that the majority of AI research and development is occurring on locally deployed AI models that are collaboratively developed and freely distributed. NIST should rework the draft's front matter and glossary to better capture the state of the AI ecosystem.
2. The practices outlined in the draft place a disproportionate burden on any AI developer outside of the small handful of very large AI companies. Mozilla believes that NIST should ensure that requirements are applicable to organizations of all sizes and capability levels, and should take into account the potential negative impact of misuse at different organizational scales.
3. The recommendations for implementing the practices outlined in the draft imply that the AI model is centrally controlled and deployed. Open-source and collaborative development environments don't align with this approach, rendering this guidance inapplicable, unhelpful, or at worst - harmful. Given the strong evidentiary basis for open-source helping mitigate risk and make software safer, NIST should ensure open-source AI is considered and supported in its work.
4. Define "gradients of access" as a way to provide a framework for AI risk management discussions and decision making. These gradients should represent incremental steps of access to an AI model (e.g. chat interface, training, direct weights visibility, local download, etc.) and each should be accompanied by its associated risks.

Our Feedback

Mozilla is committed to advancing the development of trustworthy AI around the globe and to shifting the norms and incentives governing the AI ecosystem. Guided by its Manifesto and the

vision formulated in the 2020 white paper Creating Trustworthy AI and its successor, Accelerating Progress Toward Trustworthy AI, Mozilla builds products and services like llamafile, conducts original research, funds people, startups, and initiatives, and carries out advocacy work in pursuit of these goals.

We appreciate the time and care NIST has invested in developing the objectives and practices presented in the draft and are pleased to offer our feedback and ideas on how to ensure the final publication is as broadly successful as possible. Openness and knowledge-sharing can be key building blocks of any strategy aimed at driving progress across the AI industry and the U.S. economy more broadly — not just in a handful of well-resourced big tech research and development entities. We hope our comments can help NIST develop a framework to improve the safety, security, and trustworthiness of dual-use foundation models, without regard to whether they are developed by a large corporation, an open source project, or an academic institution. We hope that NIST will revise the current draft in order to ensure that it applies to the AI ecosystem more broadly, not just a few companies centralizing AI development within their own closed ecosystems.

To that end, our feedback is organized into three categories:

1. Comments on the overall approach, scope, and applicability of the draft, particularly where it concerns open-source AI models,
2. Feedback on Glossary terms and definitions used throughout the document, and
3. Granular, lower-level input on individual objectives and practices.

General Document Comments

Mozilla believes in the power of collaborative and open-source research and development, and the AI ecosystem is no exception. Over the last several years, open-source foundation models, in many cases rivaling the performance of closed models from large tech companies, have become increasingly common for little or no cost on the internet, built and maintained by a diverse group of contributors. These models are not deployed as a service, but rather provided freely for users to download, modify, and deploy locally for their own use cases. Loosely grouped as "open-source models", these typically have publicly viewable parameter weights, are free to download, and have generally permissive licenses. Our comments to NTIA earlier this year have additional material about the definitions of open source AI and actions that we should undertake to support the development of diverse AI components, tools, and systems.

These open source AI models, and others like them, differ significantly in their management, deployment, licensing, post-deployment monitoring, and monetization. This diversity is a strength of the open source movement but can necessitate different practices and governance, especially when compared to development that occurs within the more rigid structures of a single corporation. Despite this, the current draft seems to consider only discrete AI models created by large centralized organizations, with tightly controlled deployments and highly restrictive

licensing. Mozilla believes that this shortcoming, if addressed, would help ensure that AI 800-1 is a broadly applicable set of best practices with long-lived relevancy. Many, but not all, of the following feedback focuses on concrete ways to close the gap on the draft's coverage of open-source AI models.

Below is an unordered list of comments on the overall approach, scope, and applicability of the draft.

Recognize the prevalence and importance of open-source AI models

We suggest that NIST revise this draft to recognize the importance of different and diverse models for development and delivery of advanced AI. While a significant portion of the current development of dual-use foundation models is closed and happens within large corporations, a strong and growing community of open source developers are working to create open source AI components, models, and tools. As briefly discussed above, open-source models play an important role in the AI ecosystem, and are not fully acknowledged or discussed in the current draft. In both front matter and the most granular practices, care should be taken to discuss the differences in open-source models and how that may impact or change the associated practices and recommendations.

Ahead of the United Kingdom's AI Safety Summit late last year, Mozilla organized a joint statement with over 1,800 signatories that emphasized how openness is a boon to AI safety and security. Further, in February 2024, Mozilla and the Columbia University Institute of Global Politics brought together over 40 leading scholars and practitioners working on openness and AI to explore what 'open' should mean in the AI era. And this March, Mozilla, the Center for Democracy and Technology, and a wide-ranging group of civil society organizations and academic experts wrote to Secretary Raimondo to emphasize the importance of openness in AI.

Openness in AI is vital for regulators and civil society to be able to assess AI systems and other components used in AI development to ensure they appropriately conform to all applicable laws and regulations as well as broader concerns around safety and bias. Incentivizing AI developers to work at the largest companies and behind closed doors carries a very real risk for society.

Open-source model developers could be negatively impacted if this document places undue burden on them, which could cascade into an overall loss in efficacy and harm safety efforts for the nascent AI ecosystem.

Clarify obligations across the AI lifecycle with a responsibility model

The current draft assumes that the AI model developer is a single entity, or at least a set of entities under common control, that develop a series of models that build upon each other and that keeps institutional knowledge about the risks and mitigations from those models. The misuse risk management obligations in the draft assigned to the model developer could make sense in that context, but become muddied or inapplicable when interpreted through an open-source or distributed development lens. Additionally, the draft does not contemplate the responsibility of developers to openly share their learnings.

Open-source model developers may freely release a foundation model with a permissive license, as has traditionally happened with open source development. Downstream users or developers can duplicate this model and operate it entirely locally to their own environment, or modify it to suit their needs. The original developer has no insight or control over this secondary deployment, and in most cases, has a clear legal separation from any liabilities that may follow.

In practice, open-source models often go through many such handoffs, forking off into branches and spinning into new projects with modified objectives and licenses. In addition, the model may be deployed with additional transformative components sourced from entirely different supply chains. It is clear that a single set of obligations **cannot** feasibly be applied equally to all parties in AI development.

Mozilla believes this concern would best be addressed by providing a clear definition of "open-source" models (such as the one provided by OSI), and by providing additional scoping language to each practice that targets specific actors or groups of actors in the AI lifecycle. At a minimum, the draft should discuss the complexities and nuances present in the different environments in which AI models are developed.

For further discussion on this point, Mozilla has authored an extensive look into the current state of the AI lifecycle and its implications in our recent response to the NTIA RFC.

Recognize the full AI stack and components for AI models

The document focuses risk management practices on the evaluation of a single large model treated as a discrete object, but the AI ecosystem is much more complex than that. Models, especially models developed within an open source community rather than by a single developer, are made up of numerous components, typically with different contributors and often with different licenses or requirements. In many cases, additional components are added post-deployment by users of the model, outside the purview of the original developer; it is not likely that the original developer or community of developers will have insights into these additions or changes. This decentralized

innovation has been critical for the development of our current technology ecosystem and has proven to not only provide myriad economic benefits but increase software safety and security.

As multi-modal and multi-model configurations become increasingly feasible to operate on less expensive hardware, AI "in the wild" is often several discrete AI components running either in tandem or sequentially, and there is no reason to believe that these models will be developed by the same entities.

NIST should clarify how this impacts overall applicability of the document (e.g. do four four-billion parameter models running in tandem constitute a dual-use Foundation model?) as well as how it impacts or adjusts specific practices.

This multi-dimensional approach to understanding the constituent components of the AI stack is explored in more detail in Mozilla's Policy Readout for the Columbia Convening on Openness and AI.

Recognize that openness and risk management do not need to be in tension

Throughout the draft objectives, there's a misleading tension between the need for openness in AI development and the requirement to manage misuse risks. Risk management practices need not stifle the rapid innovation common in open-source development.

In order to accomplish this, NIST should consider defining "gradients of access". These gradients should represent incremental steps of access to an AI model and the full relevant AI tech stack (e.g. chat interface, training, direct weights visibility, etc.) and each should be accompanied by its associated risks. Mozilla believes that these gradients would provide a useful context from which AI risk could be more easily discussed.

There is risk in concentrating cutting-edge research in ever-fewer research labs. Openness has, and will continue to, foster innovation - a broader risk not contemplated in this draft. Openness in AI can spur competition and help the diffusion of innovation and its benefits more broadly across the economy and society as a whole. NIST should provide clearer guidance on how to balance these competing needs, recognizing the unique value that openness brings to AI development and safety.

NIST should provide guidance on how to implement risk management practices that support rather than hinder innovation in open source AI and enhance existing open source practices which create safer and less biased AI solutions. In order to be successful, this draft needs a deliberate and well-considered approach to AI policy and openness that rewards collaboration and knowledge sharing as well as effective risk management.

Recognize the value of open scrutiny

NIST should acknowledge how open development practices can contribute to ongoing risk assessment and management for organizations. Openness in AI is vital for regulators and civil society to be able to assess AI systems and associated components. Transparency and open scrutiny can help ensure that they appropriately conform to all applicable laws and regulations as well as broader concerns around safety and bias.

While the document emphasizes the importance of transparency, it could go further in leveraging the inherent accountability of open source development. Openness in the AI ecosystem can help identify and mitigate risks and bring more scrutiny to both open and proprietary AI. NIST should provide guidance on how open source projects can leverage their inherent transparency to enhance accountability and trust, and how other projects can work with the open source community towards those same goals.

The document could leverage the collaborative nature of open source development in addressing safety and security concerns. Research drawing on open-source AI has helped advance red-teaming and safety alignment work, measuring bias and mitigating toxicity. NIST should encourage and provide frameworks for community-driven safety and security practices. Openness increases the availability of the tools that regulators need to monitor and evaluate (large-scale) AI systems and helps to create a diverse community who can lend their expertise to red-teaming and evaluating AI systems.

Acknowledge resource disparities between model developers

Well-resourced multinational corporations engaging in AI model development are better positioned to comply with certain obligations than small non-profit or academic research labs, community-driven projects, or early-stage start-ups. These smaller or decentralized projects frequently run up against restrictive resource constraints. In order to avoid negatively impacting the important work taking place outside large corporations and labs and to increase the likelihood of these objectives being widely adhered to, NIST should consider providing scalable approaches and tiered requirements based on project size or impact, especially for smaller models that may still have impacts for public safety. We acknowledge that tiered recommendations may introduce additional risk, but believe that accepting the reality of the AI ecosystem and addressing it directly is a more efficient long-term risk management strategy that brings developers across the AI ecosystem together towards the same goals, rather than alienating smaller or decentralized development efforts.

Take advantage of existing open source risk management practices

While AI represents a new frontier of computing technology and does require some special risk management considerations, it is also true that AI models share many properties with "classical" software. This is particularly true in the open-source AI ecosystem, where AI models are often distributed, downloaded, then locally executed by the user.

In this context, it would seem wise to lean on the extensive body of research that has gone into managing misuse risk in traditional open-source software. In Mozilla's comment submission to the NTIA, we discuss the importance of lessons learned from past open-source development, and make clear that we needn't start from scratch when it comes to protecting and growing open-source AI development.

Addressing global implications

The document could better address the global nature of open source AI development and its implications for managing misuse risks. NIST should consider providing guidance on managing misuse risks in the context of global, collaborative open source AI development, which happens to be where significant AI innovation is occurring today. Working collaboratively with other standards organizations and AI safety institutes to ensure consistent risk management would be helpful for the ecosystem. While this may not be appropriate to include in updates to this draft, it is an important aspect to keep in mind in the development of this guidance.

Glossary and Language Feedback

Key terms and concepts need clarification for open-source contexts. NIST should refine its glossary and key concepts to better include the realities of open-source AI development. Generally, these definitions don't take into account the open-source AI ecosystem, and run the risk of ambiguous interpretations.

1. **Dual-Use Foundation Model**: The current definition, as provided by Executive Order 14110, focuses on models with "at least tens of billions of parameters." However, it also includes models that could have high performance at tasks that "pose a serious risk to security, national economic security, national public health or safety" - which is a very broad set of considerations, and not consistent with the number of parameters in a model. Some configurations of AI models (e.g. multi-modal and multi-model approaches, increasingly common in open-source models) use lower parameter counts but can still achieve high performance at particular tasks. As it seems that the definition is potentially ambiguous,

NIST should consider augmenting the EO's definition with additional clarity around the definition of models intended to be targeted by this draft in order to provide concrete guidance on where the draft is best equipped to help manage risk.

2. **Distribution Channel**: The current definition is quite broad and includes various ways a model could be distributed, but does not contemplate open approaches to distribution of these models. NIST should consider explicitly mentioning open-source repositories and community-driven distribution platforms to acknowledge their importance in the AI ecosystem, and ensure that they are clearly included.
3. **Jailbreaking**: The current definition focuses on causing a model to act contrary to its designer's intentions. NIST should acknowledge that in open-source contexts, "jailbreaking" might have different implications, as model adaptability is often a feature, not a bug. We suggest a definition focused on violation of the model's terms of service or license, where the model developer is able to clearly define the model's intended uses.
4. **Misuse Risk**: While the current definition is straightforward, NIST should consider expanding to acknowledge that in open contexts, the line between use and misuse can sometimes be blurry and community-defined. As above, we suggest a definition focused on violation of the model's terms of service or license. In addition, we believe that "misuse risk" should be defined by the marginal risk of AI models compared to existing technologies and information, not the absolute risk. Mozilla notes the importance of taking stock of the marginal risk of this technology relative to information already readily available elsewhere.
5. **Model Theft**: The current definition focuses on unauthorized access, and this concept may need significant rethinking for open source models where "theft" is less applicable. It may be worth considering whether model theft is even applicable as a concept for open source, especially for projects where most or all resources for the creation of a model are open. NIST should consider replacing this with a more descriptive term, such as "Unauthorized Model Replication" or "Malicious Model Adaptation" to accommodate these open source contexts. As above, we suggest a definition focused on violation of the model's terms of service or license.

Feedback on Specific Objectives and Practices

Objective 1: Anticipate potential misuse risk

Practice 1.1

Provide threat profile identification methodologies and guidance: Beyond "consulting external experts", the draft does not provide any guidance on the mechanics of actually identifying a threat profile. NIST should promote transparent, reproducible threat identification methodologies that align with open-source principles. Where possible, connections or references should be made to

existing work (e.g. the NIST AI RMF). NIST should encourage the development and use of open-source tools for risk assessment and management. Additionally, the draft should acknowledge that open source developers may have significantly less information about the goals, intentions, and practices of deployers and end users.

Practice 1.2

Provide threat profile assessment methodologies and guidance: The draft does not provide any guidance on the mechanics of actually assessing a threat profile. NIST should promote transparent, reproducible risk assessment methodologies that align with open-source principles. Where possible, connections or references should be made to existing work (e.g. the NIST AI RMF). NIST should encourage the development and use of open-source tools for risk assessment and management, and take advantage of the collaborative, cooperative, and open norms of open source communities. Additionally, this risk assessment should depend on marginal, rather than absolute, risk from the model.

Practice 1.3

Address pre-development assessment for open source and community projects: Provide specific guidance on how community-driven open source projects can estimate capabilities and risks before development, especially given open or decentralized development patterns and workflows.

Promote openness, transparency, and information sharing: This practice relies on AI model developers having access to the results of analyses of other AI models. This is difficult or impossible if other entities do not share the results of capability analyses of their models. NIST should take a stronger and more explicit stance on recommending that these results be shared openly and freely. Many projects will not have significant previous context for the open source model or components that they may be modifying, making it more important to have this kind of guidance and information about AI model analysis.

Objective 2: Establish plans for managing misuse risk

Practice 2.1 & Practice 2.2

Recognize diverse governance models: NIST needs to speak to the variety of governance structures in open source projects when discussing risk management plans. The 'open source' ecosystem is no monolithic actor, be it in AI or traditional software. Its members comprise individual volunteer contributors, small organizations, and multinational corporations. The appropriate governance for these projects will be similarly diverse, and risk management will vary based on those governance models. Within those appropriate governance models, NIST should provide guidance on how open source communities can collectively determine acceptable levels of misuse risk, in line with Objective 7.

Promote diverse risk mitigation strategies: NIST should encourage the development and use of alternative strategies for managing misuse risk that don't rely solely on confidentiality, which may be more suitable for open source projects. This is an important part of creating an incentive structure that promotes openness and knowledge-sharing. Mozilla believes that secrecy should not be the default mitigation strategy, and the draft should provide a discussion of the pros and cons of such an approach.

Objective 3: Manage the risks of model theft

Mozilla feels that this objective is fundamentally misaligned with the open-source AI ecosystem, and makes the following recommendations at the objective level, instead of the practice level. This misalignment may be more broadly addressed by creating a more granular responsibility model for different roles.

Clarify "model theft" in open contexts: NIST needs to provide a clearer term than "model theft" in the context of openly available models and components. While the idea of model theft may be an appropriate concept for closed development, it is not an appropriate description of the risk intended by this section. It is not clear that model theft is even applicable as a concept for open source. NIST should consider replacing this with a more descriptive term, such as "Unauthorized Model Replication" or "Malicious Model Adaptation" to accommodate open source contexts, and rework this objective in order to more directly and clearly address the contemplated risk.

Address tension with open source principles: NIST should acknowledge and provide guidance on navigating the inherent tension between preventing model theft as a core risk management goal and the open source principle of freely sharing some or all code, components, and model weights. Restrictions on the sharing of model weights or other AI components could threaten the integrity of open science and scientific dialogue, and hinder the diffusion of progress in AI. Provide guidance on striking a balance between securing models against theft and maintaining accessibility for legitimate research and development purposes.

Consider "gradients of access": NIST should consider more measured guidance on managing risks across different levels of model access, from API-only to full weight release, which are common in open source AI projects. A 'gradient' of how AI components — particularly model weights — can be accessed: from providing no outside access to only providing access via interfaces and hosted API access to making the component available for download. See General Document Feedback for a further exploration of this concept.

Consider federated approaches: Provide guidance on, or discuss, federated learning and other distributed approaches that can help manage the risks of model theft while preserving some degree of openness.

Objective 4: Measure the risk of misuse

Practice 4.1

Encourage open benchmarks, evaluation metrics, and datasets: Promote the development and use of open, standardized benchmarks and evaluation metrics for measuring misuse risk, not just proxy models or similar previous models.

Promote openness, transparency, and information sharing: The use of "proxy model" assessment relies on AI model developers having access to the results of analyses of other AI models. This is difficult or impossible if other entities do not share the results of capability analyses of their models. NIST should take a stronger and more explicit stance on recommending that these results be shared openly and freely.

Address challenges in measuring risks of model components: Provide guidance on measuring risks associated with individual model components or techniques, which are often shared separately in open source contexts.

Allow for developer communities to determine "misuse" in their specific context: Open-source AI models are not monolithically controlled, and therefore what constitutes misuse cannot be cleanly declared. NIST should provide guidance on how to best determine and decide what misuse entails in these cases.

Practice 4.2

Encourage community-driven red teaming: NIST should promote community-driven red teaming efforts as a valuable approach to measuring misuse risk in open source contexts. Research relying on openly available AI components has been instrumental to advancing safety, security, and trustworthiness across the entire AI ecosystem, and should be strongly encouraged.

Objective 5: Ensure that misuse risk is managed before deploying foundation models

Mozilla feels that this objective is fundamentally misaligned with the open-source AI ecosystem, and makes the following recommendations at the objective level, instead of the practice level. This misalignment may be more broadly addressed by creating a more granular responsibility model for different roles.

Clarify "deployment" in open source contexts: Provide a clear definition of what constitutes "deployment" in the context of open source AI development, where code and model weights may be publicly available throughout the development process. In this context, deployment of a model is usually executed by an entity wholly unassociated with the model developer.

Consider community-driven safeguards: Provide guidance on how open source communities can collectively develop and implement safeguards against misuse. Support the open-source AI community in developing norms and practices around responsibly developing and openly releasing AI models and components.

Address challenges of continuous deployment: Offer strategies for managing misuse risks in continuous integration/continuous deployment environments common in open-source development. These risks overlap heavily (or perhaps entirely) with the risks present in traditional open-source development environments (e.g. supply chain attacks).

Address Challenges in Determining Risk Tolerance: Offer guidance on how open source projects, particularly those without a traditional organizational structure, can determine and apply risk tolerance thresholds. Discuss diversity within the open source ecosystem and the need for nuanced approaches.

Objective 6: Collect and respond to information about misuse after deployment

Mozilla feels that this objective is fundamentally misaligned with the open-source AI ecosystem, and makes the following recommendations at the objective level, instead of the practice level. This misalignment may be more broadly addressed by creating a more granular responsibility model for different roles.

Recognize that locally deployed models cannot be monitored: When an AI model is downloaded and executed locally, the original developer no longer has visibility into its operation. In such a case, data collection would likely be considered a violation of privacy or be otherwise undesirable. NIST should discuss the nuances present in information collection in this and other similarly disjointed responsibility models.

Promote open datasets of misuse incidents: Encourage the creation and maintenance of open datasets documenting significant AI misuse incidents. Promote mechanisms for sharing information about misuse incidents and effective responses across different open source AI projects. Promote the development and publication of clear protocols for responding to reported misuse in open source projects.

Support collaborative mitigation strategies: Provide guidance on developing and sharing mitigation strategies for identified misuse risks across the open source community. Collaborate on developing and sharing effective safeguards and monitoring tools. Encourage the development and use of open-source tools for detecting and analyzing potential AI misuse. Invest in and provide resources for the development and maintenance of open-source AI.

Address challenges in monitoring forks and derivatives: Provide guidance on any responsibility for monitoring and responding to potential misuse in forked or derivative models, which are common in open source contexts.

Promote responsible disclosure practices and transparency in misuse reporting: Encourage regular, public reporting on misuse incidents and responses, while respecting privacy and security concerns. These practices will likely resemble, but not duplicate, existing incident responsible disclosure best practices.

Objective 7: Provide appropriate transparency about misuse risk

Mozilla appreciates the effort NIST has taken to include transparency concerns in the draft, and encourages the authors to strengthen the points made under this objective. The importance of openness cannot be understated, and Mozilla feels that the draft could be improved by replicating similar points through the rest of the document.

Expand examples of implementation documentation: The document currently only provides a sparse list of example documentation that demonstrates implementation of this objective. Mozilla recommends the addition of further guidance on transparency, such as those outlined in CDT's Best Practices in AI Documentation or the output of Mozilla's Open-source Audit Tooling Project (OAT).

Recognize diverse transparency needs: Acknowledge that transparency requirements may differ for various types of open source projects and provide flexible guidance accordingly. "The open-source ecosystem is no monolithic actor and includes diverse participants from individual volunteer contributors to multinational corporations.

Encourage open documentation practices: Promote comprehensive, publicly accessible documentation of model development processes, risk assessments, and mitigation strategies. Openness goes beyond the release of technical and non-technical artifacts and includes documentation of development and decisions.

Support open model cards and datasheets: Encourage the development and use of standardized, open model cards and datasheets that include information about potential misuse risks to provide transparency into how they are managing these risks.

Promote transparency in governance structures: Encourage clear communication about the governance structures and decision-making processes in open source AI projects.

Final Thoughts

Mozilla wants to once again applaud NIST for its excellent work in advancing the safety and security of the burgeoning world of AI applications. The current draft represents a significant step forward in defining best practices and lays the groundwork for further research and investment in mitigating AI misuse risks.

While we believe that the work that has gone into this draft is representative of AI discourse heading in the correct direction, we feel that it is paramount that openness, transparency, and recognition of the open-source community are given more focus in the next version of the draft. To achieve that goal Mozilla recommends that NIST:

- differentiate between as-a-service AI model deployments and locally deployed AI model instances throughout the document, or at a minimum take more care in defining the differences in applicability of requirements between these two classes of deployments,

- take into account the potential disproportionate burden these practices would place on smaller AI developers,

- and provide additional clarity on how these objectives can be reached in open-source or collaborative development environments.

We hope that NIST considers the impact that this draft could have on open-source AI model developers, and addresses the gaps identified above in a future version of the draft. Mozilla thanks NIST for the opportunity to provide our feedback, and is honored to be part of this process. We look forward to working with NIST and the OS AI community to advance such practices, and we will be hosting Columbia Convening 2.0 to focus on this important work in late 2024.