moz://a

Mozilla's comments on the NIST AI Risk Management Framework

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, Mozilla conducts original research, funds people and initiatives, builds solutions, and carries out advocacy work in pursuit of these goals.

We appreciate the time and care NIST has invested in developing the AI Risk Management Framework (RMF) and are pleased to offer our feedback and ideas on how to further strengthen the framework.

Our submission focuses on the following aspects:

1. Taking a **comprehensive approach to managing AI-related risks**
2. Accounting for **upstream risks in data collection and curation**
3. Ensuring **accountability across the AI supply chain**
4. Considering the **importance of systemic transparency**
5. Providing guidance on how to enable **broad and diverse participation and input**

1. The RMF outlines a thoughtful and comprehensive approach to assessing and managing AI-related risks

We welcome that the draft RMF accounts for risks across the lifecycle of AI systems and takes an approach that looks beyond purely technical considerations. Instead, it also addresses important socio-technical aspects and the principles of fairness, accountability, and transparency. Taking such an approach is critical given that risks emanating from the use of AI are highly contextual: Not only do they depend on technical parameters and design decisions, but also on the exact purpose for which they are used as well as the (organizational and social) context of deployment. Further, we are pleased to see that the RMF considers a wide range of stakeholders, including external auditors, civil society, and affected individuals and communities. As we have argued in our previous comments on the RMF to NIST, meaningful involvement of individuals and communities in particular—as well as organizations representing their interests—should not be treated as optional. Instead, it should be considered a critical building block of trustworthy AI and occur throughout the lifecycle of an AI system. Doing so will both lead to fairer outcomes and enhance trust.

Additionally, the RMF is right to acknowledge the role independent third parties can play in evaluating AI systems and assessing risk. It's therefore important to also develop the necessary tools and processes to carry out such independent assessments. This is why, for example, Mozilla supports work to develop an open-source toolkit for algorithmic audits by Mozilla fellow Deborah Raji. At the same time, organizations should also consider novel and innovative approaches to identifying risks and potential harms. For example, we have recently seen experimentation with "bug bounties" (or "bias bounties")—an idea originally focused on the identification of cybersecurity vulnerabilities—in the context of AI, for example by Twitter. In a report for the Algorithmic Justice League, Kenway et al. provide a valuable overview of this emerging approach.

2. The RMF should account for upstream risks in data collection and curation

The RMF considers data-related risks in the pre-design stage, but it appears to focus on issues of data availability, representativeness, and suitability. However, it should also account for potential upstream risks and harms that can arise as a function of how, by whom, and for what purpose data is collected and curated.

In addition to aspects such as data quality and representational harms or bias, it is important to also consider, for example, legal concerns, data protection, and the human

labor that goes into collecting and annotating data. This includes asking questions such as: Was data collected with consent from data subjects? Do data subjects have the ability to revoke consent? Who was tasked with labeling or annotating data and under what working conditions (e.g., with regard to compensation)?

Paullada et al. provide a wide-ranging survey of important issues in this context. Further, Gebru et al. also point to the importance of such considerations and how these can be incorporated in dataset documentation in their seminal work on datasheets, now widely recognized as an example of good practice in the AI development process.

These are concerns Mozilla seeks to heed in its own work as well. For instance, in our work on Common Voice—a crowdsourced open-source voice dataset—we are working hard to responsibly steward collected data and to respect the interest of those individuals and communities from whom it is collected. Further, Mozilla's Data Futures Lab is incubating approaches to data governance that give greater control and agency to individuals or collectives, for example in the form of data cooperatives, or that enable better stewardship.

3. The RMF should ensure accountability across the AI supply chain

The draft RMF highlights that it's important to hold AI systems' (human) operators and their organizations accountable for risks and adverse impacts caused by these systems. However, as the RMF rightly points out, risk needs to be managed across an AI system's lifecycle and therefore across the entire supply chain.

As we have recently argued in our position on the EU's proposed AI Act, it is important to effectively allocate responsibility and accountability along the AI supply chain. Risk depends on the intended purpose of an AI system and its context of deployment, all of which should be duly considered by operators. At the same time, risk can also be rooted in an AI system's technical design and other factors that largely fall within the responsibility of developers.

Therefore, the final RMF should provide guidance and more clarity on how it applies to, most notably, developers and operators/deployers of AI systems and on which aspects are especially important to consider for each actor along the supply chain.

4. The RMF should also consider the importance of systemic transparency

In its definition of transparency, the RMF focuses exclusively on end-user facing transparency, stating that "[t]ransparency reflects the extent to which information is available to a user when interacting with an AI system" (p. 13). While this is important,

the RMF should also consider the importance of transparency at different levels and vis-à-vis different stakeholders.

For instance, this notion of transparency omits individuals who are directly affected by an AI system's output but do not directly interact with the system (but, for example, with an intermediary). Further, it raises the question of what information should be made available to third-party auditors so that they can effectively assess risks and potential harms, as discussed above. Additionally, effective risk assessment and mitigation might even require public-facing transparency, that is, disclosing information about an AI system to, for example, independent researchers or civil society organizations. This could include information about the model, its optimization goals, or the data used to train and evaluate it. As mentioned in our previous submission to NIST, such public disclosure can enable outside stakeholders to investigate (potential) patterns of discrimination or harm.

Mozilla's own research and advocacy work underlines the value that public-facing, systemic transparency can bring. This is particularly the case where AI systems operate at large scale and have the potential to cause harm not only to individuals or communities, but to society as a whole—like in the case of social media or content sharing platforms' recommendation engines. For example, Mozilla's YouTube Regrets research used a crowdsourced dataset to find that YouTube frequently recommends videos that violate its own policies—and particularly so in non-English speaking countries. Other research by Mozilla fellow Odanga Madung found that a foreign political organization spread disinformation and inflammatory rhetoric around reproductive rights reforms, and that Twitter amplified this through its trending topics feature. Disclosing more information about how these recommendation systems work and enabling researchers and watchdogs to better study these systems and their impacts could go a long way in identifying and mitigating harm.

Further, in our position on the EU' proposed AI Act, we endorse the proposed database in which "high-risk" AI systems would need to be registered prior to deployment. While such a database cannot be part of the RMF, transparency mechanisms like it can serve as inspiration in thinking about how the framework could enable or interact with mechanisms aimed at creating systemic transparency.

5. The RMF should provide guidance on how to enable broad and diverse participation and input

As discussed above and in our previous submission to NIST, it is a step in the right direction that the RMF highlights the importance of involving outside stakeholders, and particularly affected individuals and communities, throughout the AI lifecycle. However,

the RMF could go further by providing those following the RMF with guidance on how input from such stakeholders can be gathered and how meaningful participation can be enabled. This isn't an easy task and Mozilla itself is grappling with it, too—for example, with regard to the question of how to involve and empower language communities in our work on Common Voice. But for meaningful engagement of and learning from external stakeholders to become more commonplace in the AI ecosystem, more guidance and an overview of key considerations are needed.

Additionally, the RMF should also point to the importance of considering aspects of diversity, equity, and inclusion in the teams designing and developing AI systems as part of the risk assessment process. While through a less immediate channel, risk—and serious harms—can also emanate from a failure to consider diverse perspectives in the design and development process. At Mozilla, this is an important concern to us. For this reason, among other things, it is why we have addressed the issue as part of our Teaching Responsible Computing Playbook. But this is not purely a "pipeline problem." Being mindful of these concerns and advancing diversity, equity, and inclusion within an organization should also be considered good practice for organizations developing (and deploying) AI systems. We therefore hope that these considerations will be reflected in the final version of the RMF.

Mozilla appreciates the opportunity to comment on the AI RMF and to provide our perspective as both a non-profit foundation and a technology company. We are looking forward to future iterations of the RMF and would be happy to respond to any questions you may have regarding our comments.