Hugging Face
29 September 2022

20 Jay St
Suite 620
New York, NY 11201

Hugging Face Comments on the National Institute of Standards and Technology's AI Risk Management Framework

Hugging Face congratulates the National Institute of Standards and Technology (NIST) on the second draft of its AI Risk Management Framework (RMF). We are eager to support the RMF's development and implementation in becoming a key resource in the field. We offer recommendations to strengthen this framework based on our experiences toward democratizing good AI and characterizing risks of systems as an open platform. Our comments narrow in on specific, actionable feedback for sections and their subsets, ordered below.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good machine learning. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI. Most recently, we supported the releases of DALL·E Mini by craiyon.com and Stable Diffusion by Stability.AI, while crafting in-house documentation processes and responsible AI licenses.

1. OVERVIEW

We broadly agree with the attributes listed for the AI RMF and recommend further specificity on point 10.

> *10. Be a living document: Process for updating*
> As stated, AI progress is fast-paced. As stated in Appendix B, while there are similar lessons between AI and software risks, the novel dual-use nature of many general-purpose AI systems differentiate risk approaches. Should the same RMF be applied across all AI systems? What systems should be further scrutinized? **This living document should describe the process of how it will be updated and the projected frequency.**

2. AUDIENCE

While a broad set of stakeholders is necessary, the breadth also necessitates further delineation of roles and responsibilities and specificity around tasks. Test Evaluation Verification Validation (TEVV) experts with deep understanding of both the societal and technical aspects of a system are rare. **The RMF should specify "technical, societal, legal, and ethical standards or norms", where to source these standards and norms, and consider how implementable these standards and norms are to technical systems.** Both standards and norms are likely to differ by culture, community, and type of system. They may be high level and difficult to implement technically. A

specific organization can build and tailor its own ethical norms or charter, such as the BigScience Ethical Charter, which is based on broader international standards but also requires deep in-house expertise.

Additionally, there is no guidance on what constitutes a TEVV expert and smaller organizations may not have the infrastructure or resources to afford experts. **Further guidance is needed on what skills or qualifications are required of a TEVV expert, whether they are required to be in-house, and how their legitimacy is determined.** If consultation outside an organization is encouraged, the RMF's resources should share how an organization can find, validate, and compensate experts.

As a way to minimize duplication of effort, the RMF should also provide a path to **making TEVV work community-oriented when relevant.** TEVV expertise applied to commonly used and open-source systems should be shared either openly or in a central repository that serves as a resource available to experts analyzing other applications of the same systems.

3. FRAMING RISK

We strongly agree with acknowledging the difficulty of measuring risk, both qualitatively and quantitatively. In the adversarial setting of a real-world and post-deployment environment, different stakeholders may be closer to the system than those who conducted initial risk assessments. For example, the developers of a general purpose system such as a language model may not be as attuned to the downstream use case of that system by a company adapting it to commercial needs. **NIST should provide communication guidance for the sets of evaluators and stakeholders throughout a system lifecycle, from development to deployment.**

3.2.1. Risk Measurement

Since relying on existing pieces and developing ML artifacts out of context with a view to re-usability has become common practice, addressing this challenge should be prioritized. Analysis *a priori* can apply across a broad range of systems and levels before analysis in context. For example, Hugging Face provides tools/documentation to support:

- Tools to **analyze data** and explore its biases
- Tools to explore and compare various ML models' behaviors on tasks like **speech recognition**
- Tools to explore the various steps and various contexts of an ML task lifecycle for tasks like like **Automatic Content Moderation**
- Tools to explore the **metrics** used to evaluate models and the biases they themselves may introduce
- The Hugging Face hub also hosts a **range of community-contributed tools for evaluating biases** of various ML artifacts.

3.2.4. Organizational Integration of Risk

The RMF and Playbook should give examples of how to properly integrate the given RMF into organizations' processes. An established enterprise risk management process may differ by sector and size of the organization. **We suggest providing case studies in the**

Profiles section of the RMF in action across types of systems and sectors. These examples should show how different size organizations implement the RMF, as well as how to engage stakeholders. This can also provide examples of documentation to increase transparency, such as with Model Cards.

4. AI RISKS AND TRUSTWORTHINESS

The listed characteristics are not only interrelated but also have heavy overlap. Many of these given terms are interpreted differently across organizations. For example, AI safety is a broad field that could intersect with fairness and bias, and explainability could merge with transparency. **Further guidance is needed on how to conduct the suggested contextual assessment to address potential tradeoffs between system performance and trustworthiness.** We commend NIST on the definitions for Valid and Reliable, as well as Security and Privacy. **All characteristics should reference more in-depth resources**, such as NIST's publication on bias. The importance and order of the characteristics needs further discussion.

Human Factors

This section makes valid arguments about human-in-the-loop. We strongly disagree with AI systems being used "in high-impact settings as a way to make decisions fairer and more impartial than human". This should not be encouraged, as this can amplify harmful biases and outcomes and contradicts many arguments for assessing risk pre-deployment.

4.2. Safe

This abstract term overlaps most with the other listed characteristics and is interpreted differently among many AI organizations today. The term as it stands can refer to institutional risks, such as concentrating power among high-resource developers, or technical risks, such as unreliable outputs. **Safety should be an overarching goal of the many characteristics.**

4.3. Fair – and Bias is Managed

The complexities of fairness and bias cannot be captured in a short playbook section. We applaud the description of systemic bias and the many layers needed to address in ultimately reaching fair outcomes. This section should clearly point to either the NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, or another publication explicitly as the resource for this topic. To operationalize fairness, lessons from legal definitions of non-discrimination can also be helpful.

4.5. Transparent and Accountable

Two key questions remain unaddressed in this section: first how should stakeholders communicate about risks in a standardized and clear manner? And second, whom should be held accountable for harmful outputs from an AI system? Throughout a system's lifecycle different documentation processes are needed, which could spark intellectual property and privacy concerns. **The RMF should have optional templates for documentation and suggest the level of transparency needed throughout the system lifecycle, depending on legal considerations.** Since BigScience is an open-science organization, we documented training data as a Data Catalog. We are also transparent about what tests and results we run on our model BLOOM. This may not be the case for all organizations.

4.6. Explainable and Interpretable

Full interpretability is technically difficult to achieve in most systems and may be less efficient than contesting a decision or enabling human recourse. This section should explain the ultimate goal of explainability.

6. AI RMF CORE

As defined in section 3.1, a “risk” considers but is distinct from an impact. The RMF by nature should not encompass harms, but the RMF Core cannot operate as a circle without risks developing into impacts. Understanding how external, adversarial, and real-world lessons affect the core process will strengthen actions taken.

6.1. Govern

Both technical and organizational processes require a diverse team, as noted in the previous section. Ongoing processes should have clearly defined timelines for when to reassess each risk and system.

6.2. Map

Exhaustively mapping potential risks and benefits is nearly impossible for dual-use systems; AI systems that can be used for both good and bad. Especially for general purpose systems such as language models and multimodal text-to-image models developed without a specific downstream application in mind, mapping is an ongoing challenge. **Public institutions can guide developers by specifying sectors and use cases that are most likely out-of-scope or highest-risk.**

6.3. Measure

We appreciate measurements not being exclusively quantitative. Robust measurements and evaluations also require multidisciplinary expertise. **Tutorials and resources that can bridge gaps between social science and computer science lead to better socio-technical assessments.** For example, Hugging Face’s Evaluate library provides tutorials and no-code interfaces to run technical evaluations on our hosted models.

Conclusion

Risk management is increasingly urgent in the AI field. Further specificity on key terms, processes, experts, and metrics will make this RMF more easily implementable. We commend NIST on this second draft and look forward to supporting the RMF as it reaches its final stages.

Respectfully,

Irene Solaiman
Policy Director, Hugging Face

Yacine Jernite
ML and Society Lead, Hugging Face