Microsoft

AI Risk Management Framework: Initial Draft
Comments of Microsoft Corporation

April 29, 2022

Microsoft appreciates the opportunity to respond to the initial draft of the NIST AI Risk Management Framework. We applaud the framework's goals and direction. The draft provides a strong foundation from which to build trustworthy AI systems and incorporates many essential elements of a strong risk management framework appropriate for the unique risks that AI systems can pose and society's needs more broadly. In particular, we welcome the risk-based approach that the AI RMF adopts, and the way in which it will help organizations identify and address the sociotechnical nature of many AI risks. The framework can be strengthened by clarifying the complex nature of sociotechnical risks and what they mean for organizations developing and deploying AI and providing more detailed guidance about how organizations across the AI value chain can implement the AI RMF. Microsoft provides recommendations on these topics and other priority issues below. We look forward to continuing to support NIST in this work which will be of significant value to organizations across the United States, as well as an important contribution to the global discussion on AI governance and regulation.

Help organizations understand the sociotechnical nature of AI risks and their related responsibilities

For the AI RMF to be successful, it must ensure that all organizations developing and deploying AI understand the sociotechnical nature of AI risks, how these risks manifest, and how organizations should respond. We believe there is an opportunity to provide greater clarity to organizations about the nature of these risks and how to address them. Specifically, we recommend the following:

- Clarify that sociotechnical AI risks emerge from the interplay of technical development decisions with decisions taken about how a system is used, who operates it, and the social context into which it is deployed. The AI RMF should emphasize that addressing these risks requires action from entities developing and deploying AI throughout a system's life cycle.
- Consolidate the three classes of "risks and characteristics" (i.e., Technical Design Characteristics, Socio-Technical Characteristics, and Guiding Principles Contributing to Trustworthiness) into one list of risks that all organizations should be responsive to. Dividing the risks into different classes may give the impression that an organization need only be mindful of "technical" or "sociotechnical" risks depending on their role as developer or deployer of an AI system. Characterizing "technical" risks like "accuracy" as those "under the control of AI system designers and developers", for example, may cause confusion and fail to capture the importance of deployers testing system performance - including measuring the accuracy of human-AI teams – and taking steps on training and oversight. There is also an opportunity to rationalize the characteristics into a more compact list, removing some of the overlap between "fairness" and "managing bias," for example, or "transparency," "explainability," and "interpretability." Having a consolidated list of core characteristics will likely make it easier for organizations to apply the Framework.
- Provide more detailed guidance on how organizations should understand their role across the lifecycle of the conception, development, and deployment of AI systems and what it means for how they apply the AI RMF Core. The Core provides a comprehensive set of considerations that pertain

to both development and deployment related decisions. For example, Map Category 1 talks about both identifying a system's intended purpose and the setting into which it will be deployed. Often, the organization developing a system will be different to the one deploying it. The AI RMF should help organizations understand their respective responsibilities and incentivize those developing systems are doing so in a way that ensures they can perform fairly, safely, and reliably when deployed appropriately, and are communicating information about how the system works, and its capabilities and limitations, to facilitate responsible deployment decisions. Microsoft's *Transparency Notes*, for example, seek to provide this type of information to customers who are deploying our general purpose APIs into systems of their choosing[1]. Similarly, the AI RMF should help deployers understand their responsibilities around using a system for the uses or purposes for which it is has been developed and and to task and train individuals to appropriately operate, oversee and act on the outputs of the system while in operation.

Use the AI RMF to advance impact assessments

Impact assessments are used by many organizations, including Microsoft, to identify and mitigate AI risk. Microsoft recommends that NIST use the AI RMF to encourage organizations to conduct impact assessments for AI systems as part of implementing the AI RMF Core. Many of the elements of the "Map," "Measure," and "Manage" functions are important parts of an impact assessment process and so this adjustment would both go with the grain of industry practice and accrue to the goals of the AI RMF. We suggest NIST highlight the way in which impact assessments help organizations identify and mitigate AI risk in a structured and standardized manner. There is existing language in NIST's publication *Towards a Standard for Identifying and Managing Bias* that highlights the value of impact assessments[2], which NIST should leverage in the context of the AI RMF. We also include recommendations below on identifying impacts on affected stakeholders and adding additional guidance on processes that can support the AI RMF Core's outcomes in a way that would advance the use of impact assessments as a mechanism to operationalize the AI RMF.

Provide more guidance around how to identify impacts to stakeholders across society

Microsoft welcomes the broad view the AI RMF takes of different societal stakeholders impacted by AI, including individuals that may be indirectly impacted by the behavior of an AI system. Identifying these stakeholders and assessing the impact of a system on them is an important part of risk identification and mitigation. This section of the AI RMF could be strengthened further by the following:

- Include a more explicit reference to "individuals impacted by system performance", including "decision subjects" evaluated by an AI system, in the "general public stakeholder" box in Fig. 1
- Provide more detail about the stakeholders that should be identified as part of Map Category 1, including those who will be using the system output to make a decision, those tasked with

[1] Microsoft provides information for our AI services via Transparency Notes with a view to helping customers make informed deployment decisions. They outline the capabilities and limitations of a system as well as key considerations for how to use it responsibly.
[2] Page 35-36 of Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (nist.gov)

overseeing the system to ensure appropriate operation, and those who may be indirectly impacted by a system and the decisions it informs.

- Provide details on the process an organization can use to document potential system benefits and harms for each stakeholder group identified. This type of process is an important part of an impact assessment and would help organizations incorporate the AI RMF Core into an impact assessment process. Organizations should also ensure that those who will use and oversee a system have been enabled to do so effectively, and that an individual impacted by an AI-informed decision that produces legal effects or similarly significantly effects is provided meaningful information about the logic involved and significance of the decision, and the ability to contest it.

Orient AI RMF Core around processes and outcomes

Microsoft supports the way in which the AI RMF focuses on outcomes rather than prescriptive requirements. Achieving this outcome-oriented approach will require guidance on the processes that organizations should develop to meet the goals in the AI RMF Core, either as part of the Practice Guide or by adjusting the subcategories in each function. For example, instead of Category 1 in the Map function reading "stakeholders are defined…and outreach is conducted," it could be refactored as "Identify stakeholders who will be responsible for troubleshooting, operating, and overseeing the system, document their responsibilities, and ensure they are enabled through system design and/or training to perform this role effectively". This type of framing will help organizations more effectively implement the framework and align with the way impact assessments are conducted, where organizations follow a set of processes to identify and mitigate risks.

Provide more detailed guidance on risk evaluation

Microsoft applauds the AI RMF's risk-based framing. We believe the Framework can be of most value by helping organizations identify and address the use cases that pose the highest risk of potential harms. As part of this, Microsoft recommends building on Fig. 2 to clarify that AI systems used in a way that may pose risks of these types of harm should be classified as high-risk and subject to robust safeguards. Microsoft commends the AI RMF for articulating a comprehensive framing of potential harms spanning individuals, organizations, and broader systems. We recommend adding threats to psychological safety in the "Harm to People" category, as well as threats to "human rights," alongside "civil liberties" and "rights". We also recommend adjusting the group harm definition to read a "group of people is *unfairly impacted* or discriminated against…". We include further information on specific fairness harms in our section on the AI RMF taxonomy, below. It will also be important for the AI RMF and its companion materials to help organizations set appropriate risk thresholds. Doing so effectively will be challenging for organizations given the broad and varied nature of AI risks and the way they are highly scenario specific. Providing examples of how to address legal, societal and commercial risks across different systems and scenarios will be beneficial in helping organizations calibrate thresholds. While risk tolerance will vary by organization, there is a baseline that all organizations will have to meet to ensure AI systems are being used in a manner that does not yield significant harm and is compliant with local law. It is important to note that these baseline thresholds will likely vary from system to system and will be scenario specific. The AI RMF should help organizations identify what these baselines are for each scenario.

Provide guidance on testing AI systems and support research in this area

Testing AI systems to ensure they are performing appropriately for a chosen deployment scenario is an essential part of measuring and mitigating risk. Systems should be tested during development and in operational conditions prior to deployment and throughout the lifecycle of a system, including with humans-in the loop, given the significant impact environmental conditions can have on performance. Guidance around how to test systems, set appropriate performance ranges and conduct system monitoring throughout its lifecycle will be beneficial.

As the AI RMF identifies, more work is needed to develop testing frameworks for AI systems that can be easily deployed by organizations of all sizes for benchmark and operational testing across different types of systems and applications. Microsoft strongly encourages NIST to build on its expertise in this area, including the leading work of the NIST FRVT program, and to prioritize further work to identify standardized testing methodologies, performance benchmarks and methods of reporting of testing results to allow for easy understanding and comparison. An important part of this will be developing metrics that address the performance of broader system of human-AI decision making and helping organizations employ metrics that are appropriately aligned with the intended outcome of the system. Incorporating the "Fairness metrics[3]" section of *Towards a Standard for Identifying and Managing Bias* into the AI RMF in some way may be helpful to highlight some of the techniques that are available for fairness assessments. It will also be important to ensure measurement techniques assess the performance of the wider decision-making system and effectiveness of human oversight, including identifying where cognitive biases, such "automation bias," or an overreliance on system output, may create risks.

Build out recommendations and guidance on appropriate use and human oversight

Tasking and training individuals on the appropriate use, oversight, and interventions on system operation is an important part of mitigating AI risk. Microsoft welcomes the addition of a training provision as part of the Govern function and would recommend adding more detail around the type of training that should be provided so that individuals tasked with roles in relation to appropriate system use and monitoring are adequately able to do their job. Organizations should ensure that those performing these roles in relation to high-risk systems meet a higher bar for training and are able to demonstrate proficiency. NIST's work to develop "formal guidance" around how to implement human-in-the-loop processes in a manner that does not amplify or perpetuate any systemic or computational biases, referenced in *Towards a Standard for Identifying and Managing Bias*[4], will be valuable. Microsoft is strongly supportive of this work and is keen to contribute, where helpful.

Provide actionable guidance, supporting examples and information on tooling

It is great to see that NIST intends to create "Companion Papers," "Profiles," and "Practice Guides" to illustrate the way in which organizations can use the AI RMF. Microsoft would recommend prioritizing the development of these materials alongside the draft AI RMF framework. In our experience with

[3] Page 30 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (nist.gov)
[4] Page 38 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (nist.gov)

Microsoft's responsible AI program[5], we have found that this kind of guidance is core to teams being able to implement requirements and that the development of these materials often highlights opportunities, challenges and inconsistencies in the way in which the broader program is structured. These examples should span a range of different AI technologies (e.g., natural language processing, classification, computer vision) and customer use cases, highlighting the different types of risks that can emerge at different points in the AI value chain. The guidance that NIST has developed in its *Towards a Standard for Identifying and Managing Bias* on datasets, testing, evaluation, validation, and verification (TEVV), human factors and governance should be included in the AI RMF or cross-referenced in a way that allows for ease of implementation as part of this. Providing information on responsible AI tooling will also be valuable. Microsoft has helped develop open-source tools like Fairlearn, Error Analysis, InterpretML, and the single-pane-of-glass tool called Responsible AI dashboard[6] which may be helpful to teams looking to understand model errors and predictions, assess model fairness, and inform data-driven decisions.

Enable integration with other frameworks and emphasize need for iteration and board oversight

NIST should ensure the AI RMF can be easily integrated with Cybersecurity and Privacy Frameworks. Guidance, including explanatory visuals, on how the frameworks can interact will be beneficial. It will also be helpful to provide more detail around how the AI RMF can be used in a way that aligns to relevant international standards, including on AI management systems, and AI concepts and terminology. With a view to advancing interoperability, Microsoft would encourage NIST to continue its engagement on the EU AI Act and regulatory conversations elsewhere to ensure that the provisions of the AI RMF, AI Act and other regulatory regimes can be easily aligned. Microsoft would also suggest renaming the "Map" function "Identify" to better reflect the nature of the function. Doing so would also more closely align with the structure of the Cybersecurity and Privacy Frameworks which include an "Identify" function. In the Govern function, Microsoft would encourage NIST to emphasize the need for continuous improvement of related policies and processes so that governance can keep pace with developing technology and shifting societal expectations. Microsoft welcomes the way the AI RMF highlights the need for executive level ownership of AI governance and would encourage NIST to also emphasize the value of having board level oversight in this area, particularly in larger organizations.

Feedback on definitions in AI taxonomy

Microsoft applauds the work of NIST to build out a taxonomy of AI risk which will be beneficial to organizations looking to mitigate risk and influential in the global discussion on AI governance and regulation. In addition to the suggestions above on how to streamline the taxonomy, Microsoft provides suggestions below on some of the definitions included in this section:

- **Accuracy:** Accuracies of performance depend on the nature of the test set used to evaluate the performance of a system. Accuracies should always be paired with clearly defined test sets, and details about the case library used in the tests should be included in associated documentation. In real-world deployments, case libraries should be drawn from the local domain of application

[5] The building blocks of Microsoft's responsible AI program - Microsoft On the Issues
[6] Responsible AI Toolbox Capabilities - Microsoft Responsible AI

and be representative per context and freshness. It will also be helpful to extend the definition of accuracy beyond measuring the ML model in isolation to include accuracy of the human-AI team as, in deployment, a human is often in-the-loop and may override or edit system output. For example, in many AI-centric decision support scenarios, automated inferences are communicated to human decision makers who are responsible for making decisions. More research is needed, but studies have shown that the design of displays and workflows can affect the influences of AI advice on people.[7] Thus, testing the ML model in isolation, such as by measuring the extent to which it correctly captures a relationship that exists within training data, may not reflect accuracy in operational conditions. There is value in measures of accuracy for models per test sets and we recommend NIST promote these system-centric measures. However, the broader definition of accuracy of human-AI systems will be critically important to help all actors in the value chain understand they have responsibilities, including deployers whose system choices can affect performance of the system in operational conditions. As an example, deployers have choices to make in selecting an appropriate use case and ensuring that their data, people, and processes support the accurate performance of a system.

- **Fairness and Managing Bias:** We recommend consolidating these characteristics into a "Fairness" section that retains the strong work that NIST has done in identifying the different sources of "bias" and helps organizations understand the importance of achieving key fairness goals in their development and deployment of AI systems. These goals should include ensuring systems: 1) provide a similar quality of service for demographic groups impacted by the system 2) allocate resources or opportunities in a manner that minimizes disparities in outcomes between demographic groups impacted by the system and 3) minimize the demeaning, stereotyping, or erasure of relevant demographic groups. We also encourage NIST to help organizations understand the way in which definitions of fairness may shift depending on application, something that organizations will need to be responsive to in performing appropriate risk identification and testing.
- **Reliability:** Reliability is a goal for overall correctness of model operation under the conditions of expected use and over the lifetime of the system, as is expected in any software system. We note the importance of systems performing accurately, not just consistently, if they are to be considered reliable which is not yet reflected in NIST's definition. Measures of reliability, both on test data and while a system is in use, should include a decomposition of model quality as a function of different relevant dimensions of the input content (such as environmental conditions, etc.). They should also take into account that certain types of failures can cause greater harm and calibrate thresholds appropriately to minimize such failures.
- **Robustness**: Robustness is a goal for appropriate system behavior in a broad set of conditions and circumstances, including outside of expected or anticipated use. Appropriately robust behavior does not require that the system perform exactly as it does under expected uses but that it behaves in ways that minimize any potential harms to stakeholders if, for whatever reason, it is operating in an unexpected environment. Building robust systems requires anticipating and planning for the frequency of failure in normal operation and the ability for the

[7] R. Fogliato, et al. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging (cmu.edu). ACM Conference on Fairness, Accountability, and Transparency (FAccT), June 2022.

system to determine when it is operating outside the regime of anticipated uses. In both cases, graceful degradation of service is a property of a robust system.

- **Resilience or ML Security: "**Resilience" and "Security" are related but distinct issues. Resiliency, or the ability to return to normal function after an attack is a different quality than security, which includes resiliency, but also encompasses protocols to avoid attacks. Both are important and need to be appropriately defined and differentiated as they are in the EU AI Act and IEEE frameworks. More information around how risks related to the different security concerns listed may manifest would be helpful, and the issue of data poisoning should be added to the list.
- **Explainability, Interpretability and Transparency:** It may be useful to refocus these sections on how they accrue to the overall goal of ensuring those operating and overseeing a system are able to effectively do so, including in relation to intervening in a system's operation or modify its output. An important element of this will be ensuring system developers provide easy to understand information about intended uses, system capabilities and limitations and factors that affect performance. We also recommend rationalizing the terms "explainability" and "interpretability" into one characteristic, given the way they are used interchangeably in responsible AI discussion and with a view to simplifying the Framework for ease of use. It will also be important that those using AI are transparent about the type of technology being used and how. This is particularly important for organizations, including those in the public sector, that are using high-risk systems. It is also important to note that increasing transparency in itself does not necessarily lead to greater interpretability and can raise challenges around protecting privacy and security, including in relation to data poisoning and exfiltration of models.
- **Privacy:** Microsoft believes privacy is a fundamental right. Moreover, there is a growing volume of privacy law internationally, notably the EU General Data Protection Regulation and various U.S. state laws, that create legal obligations for organizations. Highlighting the need to prioritize privacy issues may be helpful as part of this section rather than suggesting that privacy is a subjective issue that varies "among cultures and individuals" which may be interpreted as suggesting that mitigating such risks is not a top tier priority.
- **Safety:** The safe operation of AI systems should be viewed as a priority to help ensure AI systems do not, under defined conditions, cause physical or psychological harm. Microsoft welcomes that this section highlights the importance of testing and monitoring; we also emphasize the need for responsible development practices, clear information to deployers on how to use a system appropriately, and the need for deployers to take responsible deployment decisions in line with this guidance. We recommend that measures for safety be harmonized with measures of safety used in other fields, including transportation and aerospace.