ITI Response to NIST Artificial Intelligence Risk Management Framework (AI RMF) Second Draft

The Information Technology Industry Council (ITI) appreciates the opportunity to continue its engagement with the National Institute of Standards and Technology as it seeks to develop an *Artificial Intelligence Risk Management Framework* (AI RMF). As such, we are pleased to provide comments on the *AI Risk Management Framework: Second Draft*, as well as offer feedback on the *AI RMF Playbook.*

ITI represents the world's leading information and communications technology (ICT) companies. We promote innovation worldwide, serving as the ICT industry's premier advocate and thought leader in the United States and around the globe. ITI's membership comprises leading innovative companies from all corners of the technology sector, including hardware, software, digital services, semiconductor, network equipment, and other internet and technology-enabled companies that rely on ICT to evolve their businesses. Artificial Intelligence (AI) is a priority technology area for many of our members, who develop and use AI systems to improve technology, facilitate business, and solve problems big and small.

As we have noted in prior submissions, we are engaged in AI policy conversations around the world and are committed to bringing a global perspective to conversations around fostering trustworthy AI. We have also actively worked to inform NIST's efforts to foster trust in AI technology, including responding to NIST's RFI on an AI Risk Management Framework,[1] the RFI on the AI RMF Concept Paper,[2] and provided comments on the Initial Draft of the AI RMF.[3]

As a general matter, we are pleased with the direction of the AI RMF and are happy to see that some of our prior feedback has been incorporated into the second draft. We believe that this tool will be useful for stakeholders around the world as they seek to identify and treat risks related to AI systems and will help to foster trustworthy AI. With that in mind, we have several additional suggestions that we believe will help to further strengthen the AI RMF.

Recommendations to Strengthen the AI RMF

1) Add a function that accounts for contingencies.

We have suggested this throughout our comment submissions – from the Concept Paper to the Initial Draft -- but we continue to stress the importance of adding a discrete 'Respond' function for managing incident response and contingent risks. While we recognize this would significantly alter the way in which the Framework functions are currently construed, we believe there is immense value in adding a function that is aimed specifically at addressing contingencies such as model

[1] See ITI response to RFI on AI RMF here: https://www.itic.org/documents/artificial-intelligence/NISTRFIonAIRMFITICommentsFINAL.pdf

[2] See ITI response to RFI on AI RMF Concept Paper here: ITI Comments on AI RMF Concept Paper FINAL.pdf

[3] See ITI response to AI RMF Initial Draft here: https://www.itic.org/documents/artificial-intelligence/ITICommentsonAIRMFInitialDraftFINAL.pdf

degradation, breaches, and unexpected adverse outcomes of AI systems. While NIST includes incident response in the "Manage" function under Categories 1 and 4, we continue to believe that a separate function that maps practices that organizations might undertake to respond to an AI-related incident would be useful. While we understand the intent of the Manage function is to capture activities such as incident response and contingencies, in the AI context we feel it is appropriate to include a discrete Respond function. As currently drafted, it is a bit like folding the 'Recover' function of the Cybersecurity Framework into the 'Respond' function. It is true that recovery is technically a part of response, but because the stakeholders and processes involved are a bit different, it is helpful that the CSF splits it out into two separate functions. In the same way, differentiating 'Respond' from 'Manage' would allow NIST to more specifically focus on those activities that go into responding to an incident or event, including, for example, retraining models, adjusting metrics, etc. While the 'Manage' function can capture this, we continue to believe incorporating 'Response' into this function downplays the significance of having a discrete plan in place should an incident or other adverse event occur.

Furthermore, it might also be useful to create a database with best practices gathered from the results of a Respond function so that organizations can leverage such data to proactively monitor for such incidents and deploy mechanisms (some of which may be automated, i.e., MLOps) to consistently check for risk factors. This may also help to encourage stakeholder alignment. It is also worth noting that the OECD is planning to also develop a common framework for reporting on AI incidents, and a Respond function would help feed into and help align with that process.[4] The current incident database curated by the Partnership on AI may also yield useful insights.[5]

2) Map and/or further seek to align the AI RMF with international standards.

As we suggested in our prior set of comments responding to the Initial Draft, we encourage NIST to further align with international standards to encourage consistency in the way organizations are implementing risk management processes. We particularly encourage NIST to utilize ISO/IEC FDIS 23894 AI Risk Management. While we understand that NIST does not want to fold standards that are still under development into the AI RMF given their inchoate nature, ISO/IEC standards at the DIS and FDIS stage have reached a level of maturity and stability that it is in fact appropriate to focus on alignment with standards at this stage of development. ISO/IEC FDIS 23894 is in the final draft stage and is expected to be published in January 2023, which aligns with the target publication date of Version 1.0 of the AI RMF.

We also encourage NIST to seek to further align with ISO/IEC 5338 – Information technology – Artificial intelligence – AI system life cycle processes, ISO/IEC 38507 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations, ISO/IEC 24028 – Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence, and ISO/IEC FDIS 23894 Table C.1 Risk Management and AI System Lifecycle. As we mentioned in our prior responses to the Concept Paper and the Initial Draft, it would be helpful for NIST to further illustrate the stages following deployment, including the post-market stages, which may engender certain risks across a longer period of time, and the retirement phase, which marks the end of the lifecycle and may pose a different set of risks, such as terminating the

[4] See more information on the OECD Risk Classification Framework here: https://oecd.ai/en/wonk/classification
[5] See more information here: https://partnershiponai.org/workstream/ai-incidents-database/

system and deleting data within the model in line with requirements of applicable data protection laws and frameworks. Indeed, risk management does not cease with the deployment of an AI system and continuous monitoring should underpin post-market governance. NIST should take interdependencies between risks and residual risks into consideration.

We also recommended this in our last set of comments, but ideally NIST should also seek to align the terminology used in the AI RMF with the terminology specified in ISO 31000:2018, IEC/ISO 31010:2009, ISO/IEC DIS 23894 (Clauses 6 to 6.7) and ISO/IEC 22989: 2022. Alternatively, NIST could map the RMF terminology with these international standards to support stakeholders that are adopting these international standards. By doing so, NIST would serve as an important role model for other regional efforts, demonstrating the importance of alignment with international standards. Additionally, a misalignment in terminology, nomenclature, processes, and methods with those used in international standards will make it difficult for both industry and government to understand and apply the AI RMF. By mapping and seeking to reconcile terminology, guidelines, and requirements across multiple jurisdictions, NIST can help to prevent duplication of efforts, prevent different interpretations of key terms and requirements, and help to facilitate seamless integration into existing organizational risk governance.

For example, if it is not prudent to entirely integrate the ISO/IEC FDIS 23894 terminology into the AI RMF, NIST could map the "Map -> Measure -> Manage" functions to this standard as indicated below:

- "Map" is covered by ISO/IEC FDIS 23894 under *6.2 "Communication and consultation"* + *6.3 "Scope, context and criteria."*
- "Measure" is referred to in ISO/IEC FDIS 23894 as the iterative *6.4 "Risk Assessment = Risk identification, risk analysis, risk evaluation"* cycle.
- "Manage" corresponds to Risk Treatment in ISO/IEC FDIS 23894 and ISO 31000. ISO/IEC 23894 and ISO 31000 include a response function as part of "implementing risk treatment plans", inclusive of verification of effectiveness.

NIST could also map the AI RMF to the terminology in ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology around the AI lifecycle. In particular, we point NIST toward ISO/IEC 22989 *Figure 3 — Example of AI system life cycle model stages* and *Figure 4 — Example AI system life cycle model with AI system-specific processes*:

- NIST specifies the "pre-design" stage as "Inception" by ISO/IEC 22989, and the "Data collection" activity in the AI RMF is part of the ISO/IEC Design and development stage.
- Additionally, NIST uses the term "Deployment" to describe the entire stage after release of the AI system. On the other hand, ISO/IEC 22989 breaks the post-deployment lifecycle down into several stages: "Deployment" which is the initial release to operation; "Operation & monitoring" (which is the longest, sustaining stage); "Re-evaluation"; "retirement". Each of these stages incur different risks, challenges, and opportunities.
- Finally, ISO/IEC 22989 uses the term "retirement", where "decommissioning" is only one of several retirement options of the system.

To close, even if various international standards are still under development, we think it prudent to highlight that there is work occurring in standards development bodies that may inform AI risk management practices. If NIST is not able to reference these standards in the base text of the AI

RMF, NIST should seek to undertake a future mapping exercise so that it is clear how the AI RMF aligns with ISO/IEC standards. NIST should also include these standards in the AI RMF Playbook, which may be able to be updated more easily given it is hosted online.

3) **Clarify how risks differ for human impacting and non-human impacting AI systems, as well as appropriate risk evaluation criteria**.

We suggested clarifying how risks differ for AI systems that impact humans and those that do not in response to the initial Concept Paper, as well as in response to the Initial Draft. While we appreciate that NIST added text on "human factors" to the Initial Draft, the language focuses more on human-in-the-loop measures to act as a backstop in certain instances, as opposed to the important difference between human and non-human facing risks. We urge NIST to address this in the next draft of the AI RMF, as it is instrumental to a nuanced risk management approach.

While some AI applications impact humans (e.g., face recognition systems, recommender systems, or hiring systems) many AI applications do not (e.g., analysis of weather information, defects on the factory floor, bottlenecks in networks, or state of the roads). AI systems that do not impact humans typically do not contain PII (personally identifiable information) in the data sets and frequently feed analytics to other machines, not human end users. As a result, there are distinct types of risks associated with systems that impact humans and systems that do not. For example, considering privacy risks is essential for systems that impact humans. But privacy risks are not present in weather sensor data analysis fed to another system that uses the analytics to assess climate patterns over a longer period. There are of course other risks beyond privacy worth contemplating related to human impacting systems; this is simply an illustrative example. Applying the same risk management requirements to both types of AI systems would not allow the technologists and evaluators to assess the risks for the AI systems in an actionable fashion and would also be onerous to organizations – disproportionately hindering innovation with no corresponding benefit.

Relatedly, something we have advocated throughout NIST's development of the AI RMF is establishing risk evaluation criteria to help guide organizations as they seek to establish risk thresholds and understand their risk tolerance/appetite. We note that this is still missing from the Second Draft of the AI RMF. While we recognize that establishing risk evaluation criteria is a significant undertaking, we continue to believe that such a methodology would be helpful for organizations in determining the risk-level of a specific AI use case, informing the steps that they should take to mitigate or treat the risk. Such a methodology should also identify the appropriate roles for AI developers, deployers, users, and other stakeholders in making risk determinations. These determinations are also crucial for helping stakeholders identify specific technological mechanisms for measuring, mitigating, and controlling high-risk attributes of AI systems, where applicable. We are not saying that NIST should bucket specific uses of AI into a "high-risk" category, or classify entire sectors as "high-risk," but instead that it should develop criteria that can help the relevant stakeholders with the relevant responsibilities and authorities to figure out what level of a risk a particular use case may pose. Including illustrative examples may be helpful, with the clear caveat that the examples are just that, illustrative, and not meant as a categorical determination. If NIST deems it unfeasible to include evaluation criteria in the AI RMF itself, then we strongly encourage NIST to launch a process with the goal of working with stakeholders to develop such criteria.

In this vein, we note that NIST includes 'evaluators' as a stakeholder in the AI ecosystem. They are included in the scope of two categories, both AI Design and AI Deployment in Appendix A. However, this creates confusion as to which AI actor is ultimately responsible for the evaluation of risks associated with an AI application. We encourage NIST to consider including additional guidance around the obligations and responsibilities of evaluators and how evaluations should be conducted. If it is not appropriate to include such guidance in the initial draft of the AI RMF, perhaps it should be included as an area for further exploration in an AI RMF Roadmap (see suggestion 12).

4) **Consider how the distinction that NIST makes between designers, developers, and deployers will work in the case of a general-purpose AI system.**

The AI RMF defines designers, developers, and deployers separately. Additionally, the way designers and deployers are defined by NIST suggests that designers/developers have much more control over the impact of the AI system than would be the case with a General-Purpose AI system, for example. With GPAI, much of how the AI system is used falls to the deployer (which could also be the developer in a dual role). It may be helpful to simplify the definitions included in the AI RMF to developer and deployer to alleviate confusion, or otherwise make clear that designers/developers do not have the same amount of control as a deployer. In ITI's AI transparency principles,[6] we use the terms 'developer' and 'deployer' to make this apparent, defining a **developer** (sometimes used interchangeably with **producer**) as the entity that is producing the AI system. In some cases, the AI system can be built into other products that are then deployed by a different entity and the **deployer** as the entity (sometimes used interchangeably with **provider**) that is deciding the means by and purpose for which the AI system is ultimately being used and puts the AI system into operation.

5) **Add further discussion around the fact that risks might only be able to be described in a qualitative or semi-quantitative manner and provide guidance around what to do if a risk cannot be measured.**

We appreciate that NIST has included Section 3.2 Challenges for AI Risk Management, and that it includes a discrete section around challenges in measuring AI risk, including that some AI risks may not be well-defined or well-understood. We encourage NIST to add a sentence to this section that reflects that it also may be difficult to measure AI risk quantitatively or qualitatively because of the current lack of industry consensus on robust and verifiable measurement methods that can be applied to different AI use cases.

Additionally, in our prior comments we emphasized the fact that given AI is an emerging technology, we are still learning about the range of potential risks, their likelihood, their severity and detectability, as well as how to measure them. With this in mind, we continue to believe that it would be helpful for NIST to indicate how the RMF might address a situation where such risks cannot be appropriately measured. It would be helpful to offer guidance on reasonable steps for treating that risk, without limiting innovation and investments in new, and potentially beneficial, AI technologies. Even more importantly, we encourage NIST to include text that clarifies that the

[6] See ITI's Policy Principles for Enabling Transparency of AI Systems here: https://www.itic.org/documents/artificial-intelligence/ITIsPolicyPrinciplesforEnablingTransparencyofAISystems2022.pdf

inability to measure AI risk does not imply that an AI system poses high or infinite risk. To put it another way, the absence of data should not be treated as justification for halting all use or development of a technology or use. In the same vein, not every measure of risk is meaningful. NIST should consider these inherent limitations in measuring risk which could lead to certain harms being overlooked.[7]

6) Include consideration of unintended uses in the Map function.

In the Map phase, NIST addresses the need for organizations to understand the intended purpose of the system, the setting in which the system is to be deployed, and the specific tasks supported; however, more time could be spent addressing the need to understand the foreseeable *unintended* uses of the system. How could the system be used inappropriately and/or *outside* of the bounds of its currently scoped intended purpose? If the system is in place, what else could be done with it outside of the current scope? In later phases of the AI RMF, more time could be spent addressing how likely such scenarios would be, and ways to mitigate these unintended uses of the system. This may be worth laying out in an associated AI RMF Roadmap.

7) Add a category to Govern around deciding what systems are covered under the AI RMF.

NIST should also consider the implications of including all AI systems within the AI RMF framework. Due to the ubiquitous use of AI systems across organizations, it would likely be burdensome to include all AI systems within the AI RMF. Ideally, organizations should have the ability to decide which of their systems is covered by the AI RMF. We recommend that NIST include this as a category or sub-category under the Governance Function.

8) Add more clarity around how to use the Cybersecurity Framework, Privacy Framework, and AI Risk Management Framework in tandem.

Although we recognize that NIST has added an explicit reference to the Privacy Framework to the AI RMF under Section 4.7 Privacy-Enhanced, it would be useful for NIST to add additional discussion around the linkage between the Privacy and Cybersecurity Frameworks and the AI RMF. Both privacy and cybersecurity characteristics are discussed in the trustworthiness characteristics that NIST lays out in the AI RMF, and while it is now clearer how an organization might leverage the Privacy Framework to facilitate the Privacy-Enhanced characteristic, it remains unclear how an organization might leverage the AI RMF in conjunction with the Cyber Framework. It would also be helpful for NIST to identify whether there are aspects of the AI RMF that map to either (or both) the Privacy and Cyber Frameworks. Section 1.2.1 of the Privacy Framework, for example, discusses the relationship between cybersecurity and risk management, and offers a helpful Venn diagram that very clearly illustrates where cyber and privacy risks overlap.[8] We continue to strongly encourage NIST to add a similar section on cyber and privacy risk management and AI risk management so as to help organizations understand how these risks appear in the context of AI and how they might use other Frameworks to address these risks together with the AI RMF.

[7] See Fazelpour and Lipton's "Algorithmic Fairness from a Non-Ideal Perspective" (https://arxiv.org/abs/2001.09773).

[8] See p. 3 of NIST Privacy Framework, available here: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

9) Consider ways to reach those in the audience who have "responsibilities to commission or fund an AI system and those who are a part of the enterprise management structure governing the AI system lifecycle."

In its current form, the RMF is a useful document for technologists and IT professionals and the other actors (page 6) highlighted by OECD's Framework for the Classification of AI systems. However, the RMF also notes that the primary audience includes those with "responsibilities to commission or fund an AI system and those who are part of the enterprise management structure governing the AI system lifecycle." In practice, this means C-suite leaders, so more needs to be done to reach that audience. We strongly recommend that as part of its publication and eventual promotion of the NIST AI RMF and playbook that NIST identify resources and mechanisms to reach the CEO audience. ITI would be happy to engage in conversations to help NIST achieve this goal.

10) Include a discussion of the importance of the "human-baseline approach" which sets the bar against human legacy systems, not against vague AI-related risk without important context.

In many cases, AI is augmenting or replacing what was once a primarily human activity. In those cases, effective risk management requires that risks be compared to those alternatives. As a starting point, human, manual processes are replete with risk, bias, and other ethical issues. When we look at typical harms or risks associated with existing or anticipated AI applications, they are often rooted in legal cases having to do with discrimination or harmful outcomes from manual processes. Further, leaving out the human-baseline comparison will ultimately limit AI adoption because one of the leading factors limiting adoption is entities' lack of confidence in their ability to manage risks when changing from human to automated processes.

11) In working with stakeholders to develop profiles for the AI RMF, NIST should encourage the inclusion of a variety of use cases.

While we understand development of the Profiles is a collaborative effort, we encourage NIST to ensure that there are examples for both entities that build and deploy their own models, as well as for organizations that use other teams'/vendors' models. In many instances, we have observed a focus on vendor/customer use cases, but this is not representative of the entire ecosystem.

It would be helpful for NIST to support industry development of Profiles for a variety of use-cases including:

- Human resources and hiring
- Health benefits
- Public health services
- Synthetic drug development
- Lending & credit
- Content moderation

12) Consider creating an AI RMF Roadmap.

Finally, we encourage NIST to create an AI RMF Roadmap, similar to the roadmaps created for both the Privacy and Cybersecurity Frameworks. These Roadmaps serve as a useful tool for continued collaboration as they outline areas that require additional exploration. This will be especially important in the case of the AI RMF as AI is an evolving technology. Even now, there is not a settled suite of standards and best practices that can be leveraged to achieve every Function under the Framework and it is clear there are areas where additional work needs to be done.

As we note in Section 6, for example, it may make sense to include assessing unintended uses of AI systems, for example, as an area requiring additional exploration. Indeed, it would be useful if in the future the AI RMF could provide guidance or lay down principles for AI operators to assess unintended impact.

Another area worth including in a Roadmap could focus on explainability and interpretability. Actionable guidance/principles related to explainability and interpretability, and how that guidance can be applied in using the AI RMF, would be useful to ensure that information that is disclosed can help in understanding the purpose and impact of an AI application.

Recommendations to Improve the AI RMF Playbook

We appreciate that NIST has launched the AI RMF Playbook as a complement to the AI RMF itself. Indeed, this tool is instrumental to ensuring that the Framework is actionable and implementable, particularly for organizations that may be less familiar with the scope of guidelines and best practices that are available to them. As a general suggestion, we think it would be helpful to have the Playbook appear in both text and interactive formats to maximize its use.

There are certain areas that could be clarified, however. These include:

- Governance 1.1 - As part of Govern 1.1, the Playbook states that "organizations can document the following" … "when auditing an AI system, has existing legislation or regulatory guidance been reviewed and documented?" It would be helpful for NIST to clarify whether auditing is intended to be based on this.
- Governance 1.2 - As part of Govern 1.2, the Playbook states that "organizations can document the following" … "to what extent do these policies foster public trust and confidence in the use of the AI system?" It is not entirely clear what this means as applied to Category 1.2 ("the characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures."). Again, additional clarity around what this means would be useful. At present, we are concerned that information that is documented around this question may not be appropriate to make available given the fact that nefarious actors could access it.

Additional comments

We encourage NIST to note that when it comes to AI transparency tools (e.g., system cards, model cards, etc.), the community is still determining the best approach for documentation. Certain tools may be more appropriate in some cases than others and what is most useful may depend on the audience. Developers of AI systems should be encouraged to test out different types of transparency tools. Additionally, NIST should include discussion around the limits of transparency, including the extent to which it is reasonable and feasible. Indeed, if documentation or other measures are undertaken to achieve transparency and accountability without adhering to a reasonable standard, it could be costly, time-consuming, negatively impact innovation, and serve to undermine competitiveness, without yielding impactful information. We are agnostic as to whether this should be noted in the Playbook or the AI RMF itself, but either way it should be highlighted.

We also note that the Playbook is currently geared towards technologists. Once it completes this initial version, NIST should consider replicating the playbook with different versions focused on different audiences—particularly one for legal experts and C-suite leaders which will be critical to achieving the objectives of the govern section.