Consumer Technology Association™

1919 S. Eads St.
Arlington, VA 22202
703-907-7600
CTA.tech

April 29, 2022

Consumer Technology Association
Comments on
NIST AI Risk Management Framework: Initial Draft

The Consumer Technology Association® ("CTA")®[1] respectfully submits these comments in response to the National Institute of Standards and Technology ("NIST") request for comments related to the initial draft of its Artificial Intelligence Risk Management Framework ("Framework" or "RMF").[2] CTA supports NIST's effort to create a flexible and voluntary risk management framework that will help identify and address risks in the design, development, use, and evaluation of AI products and services across a wide spectrum of types, applications, and maturity of AI systems throughout the AI lifecycle and "offer guidance for the development and use of trustworthy and responsible AI."[3] This initiative mirrors the agency's effort to create a comprehensive cybersecurity framework, which has proven to be a valuable resource for setting standards and guidelines in that area.

General Comments on the Framework

1. *Reaffirm Value of Risk-based Analysis of Opportunities and Threats Presented by AI*

Because risk is a function of adverse impacts that are likely to arise if particular circumstances or events occur, managing the risk of identified adverse impacts will help meet the goal of deploying and relying on "trustworthy and responsible AI." The outcome of the risk management framework should be to minimize anticipated negative impacts and identify opportunities to maximize positive impacts in the use of AI and the "algorithmic processes that

[1] CTA® is the tech sector. Our members are the world's leading innovators—from startups to global brands—helping support millions of jobs. CTA owns and produces CES®—the largest, most influential tech event on the planet.

[2] Artificial Intelligence Risk Management Framework: Initial Draft; released March 17, 2022. Available here: https://www.nist.gov/document/ai-risk-management-framework-initial-draft

[3] As an initial note, many of CTA's comments on NIST's initial Concept Paper have been incorporated into the RMF. For example, the Framework incorporates "safety" as one of the AI risks to be evaluated and also emphasizes the importance of obtaining input from diverse stakeholders, both suggested in CTA's comments. In addition, issues that we discussed are also included in the RMF. For example, we discussed the importance of including feedback mechanisms, and section 6.3 (Manage), includes reference to the deployment of mechanisms to receive user feedback.

learn from data in an automated or semiautomated manner."[4] Achieving that goal will require AI systems to be assessed using clear, plain, and commonly understood language, to mesh easily with other aspects of risk management, and to be useful across a wide range of perspectives, sectors, and technologies as they affect "cybersecurity, privacy, safety, and infrastructure."[5]

A foundation of managing AI risk is the evaluation of AI systems using nuanced risk assessments to ensure a balanced analysis of the risks of negative impacts versus the rewards of positive impacts. In general, CTA supports NIST's approach of allowing "organizations to specifically define their risk thresholds and [allowing] them to manage those risks within their tolerances," and encourages NIST to continue to engage with industry-specific groups for feedback. In line with that approach, risk assessments should recognize the varying degrees of risk presented by different AI systems and use cases with more attention to the risk of adverse impacts on serious or life impacting outcomes. Industry participation in this process, along with the participation of other stakeholders is essential to reach a consensus-driven transparent process that can evolve and be regularly updated to reflect AI deployment experience and the types of risks generated when using AI systems.

2. *Distinguish Risk Management Functions as Between Organizations Developing AI and Those Organizations Using AI*

CTA suggests that NIST consider including more focused guidance for organizations acquiring AI systems from third parties, as distinguished from those companies that develop the technology in-house. Section 6.4 (Governance) of the Framework states that it applies to companies that are *acquiring* AI systems and that "governance should address supply chains, including third-party software or hardware systems and data as well internally developed AI systems."

However, the Framework does not include guidance tailored to organizations that are acquiring AI systems. Such companies are likely to face unique issues when deploying acquired AI systems and may need specific guidance for managing AI risk. For example, such companies may not have access to information about the training data sets used by the company that developed the AI. As such, it will be difficult for the acquiring company to assess characteristics such as accuracy, robustness, or possible bias in the training data sets that may produce biased outcomes. Companies acquiring AI systems will also need to institute appropriate cybersecurity measures to ensure that third-party supplied algorithms do not contain vulnerabilities that would allow for malicious changes in input classification and corresponding output and also ensure data models are secure from tampering and unsupervised changes.

In addition to providing more guidance on the differences in responsibilities between actors, the Practice Guide that NIST is developing to accompany the AI RMF should provide examples that demonstrate how different responsibilities apply AI providers, deployers, and users when implementing the Framework.

[4] Artificial Intelligence Risk Management Framework: Initial Draft, at 2.
[5] *Id.*

3. *Recognize Certain Risks, Such as Bias and Fairness, Are Contextual and May Vary Depending Upon Circumstance*

CTA urges NIST to retain sufficient flexibility in the Framework to permit contextualized decisions about acceptable levels of risk. Assessments of certain risks require consideration of the circumstance, time, and situations in which such risks may be presented, and should explicitly include a weighing of risks versus benefits of a model.

Because there is no universally accepted concept of fairness, and because bias cannot be eliminated in all circumstances, the Framework should enable such contextualized decisions to ensure that steps taken to measure, map, and govern risks are reflective of unique circumstances presented in specific situations. Indeed, because decisions concerning governing bias may require tradeoffs between affected interests and intended goals of the system, developers and users of trustworthy AI systems will need to take a broad contextual approach to risk assessment and management.

Further, it would be helpful to refine the distinction between fairness and the absence of harmful bias. Section 5.3.1 of the RMF notes that "(f)airness is increasingly related to the existence of a harmful system, i.e., even if demographic parity and other fairness measures are satisfied, sometimes the harm of a system is in its existence." The section then goes on to state that "[w]hile there are many technical definitions for fairness, determinations of fairness are not generally just a technical exercise." The current description of fairness is broad and implies an expectation to do more than mitigate harmful bias. However, the RMF does not detail what additional criteria developers should consider to achieve fair AI systems. As such, CTA suggests that NIST refine and describe in more detail the relationship between fairness and the absence of harmful bias.

Relatedly, relying on the term accuracy may be misleading with regard to assessing AI systems as it has a specific technical meaning. For purposes of risk mitigation, it may be more effective to replace the term accuracy with another term, such as correctness or usefulness, to avoid confusion.

4. *Recognize That Bias Mitigation May Create Tension with Other Elements of Trustworthy AI*

As recognized in NIST's Special Publication, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,"[6] managing bias in AI has the potential to come into conflict with other elements of trustworthy AI such as accuracy and explainability. For example, NIST's Special Publication notes that achieving a high degree of statistical accuracy within an AI system may still "produce outcomes that are harmful to a social class and diametrically opposed to the intended purpose of the AI system." In addition, although explainability is a key component of trustworthy AI, the Special Publication highlighted the fact that simpler AI models, which tend to be more transparent and explainable, can actually exacerbate biases

[6] NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence; published March 16, 2022. Available here: https://doi.org/10.6028/NIST.SP.1270

"because restrictive assumptions on the training data often do not hold with nuanced demographics."

In light of these tensions, CTA recommends that NIST explicitly acknowledge that some objectives of trustworthy AI may be partially mutually exclusive, and therefore that it may not be possible to optimize all AI systems against all elements of the Framework. CTA suggests that in light of this acknowledgement, the Framework provide developers the flexibility to balance tradeoffs between the various elements of trustworthy AI to maximize the characteristics appropriate to the use of the AI system and risk of adverse or negative consequences.

5. *Purveyors of AI Should Communicate Capabilities and Limitations of AI*

Although CTA's "Guidelines for Developing Trustworthy Artificial Intelligence Systems" share a number of commonalities with the Framework, NIST may consider incorporating an additional element of CTA's guidance. In particular, CTA suggests that the Framework incorporate provisions directing purveyors of AI systems to effectively communicate the capabilities of their AI systems, and importantly, define the limitations of those capabilities. AI developers should deploy clear, useful, statements explaining the capabilities and limitations of their AI along with the specific use cases for which it was developed.

6. *Include Decommissioning in Lifecycle of AI Systems*

CTA recommends that the NIST contemplate the decommissioning and phasing out of AI systems, and offer baseline recommendations of risk management considerations when phasing-out the use of AI systems.

Comments on the "AI RMF Core": Mapping, Measuring, Managing and Governing

1. *Section 6.1 – Map Function*

Although the Map function provides a useful framework for mapping context, frameworks, risks, and goals of deploying AI systems, it does not properly consider the evolutionary nature of AI technology. One of the key hallmarks of AI and machine learning systems is that these systems evolve and "learn" as more data and outcomes are fed into the program to allow for additional training. Moreover, other attributes of this technology (such as feedback loops) impact how the algorithmic processes themselves learn from data inputs. Therefore, CTA suggests that the mapping component account for the evolving nature of this technology and be regularly reviewed and updated to reflect AI deployment experience and maximize positive impacts.

Further, the allocation of risk management responsibilities between stakeholders will vary depending on the nature of the Al system being developed and how it is deployed to customers. Specifically, while some systems are pre-trained and static at the time they are deployed, other systems can be customized around certain parameters using data provided by the customer. Further still, other systems may be tailored specifically for a particular end-user or use case utilizing data provided by the end-user. The Framework should consider these variations,

specifically how responsibility for mapping risks will be allocated between the developer and the end-user in each case.

Category 2 of the map function addresses the "Classification of AI Systems" and references "considerations related to data collection and selection." However, the selection and collection of data is a key attribute to potential risks arising later from system outputs. For that reason, CTA suggests that NIST expand the data selection and collection components to specifically include data risk mitigation strategies such as: (1) mapping or inventorying data; (2) classifying data and sourced datasets; (3) determining possible sources of corrupt or misplaced data and data sets; and (4) analyzing risks associated with the data sets.

CTA also recommends that NIST consider further defining key terms and concepts to narrow and focus the scope of the map function, including the definition of AI. Many existing proposed regulatory regimes either define AI quite broadly, or do not define the term at all. While providing a "one-size-fits-all" definition of AI is notoriously difficult, understanding what is within scope of the Framework would further define context. Elsewhere the Framework references risk taxonomy against other leading proposals (OECD, EU AI Act, Executive Order 13960) (see Figure 4, p. 9). Applying the same side-by-side analysis to other key definitions would help ensure NIST proposal is aligned with other frameworks.[7]

The Institute of Electrical and Electronics Engineers (IEEE) has released guidance for assessing the impact of autonomous and intelligent systems on human wellbeing. While it shares many similarities with NIST's draft Framework, it includes an additional element of its risk assessment that NIST may consider incorporating into its Framework. Specifically, the IEEE recommends that when mapping potential risks, developers should understand "the intended users, the unintended users and stakeholders" and assess whether they will be positively or negatively impacted by the AI as a result of incomplete or biased data sets, or algorithmic malfunctions. In line with this recommendation, CTA suggests that the Framework also propose when and how developers should consider unintended uses and impacts of AI systems.

2. *Section 6.2 - Measure Function*

The Measure function helpfully identifies various risk measurement methodologies that may be appropriate as stakeholders consider their development and/or use of AI systems. To help Framework users implement this function, CTA suggests NIST consider a few revisions to offer more concrete standards.

First, CTA recommends NIST adopt the use of impact assessments to evaluate risks. Impact assessments can be useful tools at the development and deployment stages of AI systems to assess potential risks, impacts, and intended outcomes. Assessments can, and should, address the categories and subcategories included in the draft Framework, and they are a nimble tool enabling stakeholders to make context-based and nuanced assessments, as the nature and purpose of AI systems may evolve over time. Furthermore, these assessments can evaluate the degree

[7] For example, in the Framework, NIST characterizes AI as "algorithmic processes that learn from data in an automated or semiautomated manner." Artificial Intelligence Risk Management Framework: Initial Draft, at 2.

(and necessity) of human oversight at each stage of development, and whether oversight should be adjusted based on identified potential risks.

Next, CTA notes that the use of "accuracy" as a technical characteristic as well as in Subcategory 1 of the function may be misleading and could perpetuate problematic algorithms. In the context of AI, "accuracy" is one of the many metrics that are used to assess the performance of a classification model. Incorrect choice of metric can lead to faulty assessment of the model. The use of this term in the AI RMF could mis-convey that the metric accuracy should be chosen to assess the performance of models across all cases. Instead, CTA recommends replacing "accuracy" with "metric approved for algorithmic use." When a developer chooses a metric, they must be able to stand by and defend it.

As NIST considers further development of the draft Framework, CTA suggests that the Framework explicitly recognize that not all risks can be anticipated in the initial development of AI systems. Some elements of risk cannot be assessed at inception, and instead, become apparent and/or measurable over time. Conversely, risks may decrease over time as technology improves or societal context changes. For example, risks relating to guiding principles and socio-technical characteristics can be addressed in some capacity at the development stage. However, characteristics such as accuracy, explainability, interpretability, and robustness cannot be assessed until an AI system has been at least partially developed based on learning from data inputs and outcomes. The Framework should account for the consistently evolving nature of risk measurement and encourage Framework users to engage in a continual process of risk measurement over the life of the AI system.

Relatedly, CTA reminds NIST that risks are not the same for every algorithmic model created by a particular developer or generated for a particular purpose, since each model is built differently from others. In Subcategory 2, CTA recommends clarifying that each model's risk should be evaluated independently.

Finally, CTA requests that NIST work with industry stakeholders to categorize risk levels with greater specificity. The draft Framework leaves it to organizations to assess risk and develop mechanisms themselves to mitigate such risk. While flexibility is critical to broad adoption of the Framework, collaboration between NIST and on-the-ground stakeholders to suggest of nimble categories of risk (*e.g.*, low, medium, and high) and corresponding levels of mitigation measures could improve the achievability of the Framework. Rather than introducing a vague set of standards, ungrounded in practice, NIST could draw from years of industry expertise to turn this function into a concrete, interoperable tool for better risk evaluation and targeted mitigation. Examples of "low" risk AI systems might include AI systems that select music for a listener based on their listening history. Higher risk examples might include AI systems that impact housing, health care, employment, or credit. Similarly, law enforcement uses of AI systems would also be high risk.

The Framework must not ignore that there are instances where risk cannot be measured. CTA respectfully asks NIST to provide guidance for those instances, including that the absence of an ability to measure risk does not imply that an AI system poses high or infinite risk. Additional clarity on these situations will ensure that the absence of measurement does not

automatically or necessarily result in halting the development or use of a technology—or the implementation of misplaced mitigation measures under an incorrect assumption of high risk.

But no matter the level of risk, there needs to be more algorithmic literacy to mitigate bias. All developers, deployers, and users (who are the subject of AI decision-making) "would benefit from knowledge of how these systems function. Just as computer literacy is now considered a vital skill in the modern economy, understanding how algorithms use their data may soon become necessary."[8] It has also been suggested that regulatory safe harbors could increase regulatory certainty for developers and operators of AI systems.[9] Prescribing methods for removing bias and mitigating adverse effects should be within a broad immunity so that developers and users are incented to mitigate.

3. *Sections 6.3 & 6.4 - Manage and Govern Functions*

Sections 6.3 and 6.4 of the draft Framework offer helpful standards for the management and governance of AI systems and the individuals and teams developing them.

In the Manage Function, CTA suggests two critical refinements. First, "impact" and "scale" should be defined in Subcategory 1 for greater clarity and implementability. Second, additional clarifications can also be made to Subcategory 2. CTA supports mechanisms for disengaging or deactivating AI that demonstrates outcomes inconsistent with intended uses. We recommend expanding Subcategory 2 to include "create a contingency plan for the deactivation of the AI," as such a plan is necessary to ensure there is no harmful halt in services.

CTA notes that the Govern Function—unlike Map, Measure, and Manage—is more difficult to quantify. Whereas the other categories and subcategories tie to direct actions and clear deliverables, the Govern examples offer less guidance on how fealty to the Govern function would be achieved, measured, and demonstrated. Therefore, CTA recommends that NIST continue stakeholder discussions on how the Govern categories and subcategories tie to outcomes or would be demonstrated to regulators.

To further strengthen the draft Framework, NIST should consider acknowledging the utility of training, awareness, and education and as a critical component of governance. Education of AI system enhances overall risk management and mitigation practices and supports a more holistic governance process. Such education is particularly important for personnel with risk management responsibilities and/or who are directly involved in systems development. These individuals should be empowered with sufficient authority and incentives to assess, escalate, and/or address risks. However, doing so requires sufficient training and awareness of AI risks and mitigation strategies.

The draft Framework could better distinguish between the developers and users of AI systems. CTA suggests that NIST collaborate with industry stakeholders to develop additional

[8] *See* https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[9] "For example, Section 230 of the Communications Decency Act removed liability from websites for the actions of their users, a provision widely credited with the growth of internet companies like Facebook and Google." *Id.*

guidance for allocating risks, responsibilities, and obligations between these two groups. What obligations belong to developers? Which belong to users? How should the two groups interact? Each actor in the system development, operation, and modification cycle has a distinct insight into the potential risks and has attendant risk spotting and mitigation responsibilities. Addressing both categories of stakeholders helps ensure comprehensive risk management and governance practices flow through the lifecycle of AI systems, and from developer to user (and any third-party vendors). Understanding the complex interrelationships between developers, users, and vendors will be a significant undertaking, and CTA recommends this as a particularly impactful subject for collaboration.

Relatedly, CTA recommends that NIST engage with stakeholders and advocates in the further development of Subcategory 5 (related to diversity, equity, and inclusion). Achieving actionable guidance for this laudable component of the Framework will require broad input from a variety of stakeholders.

Finally, NIST should consider explicit cybersecurity governance guidance for the auditing and monitoring of AI systems. Securing input data, models, and algorithms from tampering or unsupervised changes are necessary to further reliability (*e.g.,* that models produce anticipated outcomes) and protect against bad actors. Security governance should include ongoing audits and monitoring to confirm that systems behave as intended, have not experienced unauthorized meddling, and enjoy robust security to avoid adversarial attack. These should also address privacy and security considerations related to sharing data and models, such as between stakeholders, between private-and-public actors, and otherwise. Potential breaches or algorithmic corruption are not only an internal concern of AI developers, but broadly confront the entire technology ecosystem.

Respectfully submitted,

/s/ Douglas K. Johnson
Douglas K. Johnson
Vice President, Emerging Technology Policy

/s/ Michael Petricone
Michael Petricone
Sr. Vice President, Government and Regulatory Affairs