**Memorandum**

To: G Harris, Lauren Scheib, Gary Behrens, Donna McPartland, and John Kleeman

From: Alan Thiemann and William Snoeyenbos

Date: February 10, 2023

Subject: NIST Artificial Intelligence Risk Management Framework

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

This Memorandum discusses the key provisions of the initial version of the National Institute of Standards and Technology (“NIST”) Artificial Intelligence Risk Management Framework (“AI RMF), released on January 26, 2023, which contains voluntary guidelines for developers and users of AI (see NIST AI 100-1 AI RMF 1.0). This memorandum also compares the AI RMF with the comments submitted to NIST from the Association of Test Publishers (“ATP”) on the proposed AI RMF on March 17, 2022, and whether and to what extent key ATP recommendations were adopted.

NIST developed the AI RMF to better manage risks to individuals, organizations, and society associated with artificial intelligence (“AI”). However, the AI RMF does not compare favorably with the ATP recommendations. The ATP should consider further efforts to comment to NIST in an effort to obtain changes that will better serve the assessment industry.

1. **Definitions**
2. **Artificial Intelligence**

The final AI RMF introduces a new definition of an “AI system” that differs considerably from the draft definition of “AI”, which ATP contended was overly broad and imprecise. NIST now defines an “AI system” as “an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy” (*id*. at p. 1). This definition appears to be taken from the OECD Recommendation on AI released in 2019 and the ISO 22989 (2022) standard.

NIST partially adopted ATP’s suggested definition. The ATP proposed that an AI system “engages in learning, reasoning, or data modeling to reach outcomes.” The final AI RMF definition uses the phrase “generate outputs”, while the ATP proposed definition uses the phrase “reach outcomes.” These two phrases appear functionally equivalent. However, the final NIST definition does not seem to recognize the European Commission’s

“compromise” definition of AI from November 2021, which expressly excludes traditional software that is merely used to automate manual human tasks. That failure potentially represents is a serious issue for testing organizations under the AI RMF.

The final AI RMF definition is also narrower than ATP proposed definition because it specifies that AI is “engineered or machine based.” The AI RMF’s adoption of the term “AI system” also aligns with more frequently used technology vocabulary such as “information system.” Overall, the final definition of “AI system” is better than what NIST originally proposed – it is closer to the ATP proposed definition and more precise than the original draft definition, despite the failure to exclude automated software.

1. **Principles vs. Attributes**

The ATP comments recommended that NIST should clarify the distinction between “principles” and “attributes.” The final AI RMF does not make any attempt to clarify any distinction between these terms. Appendix D uses the word “attributes” to describe the way in which the AI RMF is intended to be used (*id.* at p. 42). However, the word “principles” also appears in the document interchangeably. Accordingly, the AI RMF offers no clarity on the meaning or differences between principles and attributes. This potentially creates a practical problem for entities trying to implement the NIST AI RMF.

1. **AI RMF Organization**

One of the most important comments by ATP focused on the fact that the proposed Framework failed to follow the structure of the NIST Cybersecurity Framework and the NIST Privacy Framework – both were built around a series of specific controls that would enable a user to evaluate its compliance with Framework. The ATP noted that the lack of a similar structure in the AI RMF would make it much harder for any organization to perform an integrated compliance against all three Frameworks.

The final AI RMF does not follow the structure of the NIST Cybersecurity Framework and the Privacy Framework, which means that NIST has not established a coordinated set of guidance around AI controls. Instead, NIST has proposed that users “fill in the gaps” by referencing other guidelines without any directions or guidance on which ones apply or how to apply them (*id*. at p. 33). The AI RMF also states that privacy and cybersecurity risk management considerations are applicable to AI and suggests organizations leverage the existing privacy and cybersecurity frameworks to manage risk – but since those Frameworks have different structures, their integration will be hard to achieve.

Instead of directly tying guidance to specific points of organizational “control,” the AI RMF is organized around four basic “functions”: 1) mapping; (2) measuring; (3) managing; and (4) governing *(id*. at p. 20). These four functions are further divided into categories and subcategories. While these categories and subcategories are not called “controls” in the AI RMF, they appear to function in a similar way by allowing a user to align its organizational development of AI to these functions, categories, and subcategories as a method of managing risk. Nevertheless,

despite the existence of these functional categories, it remains difficult to see how a testing organization will be able to neatly integrate its AI Risk Management with the NIS Cybersecurity and Privacy Frameworks in any comprehensive manner because they share no common foundation.

1. **AI Risk Discussion**

A major point of the ATP comments focused on the singular approach to risk management proposed by NIST, noting that a “one size fits all” approach is not appropriate, and urging that recognition of risks needs to be more segmented and nuanced, especially for the diversity of uses by testing organizations.

The AI RMF basically ignores the ATP recommendations, instead characterizing risk as “long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact” (*id.* at pp. 1-2). These characterizations are still completely binary and do not reflect any substantive changes to the original proposal contrary to the ATP’s response to the RFI. The final AI RMF does not introduce any sort of risk spectrum or tiered system for identifying and addressing low, moderate, or high levels of risk, or risk based on specific use cases, as recommended by ATP.

1. **AI Testing Profiles**

The AI RMF RFP proposed the concept of context- specific “profiles” and invited commenters to suggest sample profiles. This approach is carried forward in the AI RMF as a mechanism for evaluating risks; however, the AI RMF does not yet set forth any industry-specific profile templates. NIST stated that not including profile templates allows for “flexibility in implementation” (*id*. at p. 34). The AI RMF does organize profiles into two overall categories: **use-case profiles** and **temporal profiles.**

**Use-case profiles** are defined as “implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user.” The AI RMF states that examples include a hiring profile and a fair housing profile. **Temporal profiles** are defined as descriptions of either the current state or the desired, target state of specific AI risk management activities within a given sector, industry, organization, or application context. The AI RMF does not give examples of temporal profiles (*id*. at pp. 33-34).

ATP proposed an extensive assessment industry profile with various specialized applications in its RFI submission. The ATP proposal would seem to represent a use-case profile under the final AI RMF, but so far NIST has done nothing about its adoption or provided any feedback/reaction to the ATP recommendations. We understand that NIST intends to provide details on profile templates later this year, perhaps as early as the Spring.

1. **AI Actor Tasks**

Appendix A of the AI RMF contains descriptions of AI actor tasks listed in Figures 2 and 3 (*id.* at p.11). Among the most informative risk management strategies in the NIST AI RMF is guidance on the conduct of “Test, Evaluation, Verification, and Validation (“TEVV”) tasks (*id.* at pp. 35-36). NIST has identified the following five guidelines:

* TEVV tasks should be performed throughout the AI lifecycle by AI actors who examine the AI system or its components or detect and remediate problems. Ideally, AI actors carrying out verification and validation tasks should be separate from those who perform test and evaluation actions. Tasks can be incorporated into a phase as early as design, where tests are planned in accordance with the design requirement.
* TEVV tasks for design, planning, and data may center on internal and external validation of assumptions for system design, data collection, and measurements relative to the intended context of deployment or application.
* TEVV tasks for development (i.e., model building) include model validation and assessment.
* TEVV tasks for deployment include system validation and integration in production, with testing, and recalibration for systems and process integration, user experience, and compliance with existing legal, regulatory, and ethical specifications.
* TEVV tasks for operations involve ongoing monitoring for periodic updates, testing, and subject matter expert (SME) recalibration of models, the tracking of incidents or errors reported and their management, the detection of emergent properties and related impacts, and processes for redress and response.

We believe that these tasks provide useful guidance for testing organizations to follow, especially if the organization itself is developing/designing an AI system. But even if the testing organization is merely implementing someone else’s AI system, these tasks offer a detailed pathway for vetting an AI system before or during its implementation.

**Conclusion**

The ATP comments to NIST provided a strong set of recommendations for adoption of a US risk management standard that would be useful along side of the NIST Cybersecurity and Privacy Frameworks. While that result has not been achieved at this point, we will continue to share industry information with NIST in the hopes of obtaining an “Assessment Industry Profile” when NIST finally decides how to proceed in that arena. In the meantime, ATP continues to work with the US Chamber of Commerce in seeking to advance an appropriate risk management framework in the United States. Without such a result, we fear that state legislatures/regulators will impose overly burdensome AI requirements on testing organizations.