August 19, 2021

National Institute of Standards and Technology
Alicia Chambers, NIST Executive Secretariat
100 Bureau Drive
Gaithersburg, MD 20899

To whom it may concern,

Monitaur is pleased to have the opportunity to offer our responses to the NIST Artificial Intelligence Risk Management Framework. This important project has the opportunity to accelerate effective governance and assurance of artificial intelligence (AI) and machine learning (ML) systems. We believe that by creating more trust and confidence in how these technologies are used and managed, all stakeholders – corporations, regulators, and consumers – will benefit from extraordinary innovations that are more fair, safe, compliant, and robust.

Below you will find our detailed recommendations listed by question. The key takeaways are:

- AI actors need to create comprehensive programs for ongoing governance and assurance across all of their AI systems in order to protect themselves and society from the unique risks that AI presents.
- Risk management of AI demands a holistic approach that incorporates human and process oversight, not just model or data risk management.
- AI actors can use the methodology of Machine Learning Assurance, which takes advantage of the established, effective CRISP-DM framework already familiar to many organizations in order to accelerate adoption and education.
- Achieving transparency, fairness, and accountability with AI systems will require organizations to pursue context, verifiability, and objectivity as the primary goals of their governance and assurance efforts.

We look forward to participating in NIST's AI RMF in the future.

Sincerely,
Andrew Clark
Chief Technology Officer

Monitaur, Inc.

Responses to RFI Questions

1. The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks;

The greatest challenge that organizations face in managing risks is not having a holistic approach and centralized system for governing all the artificial intelligence (AI) and machine learning (ML) models throughout their lifecycles. When produced, documentation about business decisions and models themselves often resides in silos or is scattered across internal document stores, personal computers, and email inboxes.

In addition to holistic risk and control management, the lack of business understanding and data understanding are major challenges for many organizations. Heavily regulated financial institutions that are regulated by the OCC and the Federal Reserve and are subject to OCC 2011/12 have the basis for AI risk management, but often fall short in appreciating the unique challenges of assuring AI systems. Key inhibitors of successful risk mitigation include 1) the lack of business-level understanding of what risks are created by these systems, and 2) the absence of quality objective challenge by independent reviewers inside the organization. Without a solid understanding of 1st line, 2nd line, and 3rd line functions and the skill sets required for an objective challenge, it is exceedingly difficult to adequately manage AI- and ML-related risks.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

The proposed Framework and characteristics have alignment with many other proposed standards emerging across sectors and geographies and generally cover the key considerations.

Machine Learning Assurance (MLA) is a controls-based process for machine learning (ML) systems that establishes confidence and verifiability through software and human oversight. The objective of MLA is to assure important stakeholders that an ML system is functioning as expected. In more detailed terms, it assures governance and oversight of an ML system's transparency, compliance, fairness, safety, and optimal operation.

Creating assurance for AI and ML systems requires a continuous, coherent approach throughout the lifecycle of projects, as well as across enterprise operations. Careful coordination of people, processes, and systems can create the clarity, confidence, and accountability that practitioners need and regulators should expect.

Organizations can utilize the established, effective CRISP-DM steps and deploy detective controls vital for machine learning systems to drive a powerful framework for assurance and robust risk/control matrices. Three core principles empower an effective MLA program and responsible use of AI/ML.

- **Context**: MLA requires a clear understanding and documentation of the considerations, goals, and risks evaluated during the lifecycle of an ML application.
- **Verifiability**: MLA requires each business and technical decision and step to have the ability to be verified and interrogated.
- **Objectivity**: MLA requires that any ML application can be reasonably evaluated and understood by an objective individual or party not involved in the model development.

To fulfill the principles of Context, Verifiability, and Objectivity, an independent, objective party – such as a competent, unbiased, and properly compensated internal or external auditor – must be able to read the documentation surrounding the AI/ML system and understand the goals of the system, trade-offs, and why it was built the way it was. The objective party must also be able to reperform individual model predictions and verify that the results are identical to existing predictions.

Additionally, after obtaining Context, and Verifiability of the system in question, the objective party needs to perform a battery of tests against the model to verify that it is performing as expected and is unbiased. If an independent party that has not seen the model previously can perform all the steps outlined below, objective assurance can be obtained.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

Section two elaborates on the additional principles of Context, Verifiability, and Objectivity. Without these three principles, transparency, fairness, and accountability are not fully possible.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

From our vantage point, many organizations have yet to adequately incorporate AI risks in their overarching enterprise risk management. Companies with model risk management functions often lump AI/ML into their existing frameworks, incorrectly taking the approach of "nothing new here." Other organizations that historically may have relied less on models often have yet to fully incorporate AI/ML risks into their enterprise risk management program.

It is worth noting that much of the aforementioned MLA methodology derives from fundamental principles of enterprise risk management. Establishing documentation, evidence, audit trails, and objective reviews are not new concepts within enterprise risk management frameworks; however, the

concepts are not being adapted to the unique needs of AI and ML systems as intentionally they can be.

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

Organizations can utilize the established, effective CRISP-DM steps and deploy detective controls vital for machine learning systems to drive a powerful framework for assurance and robust risk/control matrices. These controls should be aligned with the risks presented across six stages of an AI/ML project.

CRISP-DM consists of six steps:

1. Business Understanding (revisited in later steps)
2. Data Understanding
3. Data Processing
4. Modeling
5. Evaluation
6. Deployment

If the process outlined above in section 2 can be carried out by an objective individual, then an organization has gone a long way toward mitigating AI/ML related risks.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

The proposed EU AI regulatory framework is the most advanced and holistic approach by a regulatory agency to date: Europe fit for the Digital Age: Artificial Intelligence.

The EU AI regulation is an extensive, all-encompassing, and ambitious proposal that strives to establish a risk-based framework for ethical, responsible, and safe AI without preventing its use or blocking its progress. The proposed framework includes several risk tiers (unacceptable, high, or low) and specifies what steps are required to manage the risk at the different tiers.

Requirements for deploying high-risk systems include items such as high-quality data sets, documentation, record-keeping, transparency, human oversight, as well as system checks for accuracy, security, robustness. An important part of the framework is the requirement for a third-party conformity assessments audit prior to placing a product into the marketplace and whenever it changes substantively.

While there may be elements of this regulation which in practice become limiting to the practice of data science and machine learning, we believe there are many principles and tactical approaches which drive Objectivity, Verifiability, and Context.

The US Federal Trade Commission (FTC) has signaled a regulatory stance on bias algorithms under the existing FTC regulations of the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA), and the FTC Act. These three regulations prohibit companies from using deceptive or unfair sales practices, credit discrimination that is biased against a protected class, and denying housing, employment, or insurance that is discriminatory of protected class membership.

The National Association of Insurance Commissioners (NAIC) has adopted the OECD's AI principles, and is looking into a stricter enforcement posture: NAIC Unanimously Adopts Artificial Intelligence Guiding Principles

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;

As mentioned above, the proposed EU AI framework should be considered. Another valuable framework was created by the Information Commissioner's Office in the United Kingdom: ICO AI Auditing Framework. The ISO's ISO/IEC JTC 1/SC 42/WG 3 "Trustworthiness" working group is developing standards as well and should be part of NIST's literature review.

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

The critical CRISP-DM Evaluation step of model construction and specific model validations is often performed by the same individuals who created the model. In financial and other companies with mature modeling teams, model risk managers may complete the Evaluation step; however, many of these individuals rely solely upon aggregated statistical evaluation techniques to determine if a model is performing as expected. In the case of AI and ML, it is vital that such aggregate techniques are complemented with audits at the individual transaction/decision level.

Instead, a robust, cross-functional team with independent compensation mechanisms should thoroughly evaluate all machine learning models prior to deployment to check for inappropriate biases and "human friendliness" of models. This cross-functional team should also be responsible for reevaluating the models regularly since the models evolve over time.

The team's work should consist of performing sensitivity analysis and "poking holes" in the model to ensure that all risks are mitigated and the model is doing what is supposed to be before it is signed off on by project owners and deployed into production. Performing independent random sampling of target demographic and socioeconomic groups and running individual transactions through the model to evaluate if equalized odds and disparate impact considerations hold is a time-consuming but extremely effective method of checking for bias.

In an ideal world, the team constructed around modeling systems can represent a diversity of thought, experiences, and knowledge to help build a system with comprehensive and varied input; however, we are likely to be in a marketplace of talent for coming years that is not as diverse or distributed as some may hope. Therefore, Objectivity from cross-functional teams ensures a distribution of review and thought beyond those few building the models.

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");

NIST's proposed attributes of transparency, fairness, and accountability are excellent foundational principles. As discussed in section 2, the addition of Context, Verifiability and Objectivity are recommended as important goals for delivering upon those principles.

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include—but are not limited to—the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and

We recommend that CRISP-DM, as discussed previously, is used as the basis for the AI Risk Framework as it has 6 main parts, is highly configurable, and maps well to other frameworks, as NIST's Cybersecurity framework does. If structured in a spreadsheet, as NIST's cybersecurity framework, CRISP-DM's 6 sections could be grouped with risks enumerated with mitigating controls outlined. More detailed technical descriptions/procedures as well as mapping to EU AI, and other frameworks/regulation could be accomplished.

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

By showing what the key risks and controls are for AI/ML, what is needed to mitigate them, and what skillsets and job functions are best suited to each role in the process, NIST can greatly help organizations build proper AI/ML development and risk management functions, as well as help to alleviate the ambiguity that exists for many organizations.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

NIST can and should provide the overarching governance structure for all industries and stakeholders to take advantage of by providing basic recommendations of how to mitigate risks of AI and ML systems and best practices for building teams, monitoring and evaluation of models, etc. that are generally applicable across industries and levels of organizational maturity.

Governance issues around grievance and redress should be tackled by companies in partnership with industry regulatory bodies since the specifics and implications vary greatly across industries.

A Framework which does not include expectations for risk mitigation across the full lifecycle of AI projects, which is more than just data and models, would be insufficient.