Deloitte & Touche LLP
1919 North Lynn Street
Arlington, VA 22209

Tel: +1 (571) 882-5000
www.deloitte.com

National Institute of Standards and Technology
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899
Submitted electronically to AIframework@nist.gov

To Whom It May Concern,

RE: **Deloitte & Touche LLP Comments on "Artificial Intelligence (AI) risk management framework"**

Deloitte[1] appreciates this opportunity to submit comments in response to NIST's Request for Information regarding considerations that should be included in an AI risk management framework. As one of the largest professional services organizations in the United States, Deloitte provides a vast array of information security services across 2,800 engagements in major commercial industries and 15 cabinet-level federal agencies. We serve our clients by helping them align their security and privacy investments with business risk priorities. Our comments reflect our deep experience with customers who are applying leading-edge artificial intelligence processes across their businesses and government departments.

Respectfully submitted,

Matt Baker, Managing Director

Deloitte & Touche LLP
Deloitte Government and Public Services
Risk & Financial Advisory

[1] As used in this document, "Deloitte" means Deloitte & Touche LLP, which provides audit and enterprise risk services; Deloitte Financial Advisory Services LLP, and their subsidiary Deloitte Transactions & Business Analytics LLP, which provides financial advisory services; and Deloitte Consulting LLP, which provides consulting services. These entities are separate subsidiaries of Deloitte LLP. Deloitte & Touche LLP will be responsible for the services and the other subsidiaries may provide services pursuant to an inter-organizational transfer. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

Contents

Introduction

Deloitte applauds NIST's efforts to develop a framework to manage risk related to artificial intelligence (AI) processes. As with the agency's efforts around cybersecurity and privacy, we expect that NIST will help to establish "guardrails" for AI that will engender a trustworthy marketplace that can innovate without causing harm. In doing so, it will be important to consider both deliberate causes of harm as well as those that may be caused inadvertently. We believe that using a framework for AI risk management is the correct approach and we look forward to contributing to its development.

Deloitte is a leading consulting firm in the field of AI, with 27,000+ AI-skilled practitioners, 300+ ecosystem teaming partnerships, and advanced analytics efforts across all 15 cabinet-level federal agencies. Our responses here are based on years of experience of supporting government and commercial clients implement AI processes. Throughout our response, we will draw upon this experience and leverage examples of Deloitte's Trustworthy AI™ (TAI) Framework.

We encourage NIST to leverage principles already incorporated into other frameworks such as the NIST Cybersecurity Framework (CSF) and the NIST Privacy Risk Management Framework, as well as the five principles embodied in the Committee of Sponsoring Organization (COSO) Framework: governance; strategy; performance; review & revision; and information, communication & reporting.[2] We may anticipate that the NIST CSF already accommodates some of the required principles for AI, as does the Privacy risk management framework. We encourage NIST to continue its outcome-based approach for AI, so that there are clear objectives for developers and integrators, and which can be satisfied in a number of ways.

1. Challenges Observed in Industry

Through Deloitte's direct engagement with the burgeoning AI industry, we have observed several challenges which might be mitigated or solved through the development of an AI risk management framework. A common denominator across many of these challenges is whether the implementing organization has a dedicated and integrated governance program. We categorize AI risks into the following three areas where a structured risk management framework with transparent, documented, and defensible processes would be beneficial:

- **Data Governance.** Data is a critical requirement for artificial intelligence, and the quality, quantity and source are important factors for AI/ML model performance. Organizations

[2] *See* Committee of Sponsoring Organization (COSO) Framework.

should pursue those data governance mechanisms and protocols as the initial phases of an effective AI risk management framework. For successful implementations, organizations must evaluate what data is needed to develop AI. Key data governance considerations include:

1) Representation of the appropriate population for the AI use case and reduction of bias;
2) Clear rules for using and disseminating data, including data collection, data quality evaluation, disclosure of use, and disposal; and finally,
3) Means of securing data assets.

- **Ethical Governance/Risk.** While AI practitioners may be aware of the importance of ethical practices, organizations are often ill-equipped with adequate tools, methodologies, or metrics for implementing them.[3] A significant challenge is the 'operationalization' of strategic level principles into tactical directives for data scientists, software developers, and section heads. This has implications for strategy and objective setting as well because it means the translation of specific strategic goals into mid-level norms down to low-level requirements.[4] Governance and training is necessary to bridge this gap at the institutional level.
- **AI Risk Models.** To account for different AI model risk profiles, organizations should perform risk assessments to solidify the business case and identify operational, reputational, regulatory, and adversarial risks related to AI initiatives to reduce exposure and identify opportunities to create value. Organizations also need to prioritize risks by assessing AI models that are part of their larger AI initiatives and determining the level of accuracy, reliability, and transparency required for the related use case(s).

Organizations adopting AI are making significant investments in AI development and implementation, and must align their AI risk management with broader risk management efforts. These include inventorying, benchmarking, and trends analysis. After implementation, the organization should practice continuous monitoring, measurement, and communication of the risks it is accepting to determine whether its practices and/or objectives might be achieved with lower risk.

2. Observations & Gaps – Characteristics, Principles, and Standards

AI and its implementation continues to mature and change. We have observed differing views on the characteristics, principles, and standards that are under development across the ecosystem, and we recommend that NIST clarify the distinction between characteristics and principles when it comes to the development of an AI RMF. Our understanding is that characteristics pertain to the distinctive qualities and descriptors of ideal trustworthy AI systems, while principles refer to the underlying objectives and philosophies that guide the development of frameworks around them. This section provides our view, observations, and highlights gaps where we have identified as they relate to the building blocks of a trustworthy AI framework using this distinction.

[3] See Brent Middlestadt, "Principles alone cannot guarantee ethical AI," p. 10, Oxford Internet University, the Alan Turing Institute (2019).

[4] Andrew Burt, "Ethical Frameworks for AI Aren't Enough," *Harvard Business Review – Technology*, (November 2020), accessible at: **https://hbr.org/2020/11/ethical-frameworks-for-ai-arent-enough**

Characteristics

NIST has identified accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI as critical characteristics of trustworthy AI. In Figure 1 we suggest definitions and provide the rationale for the importance of the identified characteristic.

Figure 1: Definitions and Rationales of Proposed Characteristics

Proposed	Definition[5]	Rationale
Accuracy	Accuracy refers to an AI system's performance. Accurate models optimize for the most relevant performance metrics for their use case.	Accurate models result have better predictive power and ability.
Explainability and Interpretability	Explainability and interpretability address trust concerns inherent in the "black box" nature of AI models, and refer to the ease in which a stakeholder can comprehend how a model makes predictions.	Stakeholders are often required to validate model inferences, particularly in settings that have regulatory impact. Explainable AI models enables decision makers to better understand and have confidence in models.
Reliability	Reliability is the ability of a system to perform its required functions under stated conditions for a specified period of time.[6]	AI is widely used in safety-critical autonomous systems. Unreliable AI models could lead to adverse consequences.
Privacy	Privacy considerations involve collecting, processing, sharing, storage, and disposal of personal information.[7] The net-new privacy requirements for AI beyond those considered in PRMF should be discussed.	Privacy is protected by many jurisdictions and laws. Organizations that fail to protect data privacy can be subject to fines and reputational damage.
Robustness	Refers to the degree to which a model deteriorates over time from development to deployment. A robust algorithm experiences less deterioration and produces consistent outputs.	A robust AI model performs more consistently across training, testing, evaluation and real-world deployment.
Safety	Safety is the condition of the system operating without causing unacceptable risk of physical injury or damage to the health of people.[8]	Unsafe AI models may have adverse consequences including destruction of property, operational service disruption, or even injury or death.
Security/ Resilience	Resiliency can be defined as the adaptive capability of an AI system in a complex and changing environment. NIST should consider the maturity of most existing Cybersecurity risk management frameworks and not attempt to establish a new AI specific cyber framework. Profiles of the NIST CSF and additional requirements in the RMF (NIST SP 800-53) should be the preferred mechanisms for addressing cyber risks related to AI.	A secure and resilient AI system supports continuity of operations that rely upon AI, requires less manual intervention to respond to shifts in data or use cases, and can help safeguard data and model outputs from unintended disclosure.

[5] These definitions are reflected in DARPA and IBM. Explainability can be related to Predictability in NISTs PRMF.
[6] (ISO/IEC 27040:2015). As mentioned in NIST Risk Management Framework, there may be triggers identified that force a re-evaluation of AI principles.
[7] This is aligned with Deloitte's Trustworthy AI™ framework as discussed in MIT Technology Review's coverage of the framework.
[8] This definition is cited in the above NIST publications from the ISO/IEC Guide 55:1999.

Proposed	Definition[5]	Rationale
Mitigation of Harmful Bias	The practice and deployment of techniques to measure, evaluate, and respond to bias in data and algorithms.	AI can perpetuate discriminatory biases in training data at scale without proper bias safeguards.
Harmful outcomes from misuse of AI	This includes unintentional or intentional harm to protected classes, unethical deployment against organizational principles, and irresponsible replacement of human decision-makers.	Misuse of AI can lead to operational service disruption, reputational damage or cause injury or death.

We propose additional characteristics of AI trustworthiness in Figure 2: Proposed Characteristics:

Figure 2: Proposed Characteristics

Proposed	Definition	Rationale
Accountability	Organizational structures and policies should be developed to clearly determine responsibility for the output of AI system decisions. Key factors to consider include which laws and regulations might determine legal liability and whether AI systems are auditable and covered by existing whistleblower laws.	Accountability for the decisions and outcomes of AI systems incentivizes stakeholder adherence to trustworthy AI guidance, and provides central points of contact for mitigation efforts.
Data governance / preparation	ISO/IEC 38500[9] defines a framework to address principles with regards to good corporate governance of IT. The framework comprises definitions, principles and a model. It sets out six principles for good corporate governance of IT: Responsibility, Strategy, Acquisition, Performance, Conformance, Human behavior.	The dependence of AI models on data precipitates the need for comprehensive data governance and preparation guidelines.
Documented and auditable	Models should be well documented such that auditors, external organizations or audiences can understand the development process and how it is maintained in production. INTOSAI defines this characteristic as an important facet of being able to audit algorithms	The development of trustworthy AI requires vigilance and effort throughout every stage of the model development lifecycle. Third-party reviewers should be able to trace and understand the steps made to create each model.

The addition of these characteristics to NIST's currently identified list of characteristics will provide a sound foundation for understanding and evaluating risk and trust in the context of AI.

3. Standards and Principles

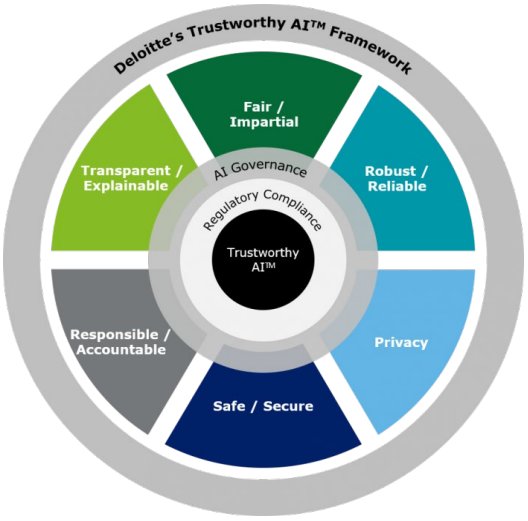
Deloitte has extensively evaluated where the market currently stands in defining, managing, and measuring characteristics of AI trustworthiness. We have identified six considerations that are critical in helping safeguard against risk and build a trustworthy AI strategy for an organization. These are reflected in Deloitte's Trustworthy AI™ framework (Figure 3): (1) Fair/Impartial, (2) Transparent /Explainable, (3) Robust/Reliable, (4) Privacy, (5) Safe /Secure, and (6) Responsible/Accountable. While the first five pillars overlap with the NIST characteristics, the additional Responsible/Accountable pillar focuses on ensuring that policies are in place to determine who is held responsible for the output of AI system decisions. Our extensive experience with Model Risk Management[10] has led us to focus on this key area. Risks that may rise from this

[9] https://www.iso.org/standard/62816.html

[10] In particular, the Supervisory Guidance on Model Risk Management (SR 11-7) issued by the Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency

pillar include poor AI model oversight, poorly defined roles and responsibilities, uncontrolled or undocumented model changes, and a general inability to consistently apply AI across an organization.

Figure 3: Deloitte Trustworthy AI™ Framework



Best Practices and Principles

While the Trustworthy AI Framework helps define important characteristics in trustworthy AI, there are foundational prerequisites that must be considered before organizations can begin to address trustworthy AI issues. They are as follows:

1. **Use of Data as a Strategic Asset:** It is crucial for organizations to treat and identify data as a strategic asset. The appropriate treatment of it will ensure that data is securely transported between hardware, software and personnel. Accordingly, organizations can take steps to protect the strategic assets as any physical, monetary, or intellectual assets.
2. **Overcoming Technological Barriers to Entry:** There are numerous technical barriers to entry into implementing trustworthy AI principles. Driving down these barriers can allow organizations to begin ethical and responsible AI considerations. Platforms such as Deloitte's Trustworthy AI Platform that operationalize the Trustworthy AI™ Framework can be leveraged to help organizations address trustworthiness in a single platform.
3. **Access to Artificial Intelligence Solutions:** Organizations need to be able to develop and leverage a diverse set of solutions with risk-appropriate scale and speed.
4. **Cultivated and Informed Workforce:** To promote that the consideration of these principles are incorporated throughout the entire AI development process, an informed workforce with a cultivated culture of trustworthy AI among technical and non-technical staff is required.

Tools

As ethical AI lies at the intersection of many disparate disciplines, the path to trustworthy and responsible AI at scale will require not only qualitative management frameworks (such as an AI RMF, standards, and leading practices), but also powerful supporting technical tools. NIST's development of the former will provide the appropriate standards by which to evaluate and mitigate risks created by deployed models. The NIST AI RMF can also lay the foundation for helping organizations develop transparent, documented, and supportable processes for understanding and scrutinizing how models perform. At the more granular level of model development, the ability to leverage open-source and commercial packages and tools to operationalize AI risk frameworks becomes critical.

The programmatic and algorithmic tools that monitor and remediate AI risks can be classified by the AI model development lifecycle phase they target: the data input phase, when data is collected, processed, and collated for downstream consumption and usage; the modeling phase, during which an AI model is trained and built; and the model disposition phase, the point at which a model is applied to new data to produce an output, such as a suggestion or predicted likelihood.

During the data input phase, pre-processing techniques such as reweighting and fair representation techniques can be applied to allow model developers to curate the input data that are used to drive predictions. These can de-correlate or de-bias a dataset for use in any arbitrary machine learning algorithm to make predictions about a given outcome. In the modeling phase, in-processing tools

can leverage optimization and regularization techniques to address bias risks during AI model training. Given a protected attribute and a predicted outcome, these techniques penalize a model if the predicted outcomes disproportionately favor one group over another. This penalization is then used to augment the model training process by requiring the model to not just learn relationships that minimize prediction error, but also reduce inequality among disparate groups of data points. Finally, post-processing techniques can be employed to mitigate AI risk by editing model inferences. These techniques can help make informed decisions of where to draw a threshold on inference scores to maximize parity, prediction accuracy, or other relevant metrics for the use case. Post-processing techniques can also be employed to automatically adjust predictions to attain specified balance goals.

5. Relationship to Enterprise Risk Management, Risk Management Standards, Frameworks, Guidelines, and Models

AI risk management cannot and should not exist in a vacuum. Any AI Risk Management Framework needs to acknowledge and operate in concert with overarching Enterprise Risk Management frameworks such Committee of Sponsoring Organizations (COSO) enterprise risk Management framework. Additionally, NIST should build the AI RMF to align with and even build off of existing domain specific risk management frameworks such as the NIST Cybersecurity Framework, NIST Privacy Risk Management Framework, the Federal RMF.

The NIST Cybersecurity Framework & The NIST Privacy Risk Management Framework

We recommend structuring the Trustworthy AI framework as a flexible risk management tool, similar to the NIST Privacy and Cybersecurity Frameworks, to help organizations adopt and create innovative AI solutions while minimizing adverse consequences for individuals. Each principle in the framework (see DAI and our recommendations for Trustworthy AI principles in question 3) should correspond to specific outcomes and metrics that organizations can implement and measure in alignment with their priorities and risk thresholds to track progress and maturity over time.

While adoption of AI across organizations creates new sources and degrees of risk, key privacy and security concerns continue to revolve around proper processing and protection of sensitive information, making NIST Privacy and Cybersecurity Frameworks applicable. For example, use of AI would still require organizations to be transparent about data processing (Communicate Function in PRMF), decrease association/identification of individuals (Control Function in PRMF), and protect data used for the algorithm from unauthorized uses (Protect Function in PRMF). Therefore, structuring the AI RMF framework like the NIST Privacy and Cybersecurity Framework while updating the latter to accommodate any additional AI nuances will facilitate the use of all these frameworks together. A depiction of how the existing NIST frameworks apply to the Trustworthy AI RMF can be found in *Figure 4*.

Figure 4: Interaction of Existing NIST Frameworks

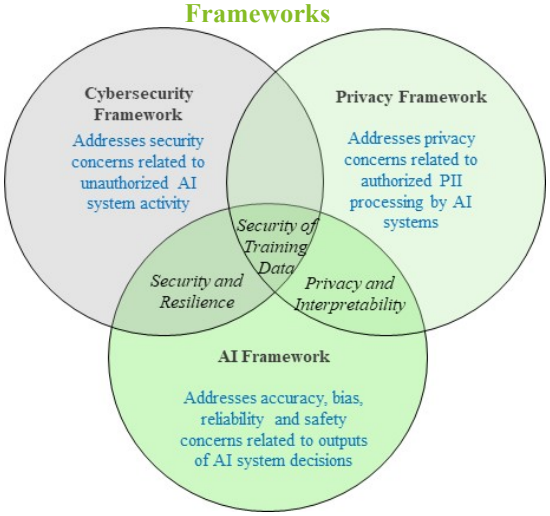


Figure 5: Mapping to NIST RMF

AI Lifecycle	RMF Steps	Example Security Activities
Research and Design	**Release Planning** (RMF Step 1 and 2)	Embed AI security & privacy SMEs; prioritize user stories based on current environment and mission risk scoring; facilitate design sessions including security architecture reviews
	Continuous Build (RMF Step 3)	Data Scientists access to certified, approved containers and ML libraries for common user risk stories from code management tools; security enforcement and automated configuration checks for container images hosts, and serverless functions
Develop, Train and Deploy	**CI/CD** (RMF Step 4 and 5)	Evaluate code of AI systems for test coverage; identify vulnerabilities with secure code review tools, e.g. underlying infrastructure scans, penetration testing, and dependency analysis; break builds on security flaws or exceeded threshold of flaws to provide a near-real-time feedback to Data Scientists; validate data sources monitor registry information to ensure secure deployment of code and implement AI model image signing.
Operate and Maintain	**Operations and Monitoring** (RMF Step 6)	Protect and monitor running AI systems with firewalls, network inspection, access control, runtime defenses, and robust logging.

Risk Assessments Guidelines and Frameworks (e.g., NIST SP 800-30)

As described in the first section, effective AI Risk Management requires an approach where AI risks are incorporated into organization's enterprise risk management initiatives where the severity of AI risks are measured to be prioritized for remediation. Risk details are aggregated into a statistical model, such as a Monte Carlo simulation to model various events based on the latest threat intelligence, tactics, techniques and procedures (TTPs) and vulnerability assessment results.

AI models have different risk profiles – AI models that require a high level of accuracy, reliability, or transparency to achieve success likely have a high risk profile. AI models that are used to provide recommendations for a low-impact decision (e.g., music recommendation) may have a lower risk profile than an AI model that is being used to automate decisions previously undertaken by humans (e.g., deciding on underwriting terms for an insurance policy).

NIST may design its own AI RMF related to the following ***risk response categories***:

- *Accept*: This may be appropriate when the risk to strategy and business objectives is within risk appetite limits. Risk that is outside the organization's risk appetite and that management seeks to accept will generally require approval from the board or other oversight bodies.
- *Avoid*: Action is taken to remove risk, which may mean not using the AI model, limiting the scope of use of the model, or modifying the functionality of the model to reduce complexity.
- *Pursue*: Increased risk to achieve improved performance is accepted. This may involve expanding the scope of use of AI models or modifying the functionality of the AI model to increase complexity. When choosing to pursue risk, the nature and extent of any changes required to achieve desired performance while not exceeding the boundaries of acceptable risk tolerance must be understood.
- *Reduce*: Action is taken to reduce the severity of the risk. This involves, establishing business processes and controls that reduce risk to an amount of severity aligned with the risk profile and appetite. *See the following paragraphs for actions that organizations may take that can reduce risk associated with AI models.*
- *Share*: Action is taken to reduce the severity of the risk by transferring or otherwise sharing a portion of the risk. A common example is outsourcing development, implementation, or monitoring of AI models to specialist service providers.

Integrating ethical considerations into an RMF would give NIST the ability to monitor, inform, and set policy according to trustworthy and responsible AI principles already adopted by the Federal government. Through transparent, documented, and defensible processes, an ethical framework would be an approach to help inform those organization who come to NIST for its expertise where either latent or obvious ethical equities exist across an enterprise and how to mitigate them.

Committee of Sponsoring Organizations (COSO)

The COSO enterprise risk management framework is an established and effective framework to identify risks, especially those related to compliance. The five core principles mentioned in the introduction of this response, can and should be applied to AI initiatives, to help organizations to begin addressing risk.

6. Recommendations for Establishing, Implementing, and Governing an AI Framework

NIST has a long history of creating flexible, outcome based, and highly adoptable risk management frameworks. Deloitte highly encourages NIST to make use of leading practices implemented during the development of similar risk management frameworks, including the NIST Cybersecurity Framework and the NIST Privacy Risk Management Framework. As a Frequent Partner in the development of such Frameworks, Deloitte provides inputs on the framework attributes and governance.

Framework Attributes

These attributes capture essential needs of an AI RMF and are adequately general that the framework can be used in a sensible, practical, and principle-based fashion across organizations spanning all sizes, background and domains. The attributes align with the considerations behind the development of the Deloitte Trustworthy AI™ framework, which we have found is resonating deeply both internally and with clients across many industries in the market.

Figure 6: Comments and Considerations on Framework Attributes

Comments and Considerations on Framework Attributes	
Consensus-driven; developed and regularly updated through an open, transparent process	The field of AI ethics and trustworthiness moves at a rapid pace, and a common framework must adapt to new findings and suggestions from industry and academia. The process of consensus will spark discussions and refine theoretical trustworthy AI theory into actionable guidance. From the perspective of AI practitioners and institutions, it is equally important to understand how such guidance is formed.
Provide common definitions	Being a nascent field, terminology in AI ethics is often confusing, redundant, and overwhelming to the uninitiated. Common definitions must be established for an AI RMF to be valuable across institutional endeavors. Using AI fairness as an example, it would be beneficial to have standardized mathematical and plain language descriptions of common fairness metrics along with examples of their significance.
Combining plain language with technical depth	The AI development lifecycle must be documented in a language that is understandable to a broad population of sponsors, SMEs in other fields, developers, clients, end users, and other invested parties that may not be familiar with the inner workings of advanced AI models.
Adaptable to different organizations, AI	AI models are deployed across industries and use cases, and an AI RMF should aim to be generally agnostic of such, focusing on principles of ethical AI. We note that the EU's approach in the European Commission's Artificial Intelligence Act[11] which

[11] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Comments and Considerations on Framework Attributes	
technologies, lifecycle phases, sectors, and uses	*"imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety"* and identifies said potential sectors and uses cases can provide value. AI issues such as bias may be subtle and difficult to notice or identify, and explicit guidance for certain use cases may effectively spread awareness for developers and stakeholders in those industries.
Risk-based, outcome-focused, voluntary, and principle-based	These attributes are consistent with our approach to Model Risk Management (MRM) and quality risk management, and well-aligned to successful adoption of the NIST CSF.
Integrated with enterprise risk management strategy and processes	Deloitte's Trustworthy AI™ Framework draws from our expertise in model risk management and quantitative risk modeling. This integration enables us to intuitively and methodically approach ethical AI issues both within and outside of our organization.
Consistency with other AI risk management frameworks	AI frameworks are being developed by institutions, governmental agencies, and countries. Regular surveys of regulatory developments across these and integration where possible would reduce confusion and promote adherence.
An evolving framework	The fast-moving pace of the field of ethical AI requires a iterative approach to framework development. A living AI RMF will allow NIST the ability to continuously and quickly respond to changes and needs.

Governance

Organizational governance is embedded in organizational structure, culture, and technology. *Structure* includes software and engineering development processes, their integration with larger business development cycles, and organizational incentives and hierarchy. Organizational *culture* may be nuanced by corporate vision, employee interaction, and the organizations values, but what is important is that the organization considers these issues holistically. *Technology* considers tools used to create, store, transfer, and apply knowledge. Organizations should choose carefully the tools and mechanisms pursuing initiatives to improve governance.

Governance of a risk management framework means understanding several issues. *First*, governance should account for the complexity of AI and its possible impacts, such that governing policies articulate and anticipate issues necessary to guide decision-making and action. *Second*, governance should require easily retrievable data, documents, and repositories so decision-makers can easily codify (and locate) the corpus of governing literature, and thereby promulgate new governance policy. *Third*, the organization should be able to evaluate, and mitigate risks of technology deployed within its ecosystem and should have transparent, documented, and supportable processes for understanding and scrutinizing how that technology behaves and what the organization expects of it.

The Governance Framework has three phases which correspond to the model lifecycle. In the *Development & Implementation phase*, the MRM committee and senior management will review the rationale for a new AI solution, assess model risk, and define conditions and outcomes for testing. In the *Validation phase*, teams gather and prepare data, build and train models, test to confirm business outcomes, and evaluate performance, replicability, and risk management prior to full-scale production. during this phase, MRM committees and senior management will verify that the conditions which were set out in the first line have been met and solution is cleared for deployment. Further, regular touchpoints are established with subject matter experts (SMEs), data scientists, and data engineers. After a model is deployed, an internal audit team supports as part of

the third line of defense during the *Use & Ongoing Monitoring phase*. During this phase, teams will monitor, maintain, assess, and re-train on an ongoing basis with considerations for unforeseeable AI risks and potential feedback loops.

We have helped implement this framework in the Regulatory space for state governments, federal agencies, as well as internationally. We use AI to accelerate the analysis of regulations using methodologies such as classification modeling, clustering, entity extraction, and network diagraming – all with the constant feedback of a regulatory expertise. These iterative checks allow us to determine the threshold of where AI is useful and when augmenting human judgement is best. As the organization matures in capacity and accountability structures are put in place, Deloitte's involvement may lessen or shift to new areas as the organization is empowered to maintain its own controls.