AI Risk Management Framework:

Guidance

For Corporate Legal Teams

DATE: February 2025

Contents

1. **ABOUT RAILS** 3
2. **INTRODUCTION: THE PROMISE OF AI** 4
3. **OBJECTIVE** 5
4. **PRINCIPLES** 6
5. **UNDERSTANDING AI RISKS** 7
6. **DEVELOPING YOUR FRAMEWORK** 11
7. **IMPLEMENTING YOUR FRAMEWORK** 24
8. **CONTINUOUS IMPROVEMENT** 27

APPENDIX 1 29

APPENDIX 2 35

1. ABOUT RAILS

Responsible AI in Legal Services (**RAILS**), hosted by the Duke Center on Law & Technology, is a collaborative network dedicated to advancing the responsible, ethical, and safe use of AI in legal services. RAILS unites stakeholders across law, technology, academia, and civil society to create actionable solutions that balance innovation with accountability.

The **RAILS AI Risk Management Framework** (**RMF**) Working Group exemplifies this mission, bringing together experts to develop a practical resource that empowers legal professionals in corporate legal departments to integrate AI responsibly while maintaining public trust.

RAILS at a Glance - RAILS drives its work through three core approaches -

(i) **Guidelines and Guardrails** - Developing voluntary standards, frameworks, and policy recommendations to responsibly embed AI across legal services, from client engagement to courtroom processes.

(ii) **Cross-Sector Collaboration** - Proactively fostering diverse, interdisciplinary, networks that encourage meaningful dialogue and partnerships.

(iii) **Education and Accountability** - Promoting transparent research, public awareness and principles that emphasize equity, fairness, and accountability in AI adoption

Much of RAILS' work takes place through dedicated working groups—volunteer networks of stakeholders collaborating to advance responsible AI practices. Working groups include experts across law, technology, policy, and academia.

RAILS extends its gratitude to all contributors for their thoughtful and generous contributions to this resource. As AI technologies evolve, we welcome additional participants to engage in refining this framework and contributing to ongoing discussions.

By working together, RAILS aims to foster a legal system where AI innovation strengthens—and never undermines—equity, accountability, and access for all.

To learn more and join upcoming RAILS and available working groups, sign up as a participant at https://rails.legal.

2. INTRODUCTION: THE PROMISE OF AI

2.1 Unlocking AI's Potential.

AI technologies present transformative opportunities across industries, enabling organizations to enhance productivity, streamline operations, and innovate in ways previously unimaginable. From improving decision-making through advanced data analysis to automating complex workflows, AI solutions can increase efficiency, reduce costs, and deliver new services that improve customer and stakeholder experiences.

2.2 Capturing Value Responsibly.

To fully capture these benefits, organizations must approach AI adoption with a strategy that balances ambition with accountability. By fostering trust, transparency, and adaptability, businesses can strengthen their competitive edge while demonstrating a commitment to ethical innovation.

2.3 The Role of Risk Management.

While the potential gains are significant, so too are the risks—ranging from bias and data security challenges to regulatory concerns and operational vulnerabilities. However, these risks need not be barriers to progress. A well-designed AI Risk Management Framework (**RMF**) ensures that risks are identified, assessed, and managed effectively, enabling organizations to pursue AI-driven growth with confidence. In the spirit of **Responsible AI in Legal Services**, this document's focus on risk ultimately aims to support that growth and confidence and to enhance trust.

2.4 A Path to Sustainable AI Success.

By embedding AI governance and safeguards into their operational frameworks, organizations can build resilience, foster stakeholder trust, and set a foundation for continuous improvement. The **RAILS AI Policy Guidance Framework** provides a roadmap to achieving this balance—empowering legal teams and decision-makers to capture AI's potential while mitigating potential harms.

3. OBJECTIVE

3.1 This document is intended to provide corporate legal teams with practical guidance on how to establish, implement, and maintain an AI risk management framework (**RMF**) for their business. The public release of ChatGPT in 2022 opened the doors of generative AI and triggered a spectacular growth in AI-focused products, regulatory attention, and discourse in the media. The speed, size, and evolving nature of that growth, being fueled again with the emergence of Agentic AI[1], have presented corporate legal teams with a significant task as they seek to work with their clients to develop a robust approach to risk management concerning AI. These teams face three challenges in particular:

(i) The rapid evolution of the marketplace for AI products and services, tool functionalities, and industry practices – including technical safeguards designed to reduce risks "at source."

(ii) The emergence of nascent regulatory regimes across jurisdictions and industries, where most of the regulatory landscape is evolving, but with almost no enforcement or jurisprudential background to inform regulatory understanding of the risks and benefits of AI.

(iii) Overlap in policy and frameworks between AI and other domains, such as privacy regulation.

3.2 The regulatory directives and guidance that have emerged already in different territories[2] all generally highlight the need for RMFs to be context-aware and appropriately modulated for the levels of risk involved. For that reason, this Guidance is designed more as a "how-to" guide than as a prescription for what a particular framework should look like. This approach is necessary because best practices are barely emergent at this stage. For example, the International Organization for Standardization (ISO) has developed standards for AI risk management,[3] which should be factored into corporate risk management planning –these standards continue to evolve and organizations will need time to familiarize themselves with them and determine an approach for implementation.

3.3 The primary audience for this Guidance is the legal professional, or team of professionals, with responsibility to their business client for the production of a corporate AI risk framework. In many organizations, responsibility for risk management is distributed across a number of executive and departmental roles, for which the legal team may play a contributory role. So, although this Guidance addresses AI risk management from a legal and regulatory perspective, it is designed for use by the team primarily responsible for the creation and maintenance of the RMF – and in this document we refer to the members of that audience as **users** of this Guidance.

[1] https://hbr.org/2024/12/what-is-agentic-ai-and-how-will-it-change-work
[2] e.g., the EU's AI Act of 2024 and the US National Institute of Standards & Technology AI RMF NIST AI 100-1.
[3] e.g., ISO/IEC 23894:2023 and ISO/IEC 42001.

4. PRINCIPLES

This Guidance has been developed using four principles:

4.1 Facilitation, not prescription.

The Guidance provides the users with steps and considerations for the production of an RMF suitable for their business, without mandating specific form and content. Illustrative examples and/or use cases will be given where helpful.

4.2 Agnosticism.

Corporate RMFs will vary by territory, industry, type of business, and even the use cases being deployed. This Guidance aims to provide generic direction, regardless of the enterprise specifics.

4.3 Pragmatism.

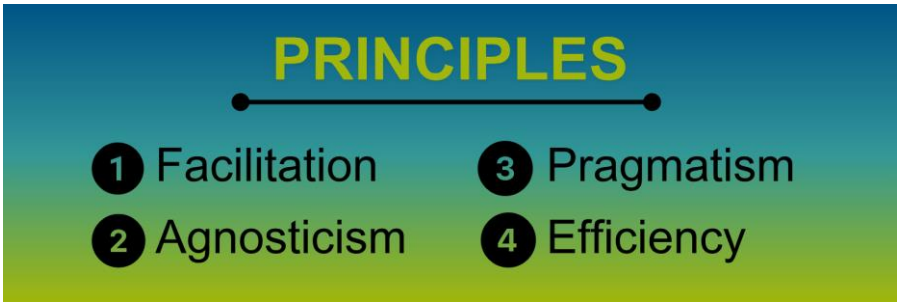
As a how-to guide, this document provides an outline of a roadmap that users can follow in order to build their own RMF. The framework itself is more than a policy: it is a dynamic, evolving set of principles, processes, roles and activities. This Guidance looks beyond the creation and implementation of the RMF, towards ongoing maintenance and improvement.

4.4 Efficiency.

AI technology and its use in business are, as mentioned earlier, developing and expanding at an extraordinary pace. So too are the efforts of lawmakers and regulators to keep up, and all these factors could lead to delay or even decision paralysis for organizations seeking to create their own RMFs. With that in mind, this Guidance seeks to do the following:

(i) Help users to prioritize AI risks appropriately for their industry, their planned technology deployments, and their primary operating territories.

(ii) Allow users to take advantage of existing RMFs and infrastructure within their organizations, where there is overlap or the ability to repurpose existing frameworks for the purposes of an AI framework.



5. UNDERSTANDING AI RISKS

5.1 The starting point for an RMF is to establish a clear understanding of the risks that need to be managed. The central question is ***what type of risks*** may be faced by the organization in building or deploying AI technology? That categorization then provides the foundation for the user to calibrate and prioritize those risks as they develop the RMF itself.

> ***What type of risks*** may be faced by the organization in building or deploying AI technology?

5.2 AI risks must be understood in the context of the legal and regulatory environments most applicable to the organization in question. Through all risk dimensions – categorization, calibration, and prioritization – an organization will need to ensure that its approach aligns with particular laws and regulations that apply to the organization's territorial reach and / or type of business.

5.3 The risks associated with the use of AI technology have been widely discussed in both academic papers and institutional resources. Some of these risks are largely the same as for other types of software-based systems - for example, where a program produces incorrect or unreliable output due to code or algorithmic errors. Some of the resulting risk depends on the magnitude of the harm that the errors can cause. Generative AI, however, raises the prospect of additional risks because of its ability to produce human-like digital content that may then be relied upon to influence the behavior of individuals, groups, or potentially societies as a whole. Also, the ability of generative AI to create content in the form of software code means that the technology can, in response to natural language instructions, be used to operate autonomous agents controlling machines and other software-powered systems.

5.4 Furthermore, the data processing systems, training data, and logic used by generative AI to produce content operate at a scale and complexity that obscures the technology's decision-making "choices." So, quite often even AI experts do not know exactly why an AI application answers a question or produces a picture in the way that it does. Compounding this problem is the fact that the data upon which AI models are trained may be flawed or out of date by the time the model is being used.

5.5 AI risks – in fact, risks in general – can be categorized according to both cause and effect, i.e.:

(i) The risk is that X happens (cause).

(ii) The risk is that X happens (cause), resulting in Y (effect or harm).

In order to help an organization deal with the risks, it is important for the risk management framework to understand both forms of categorization but not to conflate them.

5.6 From an ***effects*** (harms or financial impacts) perspective, AI risks can be broadly categorized as follows:

(i) **Human** – the risk that individuals or groups of people may be harmed, including social and environmental impacts.

(ii) **Operational** – the risk that the organization cannot function either fully or partially, for a duration that is material.

(iii) **Regulatory** – the risk that applicable laws or regulations may be breached.

5.7 The challenge with this approach lies in the fact that risks often – perhaps even usually – sit within multiple categories in practice.

Example 1 illustrates this:

Example 1

Your company rolls out an AI tool to assist with the screening of employment candidates. The tool unwittingly rejects candidates who are female, ethnically Afro-Caribbean, and over 50. The harms involved are:

(i) Human: rejected candidates are subject to unfair discrimination that impacts their ability to obtain employment.
(ii) Regulatory: this discrimination breaches various labor and privacy laws in the relevant territory.
(iii) Operational: significant cost and management time are incurred in investigating and remedying the situation, and meanwhile, the project that was designed to improve the speed and efficiency of the hiring process is hit by months of delay. The business also suffers a reputational impact.

5.8 A ***causal*** approach to risk categorization is frequently taken in discussion papers and regulatory instruments[4], perhaps because much of the focus has been on the fact that generative AI can "go wrong" in ways that are different from traditional software or even classical machine learning systems.

A good summary of causal risks can be drawn from the NIST AI Risk Management Framework[5]:

[4] See, for example, the OECD Initial Policy Considerations For Generative Artificial Intelligence, September 2023.
[5] See reference in footnote 2 above.

NIST Risk Category	Illustrative Concerns
Data quality and integrity	• Unknown input data quality • Unknown input data sufficiency • Risk of hallucinations • Reliance on unknown or third-party data sources
Transparency & use	• Difficulty understanding how the AI works • Difficulty interpreting output • Challenge mapping functionality to business need • Level of training required to implement AI effectively
Data privacy & security	• Challenge allocating responsibility between AI vendor and deployer for personal data safeguards • Protection of training and other proprietary data utilized by the AI tool • Risk of cyberattack using or penetrating an AI tool
Regulatory	• Burden of keeping pace with complex and rapidly changing law and regulations across multiple jurisdictions • Challenge allocating compliance liabilities between parties
Ethical	• Risk of bias in output • Risk of inappropriate use of AI in analysis or high-risk use cases • Social and environmental impact of the AI
Operational	• Risk of downtime • Unstable system performance • Resourcing involved for system maintenance • Use by the uneducated • Vendor concerns

5.9 Risk categorization is a matter of judgment, industry practice, and organizational approach rather than established taxonomy. In deciding what approach to take, this Guidance recommends the following steps:

5.9.1 Consider whether your organization maintains an existing RMF covering technology and / or data protection risks, which can provide at least the starting point for an approach to AI risk categorization.

5.9.2 Ensure that standards and operational guidelines relevant to your industry are reviewed and factored into your categorization methodology – for example, see the reference to ISO standards in section 3.2 above. It is highly likely that major industry

sectors such as finance, healthcare, law, and media will all produce their own regulatory guidance concerning AI risks.

5.9.3 Taking those foundational considerations into account, select an approach to risk categorization that can be applied intuitively within your organization. Think about other areas of corporate risk and how the organization approaches them, and in particular how it distinguishes between risk cause and risk effect as a matter of categorization.

5.10 The Appendix 1 sets out a series of risk identification questions designed around the NIST framework, which can be used to assist an organization in categorizing AI risks as they relate to that organization.

5.11 Also in the Appendix 2 is a chart providing examples of high and low risk use cases specific to the implementation of generative AI, to assist in determining if the organizations planned use case falls into an area already identified as having a higher or lower likelihood of risk.

5.12 Once the understanding of AI risks and how they should be categorized has been established, build the RMF.

6. DEVELOPING YOUR FRAMEWORK

6.1 Risk calibration

6.1.1 Risk calibration is the next step in building the RMF. This is focused on determining the ***intensity of risk*** for each category: what is the ***likelihood*** of the risk arising and what is the ***potential impact*** if it does?

6.1.2 Organizations may find it easier to determine the potential impact of AI risks than to assess risk likelihood, particularly given the relative novelty of generative AI tools at this stage. Subject to that caveat, there are a number of methodologies used for risk calibration that can be instructive for this process, including some that have been developed specifically for IT risks.

6.1.3 Many of these methodologies seek to calibrate risks using a matrix-style approach, plotting likelihood against impact and adopting a straightforward scale such as Low, Medium, and High across both dimensions.

This method is known as a qualitative risk assessment, and it will require a degree of expertise and experience in calibrating both likelihood and impact. Assessing impact can be done by drawing on relevant past experiences, both within the organization and across the applicable industry, and then working through scenarios as thought experiments. Organizations can even run specific tabletop exercises to assess risk likelihood and impact. Another risk modeling technique identifies and ranks AI risks and then assesses the severity and likelihood of harm for each as described in Bloomberg Law.

As mentioned earlier, assessment of likelihood may be more challenging and require heavier reliance on the guidance and technical parameters proposed by the AI tool or supplier. Researching published studies, incident reports, and legal claims may be helpful, although any statistical assessment of risk likelihood must pay close attention to the difference between relative and absolute statistics. An absolute statistic shows the actual difference in numbers, while a relative statistic shows the proportional difference compared to a baseline. A relative statistic may seem more significant by focusing on proportions while the absolute statistic reveals the actual- and potentially much smaller – real world impact.

Another reference point that will grow in importance is from the insurance industry, which is precisely in the business of calibrating risks. All indications are that insurance revenues, policies, and underwriting practices around AI risks will grow significantly and rapidly over the next few years[6].

Example 2 below provides an illustration of a risk calibration matrix. In this example, the organization is deploying an AI copilot to assist its lawyers in the review and redrafting of commercial contracts.

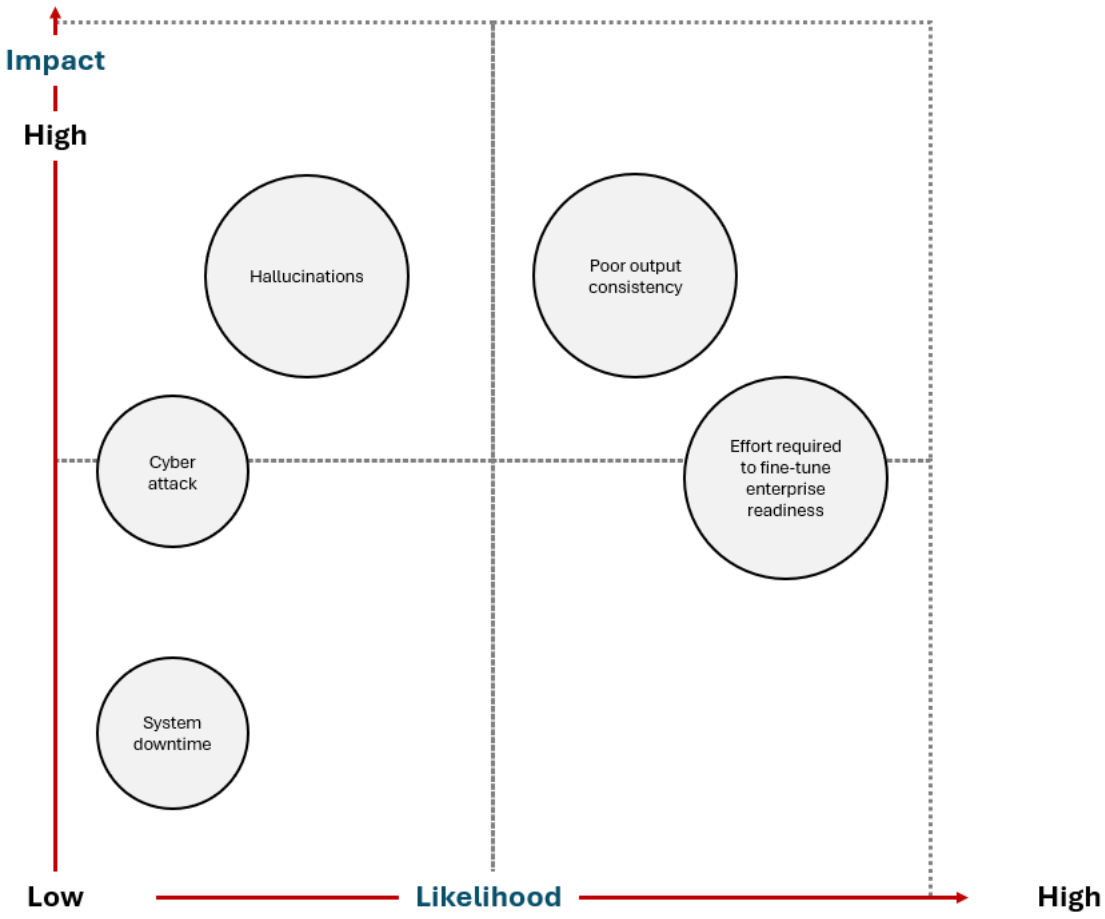
The risk analysis breaks down as follows.

System downtime:	The tool is intended to accelerate the first-pass contract review and production process, but if it fails to work, the contracting process itself should not be affected, only the levels of efficiency – Low Likelihood, Low Impact.

[6] See, for example, Deloitte Center for Financial Services report on AI risk insurance, 29 May 2024.

Cyber-attack:	The tool is cloud-based. The supplier and its cloud hosting provider deploy high-grade security features to prevent malicious access to end user systems. Low Likelihood, but if an attack manages to get through, the Impact could be Medium or High, depending on the nature of the incident.
Hallucinations:	The tool has an “anti-hallucination” feature, which minimizes Likelihood. However, hallucinations cannot be ruled out given the nature of LLMs, so while there is Low to Medium Likelihood the potential Impact could be High if a lawyer fails to spot the error.
Poor output consistency:	There is a relatively High Likelihood that the same prompts may generate different outputs, although the tool supplier is working on technical solutions to reduce inconsistency over time. Initially, however, there is both a relatively High Likelihood and High Impact with this risk, especially if the legal team does not perform proper quality checks of the output.
Effort required to fine-tune for enterprise readiness:	The organization will need to dedicate significant time and resources to the initial training and rollout of the tool, so this is a High Likelihood. Impact is Medium because the effort has been factored into the overall project and budgeting.

Example 2



6.1.4 The qualitative risk assessment method can be susceptible to criticism on the grounds of subjectivity and lack of precision. Another approach is often referred to as the ***quantitative*** risk assessment method, in which numerical values are assigned to likelihood and impact using probability distributions and historical data to determine expected scenarios[7]. For larger organizations with established risk management frameworks, complex quantitative assessments can be conducted using specific risk modeling tools and advanced techniques, such as Monte Carlo simulations[8]. However, the challenge for all types of organizations in adopting the quantitative approach is the current paucity of historical data around AI risks, as well as the effort and complexity involved in building the analysis.

6.1.5 In summary, at this stage in the evolution of AI technology deployment and regulation, a qualitative risk assessment is likely to be the most feasible for a majority of organizations, and a relatively simple matrix approach ought to be sufficient.

With such a matrix, the organization will have a ready reference guide identifying relevant AI risks, how likely they are to arise and what the impact could be in the event that they do arise.

6.2 Prioritization

6.2.1 If calibration is a way to standardize the risk assessment, prioritization is how to organize risks in order of importance or urgency for action. For that reason, prioritization is also a step along the framework path from theory about AI risks to practical execution.

6.2.2 Each organization must come to its own conclusions about how to prioritize risks. But in essence, the calculation comes down to two simple questions: ***how much could this risk hurt us, and how quickly***? Calibration will naturally inform this calculation, because the risks that carry the highest impact and likelihood will always be the top priority for action. But urgency is an additional factor in the prioritization process – i.e. how ***quickly*** a risk will hurt the business.

6.2.3 The purpose of prioritization is to help the organization work out how it should ***concentrate its resources and efforts*** according to the calibrated risks.

6.2.4 Here is a suggested method of establishing risk prioritization within the RMF:

Step 1: **Align risks with strategic objectives**. Identify how each categorized risk affects the organization's strategic goals. Risks that directly threaten high-priority objectives should generally be prioritized over those with less strategic impact.

Step 2: **Map strategic alignment to risk calibration**. Using the calibration matrix (to score risks by likelihood and impact), cross-reference calibration scores against the strategic assessment.

[7] NIST have produced a framework for risk assessment in IT systems (SP 800-30) built on this approach and widely used for compliance with U.S. federal standards and structured assessments of risk likelihood for security-conscious environments.
[8] A Monte Carlo simulation is a mathematical technique used to predict risk by running thousands of "what if" scenarios on what could go wrong and assess likelihood of each. As an example of the use of Monte Carlo simulations in the finance sector, see this article in the Journal of Accountancy: https://www.journalofaccountancy.com/issues/2017/nov/risk-assessment-using-monte-carlo-simulations.html?utm_source=chatgpt.com

Step 3: **Define risk appetite and tolerance**. Building on this strategic calibration, establish the level of risk the organization is willing to accept for each risk category.

Step 4: **Consider risk interdependencies and compound effects**. Analyze how risks might amplify or ameliorate each other. Interconnected risks may require joint prioritization to prevent cascading impacts or to leverage combined mitigation efforts.

Step 5: **Assess resource availability and constraints**. Conduct high-level review of the resources (budget, time, personnel) available to deal with risk events.

Step 6: **Conduct scenario analysis**. Explore risk event scenarios to understand how situations could evolve under different circumstances (e.g. industry-wide cyberattack, third-party IP claim against AI vendor, regulatory change). Plot available resources against prioritized risks in each scenario. Note that the speed at which different events might evolve becomes a critical factor in the impact analysis.

Step 7: **Stakeholder validation**. Gather input from leadership, operational teams, and other stakeholders to validate prioritization and scenario analysis, ensuring that the collective insights and concerns of the organization are factored in.

Step 8: **Finalize the calibrated, prioritized risk analysis**. The prioritized risk matrix and any relevant accompanying notes should be documented, following the organization's existing risk management methodology practices, as appropriate.

A prioritized risk matrix is often referred to as the **risk register**, which will constitute a foundational document for the RMF itself.

6.3 Establishing risk management governance

6.3.1 Framing AI risks – understanding, calibrating, and prioritizing them – can be done in parallel with the establishment of the organization's risk management infrastructure. Once again, an organization with an established RMF would look to that framework as the starting point to "plug in" new elements required for AI risks.

For context, this Guidance assumes that the organization needs to build an RMF from scratch, before going on to consider the steps necessary to update an existing RMF for AI risks.

Best practices in organizational risk management must adapt according to the size, complexity, and resourcing of the organization in question. For example, a large company will be able to mobilize a risk management sub-committee reporting to the board of directors, while a small business will typically have to deploy the senior leadership directly in risk management roles.

6.3.2 The first step in risk management governance is to ensure that risks are ultimately subjected to executive oversight. Overall responsibility for risk management should be assigned to the board of directors or a dedicated risk management committee with director-level leadership and reporting. A large organization may have a Chief Risk Officer or similar position in place, while for small and mid-sized companies, the responsibility could lie with the COO or CFO.

6.3.3 At or just below the executive level there should be a risk management committee with reach across the various functions and departments of the organization. The committee could include members such as the following (always allowing for limitations due to the size, structure, and resourcing of the company):

- The Chief Risk Officer (**CRO**), who would chair the committee and assume responsibility for overall risk oversight in the organization.
- The Chief Technology Officer (**CTO**), tasked with ensuring that AI deployments are technically sound and secure.
- Potentially, an AI Risk Officer focused on monitoring and mitigating risks specific to AI systems. This person should have expertise in AI ethics, data governance, and risk management.
- The Chief Legal Officer (**CLO**), ensuring compliance with legal frameworks, such as data privacy laws (e.g., GDPR) and AI regulations.
- A Head of Compliance, who monitors adherence to internal and external policies and regulations.
- The Head of Internal Audit, who conducts internal audits on AI systems and provides independent reports to the committee.

6.3.4 Expanding or supporting the work of the risk management committee, there may be additional roles to deploy:

- A specialist AI Ethics and Compliance lead, who would specifically focus on the development of ethical AI use guidelines to ensure fairness, transparency, and respect for personal and data privacy rights.
- AI model owners within relevant business units and functions, who are responsible for the AI models deployed within their respective corporate areas (e.g., product / service lines, HR, marketing, finance, etc.). They would typically be tasked with monitoring AI models for accuracy, fairness, and compliance, and to ensure the explainability of models to non-technical stakeholders.

- The Data Protection Officer (**DPO**), who would oversee personal data governance and privacy policies concerning AI systems and conduct privacy impact assessments for AI-driven projects.
- The AI development team, working with AI model owners, the CRO / AI Risk Officer and CTO to develop AI models that are robust, secure, and transparent.
- Business unit risk champions, representing different departments to promote AI risk awareness and acting as liaisons between business units and the AI Risk Officer / CRO.

6.3.5 If the organization has a mature RMF already in place, a specialized committee can be set up with specific focus on AI risks. This committee would bring in expertise across all relevant areas, including AI ethics, data security and privacy, and legal and regulatory compliance.

An “AI accountability framework” could be introduced as an extension to the existing framework, identifying the roles and responsibilities for AI risk ownership (including the appointment of an AI Risk Officer, for example).

6.3.6 As with any organization, accountability and direct / indirect reporting structures should be carefully plotted. On a practical level, the responsibility matrix and reporting structure need to find a balance between organizational logic and the actual demands of execution.

The risk management ***policy*** – which is covered in the next section – should have an identified owner within the governance team.

6.4 Creating an AI risk management policy

6.4.1 The risk management policy is a document that captures all relevant features of an organization’s RMF, from the governance structure to how the organization identifies, prioritizes, and responds to risks. It is best practice for such a policy to start with an introductory explanation of its purpose and scope. The policy would then typically cover the following elements.

6.4.2 Governance structure

The risk management governance structure worked out for the organization is detailed with reference to role, responsibility and, where appropriate, direct and indirect reporting lines. It is good practice to include individual contact details, at least for team members included in the incident response sub-team (see section 6.4.7 below). As mentioned above, the policy document should have an identified owner from the governance team – often, this will be the CRO, if there is one, or otherwise the COO or CFO.

6.4.3 Risk identification

The risk management team must document the main AI-related risks faced by the organization, which requires an assessment of how AI is being used or may in the

future be used in that organization. Information about AI usage should cover the following:

- Where AI is or might be deployed (e.g., in decision-making, process automation, customer service, data gathering, etc.).
- The types of AI models or tools to be used.
- The scope of AI's impact or degree of touch on business operations and its potential for harm and the probability and potential size of that harm.

6.4.4 Calibration and prioritization

Using the risk calibration matrix and prioritization methodology established by the organization, the organization should grade the identified risks accordingly. This grading classification should be applied both to AI tools deployed for internal use and the development and commercialization of AI products and solutions being built or proposed by the organization for supply to customers. For both "internal" and "external" use tools, the management team should consider associated supply chain risks in which AI capabilities are being bundled into broader solutions deployed by the organization.

This information feeds into the risk register, which will either form part of, or be linked to, the policy.

6.4.5 Risk monitoring

(i) The organization should establish Key Risk Indicators (**KRIs**) to provide early warning signals for potential risk exposure. KRIs should be measurable and relevant to the organization's risks according to the calibration and prioritization methodology used.

Here is an illustration:

Example 3

KRI	Risk
Significant drop in model accuracy (e.g., accuracy falls below a predefined threshold, like 90%).	Poor decision-making due to inaccurate AI predictions, leading to operational errors or customer dissatisfaction.
Detection of biased outcomes (e.g., disparities in prediction outcomes for different demographic groups).	Legal, ethical, and reputational damage due to biased or discriminatory AI results.
Increasing percentage of incomplete or poor-quality data fed into AI systems.	Poor data-in / data-out, leading to flawed decision-making, operational inefficiencies, or legal risks.
Number of AI processes not aligned with existing regulatory frameworks (e.g., GDPR, CCPA, AI Act).	Fines and penalties from non-compliance with data privacy, ethics, or AI-related regulations.
Increase in AI-related security vulnerabilities or cyberattacks, such as model hacking or adversarial attacks.	Data breaches, compromised systems, or malicious misuse of AI models for fraud or attacks.

KRI	Risk
Percentage of AI decisions that are black-box (i.e., not explainable).	Difficulty in justifying or understanding AI decisions, leading to decreased stakeholder trust and potential legal challenges.
Percentage of business-critical decisions made exclusively by AI without human intervention.	Over-reliance on AI without human checks could lead to systemic failures if AI models are flawed or behave unexpectedly.

(ii) A protocol should be established to implement regular monitoring and tracking of risks and KRI performance, ideally using tools such as risk dashboards to allow for real-time visualization. The protocol should be designed to ensure that there are regular reviews to identify new risks, keeping track of evolving AI technology and usage patterns. Use of tabletop exercises (scenario / gaming) should be encouraged wherever new risks are identified.

In designing the monitoring protocol, the organization may need to consider the following factors:

(a) Opportunities to take advantage of existing risk management structures, i.e., to align AI risk monitoring with established processes such as internal audits, compliance checks, and operational risk assessments.

(b) Because AI systems rely on vast amounts of data, the organization must monitor the quality, completeness, and timeliness of the data closely. This monitoring may require mechanisms to be implemented to track data quality in close to real-time.

(c) The KRI monitoring system should also extend to regulatory observation, tracking changes to applicable laws and regulatory guidance for compliance purposes.

(d) Human oversight of AI output is essential, so the organization should ensure that human quality control mechanisms are built into use case workflows, with clear accountability for AI decision-making processes and escalation paths when AI systems behave poorly or unexpectedly.

6.4.6 Reporting

(i) Very broadly, the aim of the risk reporting system will be to identify AI risk exposure, track and support response / mitigation efforts, and provide actionable insights into AI performance so that the risks can be addressed proactively.

(ii) The reporting system should be built upon consistent and standardized parameters, to ensure that AI risk data can be understood with clarity and comparability. The risk categories should be clear and drawn from the risks matrix already developed, and the metrics used should be consistently applied (e.g. as to accuracy rates, model drift etc.). Similarly, there should be discipline around the reporting frequency and formats (with reports being as concise and focused as possible), and where appropriate data visualization techniques should be used.

(iii) AI risk reporting should not be siloed from broader organizational risk reporting processes, so the AI risk reporting system would ideally integrate at some level with existing enterprise risk management (ERM) systems, regulatory reporting

frameworks, and compliance dashboards. Data-sharing protocols and APIs may be required to ensure seamless integration and automated data flow between systems.

(iv) Where possible, real-time monitoring tools should be deployed for dynamic risk assessment, with automated data collection from key sources such as performance metrics from the AI models (accuracy, precision, etc.), external threat intelligence (e.g. for cybersecurity risks) and internal audit findings. AI-powered analytics can even be deployed to detect patterns or anomalies in AI performance.

(v) Although reporting standardization and simplicity are important aims, different stakeholders may require varying levels of detail or focus, so the reporting system should allow some flexibility for customization - for example, by giving executives only the high-level summaries but giving risk managers and expert contributors more detailed and technical reports.

(vi) One technical consideration is whether it may be possible, at least for certain AI products, to integrate explainable AI tools into the system so that AI decision-making processes can be interpreted. This integration could allow some risk reports to include narrative explanations of AI behavior, particularly when risks such as bias or model drift are detected. For example, Explainable AI (XAI) is a set of processes or methods that enable humans to understand the results of the AI. NIST has identified four principles of XAI found here and described at a high level as follows:

Explanation	How conclusion was reached by the AI
Meaningful	Explanation of the conclusion should be understandable
Explanation Accuracy	AI explanation aligns with underlying process/methods
Knowledge Limits	AI should be able to identify when there is insufficient information or lack of confidence in output

(vii) Another consideration when designing the reporting system is auditability. Ideally, the system will allow the construction of a complete audit trail for incidents being captured – including when the log was generated, who accessed or modified it, and the response actions taken.

(viii) Finally, the reporting system should provide clear escalation pathways, with reports to senior leadership automatically triggered by high-risk incidents.

6.4.7 Incident response

The incident response system will have a number of components:

(i) An **AI incident response team**. This team will be drawn from the AI risk management committee (see sections 6.3.2 to 6.3.4 above), potentially supported by staff with relevant technical or operational expertise (such as data scientists, risk managers, etc.).

(ii) An **AI incident response playbook**. This playbook can be built using the risk framework enshrined in the risk register, with scenarios developed from the various analyses used to produce the risk register. The immediate responses must focus on situation containment to minimize further impact from a given risk event. For instance, a designated response may include mechanisms to temporarily suspend

the AI's decision-making capabilities, revert to a previous model version, or insert human intervention at appropriate steps in the workflow.

The playbook guides the response team through the range of AI incidents based on the risk register's calibrations and priorities, identifying whom to contact, what containment actions should be taken, and the communications protocols to follow.

(iii) A **communications protocol**. This protocol should address both internal and external communications:

(a) Clear internal comms channels will inform relevant stakeholders on incident status, which, in appropriate circumstances, may involve real-time updates.

(b) Communications with external stakeholders and agencies must also be controlled – customers, regulators, suppliers, and the media should all be taken into account. Transparency has to be balanced against the risk of creating noise and heat while information about an incident is still being collected and organized. Reporting practices mandated or recommended by regulatory authorities should be given paramount consideration.

(c) If the organization maintains a cross-risk crisis management team, AI incident-related comms should be coordinated accordingly.

6.4.8 Incident review and analysis

The RMF must include arrangements for incident review and analysis. These arrangements would comprise a number of elements:

(i) **Forensic capability**: the ability to audit the AI models in use to identify foundational contributors to the incident.

(ii) **Bias and fairness assessment**: this feature will be particularly important for AI decision-making systems that are involved in an incident.

(iii) **Data tracing**: the incident review team should be able to trace back data inputs and transformations that may be implicated.

(iv) **A defined incident review team**, whose composition may vary depending on the nature of the incident. For example, where personal data or decision-making that affects people may be involved in the incident, the organization's responsible officer for privacy matters will be a key member of the team. But that officer might not be required where the incident involves breach of customer service levels under

commercial contracts, while both the Chief Legal Officer and a senior representative of the service delivery team may need to be brought in.

Incident reviews need to be thorough and tasked with assessing root causes, the particular sequence of the events, the effectiveness of the response, the adequacy of the reporting system, and any gaps in detection or containment.

Final reports should be produced with input from all key stakeholders to ensure a diverse range of insights, and conclusions will then feed into the continuous framework improvement loop.

6.5 Risk mitigation

6.5.1 Risk mitigation is about taking steps to ensure that the impact of AI risks is either avoided or minimized. So in the broadest sense, all of the steps described in this Guidance for the creation of an RMF serve this aim of risk mitigation. But there are a number of measures that can be taken specifically to mitigate AI risks through proactive planning and preparation.

6.5.2 Automation and analytics

Automated detection and response mechanisms have been mentioned already[9], but the speed and complexity of AI incidents may require the use of these mechanisms. For example, automated monitoring can track model performance, data quality, and unusual system behaviors to trigger alerts before a risk escalates into a full-blown incident.

Using sophisticated risk analytics tools can also allow for the continuous assessment of the risk landscape, with AI engines helping to predict incidents based on historical patterns and model behaviors.

6.5.3 Technical measures

(i) A standard mitigation in IT security, the use of back-up systems will be an obvious measure to take in the case of suitable AI risks. Alternative systems and fallback protocols should be designed to kick in as soon as an AI system fails critically.

(ii) A more AI-specific measure is to ensure (directly or via the provider) that AI models are regularly retrained with fresh data to reduce the risk of model drift and maintain relevance.

6.5.4 Data Preprocessing Techniques

Data preprocessing improves the quality of the data and involves techniques such as data cleaning to identify and correct errors or inconsistencies to make it more suitable for analysis.

6.5.5 Simulation drills

Where feasible, AI incidents should regularly be simulated to test the efficacy of the response playbook.

[9] See section 6.4.6.

6.5.6 Data privacy and security

Again, overlapping with an organization's IT security arrangements, AI incidents should be factored into the planning and design of data governance and processing structures. For example, data anonymization, masking, or encryption can be deployed to protect sensitive personal information that may be processed by an AI system either in training or live production. Other examples include the use of tight access controls and the use of PETs (Privacy Enhancing Technologies).

In this context, liaison with technology partners will be an important element of the risk mitigation strategy – particularly around vulnerability probing, model security and resilience testing.

6.5.7 Legal and regulatory compliance

It is not possible with AI systems to entirely rule out user misuse, technical malfunction, or cybersecurity attack. Nevertheless, an organization's exposure to the ***consequences*** of these risks – especially in terms of legal liability and reputational impact - can be mitigated in practice by taking a rigorous approach to legal and regulatory compliance. Specifically:

(a) The organization should stay informed of evolving AI regulations and take proactive steps to ensure compliance – across AI-specific legislation such as the EU AI Act, relevant sector regulation (healthcare, financial services etc.) and data protection / privacy rules. Outside counsel may need to be engaged to assist with this analysis.

(b) Legal teams should be engaged early in an AI project lifecycle to look at the contracts, intellectual property rights, liabilities and commercial risks associated with the tool's procurement, development, and deployment.

(c) Where possible, systems should have accountability built in – especially in relation to decision-making AI tools.

6.5.8 Ethical AI and Corporate Social Values (CSV)

The approach to regulatory compliance extends to the organization's observance of ethical and social standards. Three principles can be followed here:

(a) **Alignment with values**. AI initiatives can be demonstrably aligned with the organization's ethical standards and CSV goals.

(b) **Inclusive design**. Stakeholders in the AI tool design process should represent divergent backgrounds, including under-represented groups.

(c) **"AI for good."** The use of AI by the organization can be promoted for positive societal impacts and the minimization of harms, supporting aspirations of sustainability, fairness and social benefit.

6.5.9 Training and awareness

Training and awareness are key to risk mitigation; see section 7 below.

6.5.10 Industry engagement

Engaging with experts, regulators, and industry peers will ensure that the organization keeps up with best practices and contributes to the shared pool of experience that will ultimately drive better, more robust AI systems and risk controls. Specifically:

(a) **Engage with regulators**. Remain actively engaged with regulatory bodies to establish compliance with the evolving regulatory and legal landscape and to inform the thinking of those bodies.

(b) **Collaborate with industry**. Collaboration can extend to industry peers, academia, and external experts with a view to sharing best practices, surfacing new technologies and evolving AI risks.

(c) **Adopt industry standards**. Implement and adhere to industry guidelines for AI risk management and safe, ethical AI development[10].

[10] E.g. IEEE, ISO and NIST frameworks.

7. IMPLEMENTING YOUR FRAMEWORK

7.1 Building a risk management culture

7.1.1 With organizational RMFs, there is a danger over time that the systems and processes deployed become exercises in checklist ticking. As with other significant risks that reach across the organization's functions, technical domains and departments, the aim should be to instill a culture that combines awareness of AI risks with a determination to tackle those risks proactively, effectively and with sensitivity towards the interests of both the organization itself and the people who can be affected by them.

7.1.2 Building such a culture is not a one-time event, nor is there a single mechanism for making it happen. There are many different approaches, but the main pathways can be grouped under these headings:

- Leadership
- Training
- Integrated communications
- Practical reinforcement

7.2 Leadership

Leadership is required at every stage in the creation and implementation of an AI RMF:

7.2.1 The RMF governance structure must ultimately be overseen at the executive level.

7.2.2 Cross-functional stakeholder involvement should be directed by the engagement of senior leaders for each function, and this engagement must be visible to the organization.

7.2.3 Key personnel can be appointed across departments as "risk champions", specifically tasked with leading awareness campaigns and adherence to policy.

7.2.4 Behaviors and communications around AI risks and risk management should be consistent across the leadership team, while being adjusted as necessary for the relevant audience. Essentially, every member of the organization needs to feel that it is "my job" to support the RMF in the ways that have been tailored and communicated to each person's part of the organization.

7.2.5 Strategic decision-making by executives should integrate AI risk considerations wherever relevant.

7.2.6 Leadership must encourage transparency, building an environment where employees feel comfortable reporting potential risks and concerns without fear of retribution. This approach will be facilitated by the implementation of straightforward risk reporting channels and mechanisms.

7.2.7 Positive risk management behaviors should be clearly incentivized, i.e., by:

- Linking performance metrics to AI risk awareness / risk management participation, especially for teams that are working directly with the technology or associated data.

- Rewarding teams or individuals who innovate within the bounds of responsible AI practices and who demonstrate a strong commitment to AI risk mitigation.

- Ensuring that all staff understand their roles within the RMF and are held accountable accordingly.

7.3 Training

7.3.1 When ready for launch, the RMF and policy documentation should be announced to the organization. The announcement will be a prelude to formal training, and it sets the scene for the leadership engagement and prioritization referred to in section 7.2 above.

7.3.2 Training itself should be multi-tiered to ensure that the right level of training is being given to the right groups in the organization. For example:

(i) **Basic AI literacy**. As a starting point for understanding the basics of AI, its risks, and socio-ethical considerations, AI literacy programs would be run for the whole organization. This literacy training would include training non-technical teams to recognize the implications of AI-powered processes in their respective domains.

(ii) **Targeted risk management training**. More advanced training would be designed to equip AI practitioners, risk officers, and compliance teams with a detailed understanding of the AI risk management policy, regulatory requirements, and risk management and mitigation techniques.

(iii) **Reporting and incident response training**. Specific training is given on the risk / incident reporting system, for all members of the organization, with concentrated training given to the members of the AI incident response team - including those responsible for executive oversight - to embed knowledge of the response methodology and how it fits into the overall RMF.

(iv) **Scenario-based learning**. Scenario-based learning (or “tabletop exercise”) is a method that can feature in some or even all the training programs, using simulations and case studies that position AI risk management in the real world and provide very practical situations to illustrate how to detect, address, avoid and mitigate AI risks. The scenarios themselves would be adjusted for relatability according to the training audience.

7.3.3 It is not enough for training simply to be made available to the organization. Engagement with the training should be mandatory where necessary, strongly encouraged in general, and tracked as regards adoption and compliance levels. Staff managers should be incentivized to ensure full adoption of the relevant programs, but also to use staff meetings as an opportunity to validate that the training has been engaged with meaningfully.

Best practice also dictates that the organization implement a meaningful feedback loop, so that staff reaction to training can be factored into the training refresh and improvement initiatives.

7.4 Integrated communications

7.4.1 An integrated communications system has certain characteristics. Communication about AI risks and the RMF:

- flows freely and consistently through the organization;

- is tailored to audiences for maximum relevance and effectiveness (e.g., communicated via OGC/engagement letters when communicating with law firms or board policies for internal legal departments and clients);
- incorporates feedback upwards through the organization so as to inform the evolution of the RMF; and
- encourages the inflow of information from external sources and the wider industry.

Open dialogue is fostered, with staff being encouraged to discuss AI risks and socio-ethical concerns in both formal and informal settings – for example, through workshops, town halls, and other suitable forums.

7.4.2 One challenge for larger and more diffuse organizations is the difficulty of maintaining communications consistency and clarity across departments and locations. This is not unique to AI risks, nor even to corporate risk management in general. The communications strategy can be designed specifically to meet this challenge, and where an organization has had good success with such strategy in other areas, it should capitalize on the methods that delivered that success.

7.4.3 Regarding information from external sources, as mentioned in section 6.5.10, it is important to collaborate with industry peers, academia, and regulatory bodies to stay updated on AI risk trends and emerging best practices.

7.5 Practical reinforcement

Practical reinforcement refers to all the observable behaviors in the organization that support the communications and training around AI risk management. In a sense, this is about bringing leadership to life and adhering to the risk management policy, and encouraging their teams to do the same. There are many examples of how this can be done:

(i) **Encouraging experimentation within guardrails**. The aim is to promote innovation and familiarization with AI, but also respect for the strict protocols and controls that define acceptable activity and risk boundaries.

(ii) **Risk-aware workshops**. Technical workshops on the design and development of AI-powered solutions can incorporate specific agenda items focused on AI risks (for example, ethical decision-making), leading to informed trade-offs between innovation and risk mitigation.

(iii) **Celebrating success**. AI success stories can be publicized, and in doing so reference can be made to the risk-sensitive approach taken by the team in question.

(iv) **Incentivization**. Incentivization is covered in section 7.2.7 above and is probably the most direct management tool available to reinforce messaging around AI risk management. Positive incentives are typically more effective - if the incentives are meaningful and valued by members of the organization - than those that are negative or punitive. “Catching someone doing something good” can be effective and handled in every department of the organization.

8. CONTINUOUS IMPROVEMENT

8.1 RMFs need to be reviewed and updated regularly. Given the speed at which AI technology and use cases are developing – and indeed the pace of the chasing regulatory regimes – there is a particularly compelling requirement for AI RMFs to evolve dynamically. There is both a reactive and proactive element at work here:

- ***Reacting*** to technological, regulatory and industry developments to ensure that the RMF remains relevant and fit for purpose.
- ***Proactively*** seeking ways in which to improve the RMF in functional terms.

Together, we can refer to being reactive and proactive as creating a continuous improvement program.

8.2 This Guidance has already referred to some features of a continuous improvement program, such as the internal audit process. In this section, we describe each element of the program and how each fits together with the whole.

8.3 Incident review

The incident review process described in section 6.4.8 should produce detailed reports for specific risk incidents, but a good incident review process will also allow the review team to look at patterns and trends. A mature RMF will allow for formal reporting at executive level, covering incident patterns and overall conclusions.

8.4 Audits

Internal and external audit processes can provide significant input to the continuous improvement program, focusing outwards on compliance with applicable regulatory standards and inwards at the levels of industry best practice displayed by the RMF. Audits can assess the conformity of the organization's processes with their own framework requirements – for example, the extent to which incident responses are aligned in practice with the risk calibration methodology.

The findings from both internal and external audits should feed into the framework update process on at least an annual basis, prompting improvements in the risk policy, internal controls, monitoring and reporting and even the training programs.

8.5 Reporting systems

Weaknesses in the AI risk reporting systems should be picked up by incident reviews and audits. But there should also be an ongoing focus on the quality of reporting by the responsible stakeholders – those tasked with generating reports and using the reporting systems, and those with accountability for decision-making based on the reports. Indeed, the purpose of the reporting systems is to identify AI risks and support response efforts, so the reports should be generating data about those risks that will directly inform updates to the RMF.

8.6 Benchmarking

Benchmarking an organization's RMF can be a targeted outcome from designated audit processes, particularly if external auditors are brought in with a perspective on industry practices. Benchmarking can also be performed using dedicated benchmark

service providers, although this might prove to be challenging or premature at this stage in the evolution of AI risk management practices.

8.7 Informal feedback channels

8.7.1 There is an informal route to benchmarking, which is particularly relevant for organizations at this early stage in the development of corporate AI risk management: information sharing among peers and across industries. Discussions, work groups, conference presentations, or roundtables, and participation in standards initiatives can all provide an insight into AI risk management challenges and successes.

8.7.2 Feedback channels inside an organization will be as important as those available outside. Section 7.4 of this Guidance discusses the importance of integrated communication and the role of bi-directional information flow between management and staff. A successful RMF will foster an open, risk-aware culture in which constructive feedback is freely given on everything from AI tool deployment to process and design issues.

It will be the job of team heads, managers, and the nominated leadership of the RMF to ensure that this free flow of feedback can be harnessed, with relevant information being organized and fed into the formal framework update process.

8.8 Framework updates

The risk management policy should specify a formal process by which the framework will be reviewed and updated, at least on an annual basis. The policy owner (see section 6.4.2) would nominally be responsible for the implementation of updates, and for organizing the decision-making process around those updates.

Ultimately, the RMF should be seen as a dynamic and evolving system, articulated within the risk management policy and driven by a sense of shared ownership across the organization.

APPENDIX 1

Risk Identification Questions by NIST Risk Category:

NIST Risk Category: Data Quality & Integrity

Concerns:

- **Unknown input data quality**
- **Unknown input data sufficiency**
- **Risk of hallucinations**
- **Reliance on unknown or third-party data sources**

Questions:

1. What are the sources of training data for the Generative AI (GAI)tool? (Understanding the data sources helps assess the quality and potential for hallucinations in the AI output.)
    a. Is source data defined and limited? Can it or should it be preprocessed to avoid bias or improve accuracy?
    b. Where is it from?
    c. How recent is the data and how frequently refreshed?
2. Has the GAI vendor made available documentation that confirms data sources have been verified/vetted and within what timeframes?
3. How does the AI tool integrate with existing systems and workflows?
    a. Are input data sources appropriate?
    b. Is output validated before use by downstream tools or organizations?
    c. Have integration points been tested for security and data validation?
4. What are the cycles for data quality and security checks?
5. What are the vendor's policies on intellectual property and who owns the source data, training data, inputs or prompts, and AI-generated outputs?
6. What are the Vendor's data use limitations, model access rights, and input/output ownership or licensing terms?
7. Is the AI vendor license with a third party and what rights and protections exist and are passed through?
8. Will the vendor guarantee or indemnity for the 3rd party's performance?

NIST Risk Category: Transparency and Use

Concerns:

- **Difficulty understanding how AI works**
- **Difficulty interpreting output**
- **Challenge mapping functionality to business need**
- **Level of Training required to implement AI effectively**

Questions:

1. How transparent and explainable are the AI system's decisions? (If the AI tool decision basis cannot be clearly explained then you must understand this is a risk that must be absorbed to use this technology.
2. What data mapping, bias or other transparency tools does the AI vendor use in its operation, if any?
3. What performance metrics and benchmarks does the vendor provide to assess the AI tool's effectiveness and reliability based on measurable criteria?
4. Has your team defined KPIs for each use case with a cadence for testing output, measuring performance, and providing feedback?
5. Is the business need fully defined, and does it align with design of AI application?
6. Is there an expected ROI for the project and how will it be measured?
7. Does the team implementing have the requisite training and skills to manage the AI and understand its environment, inputs, and outputs?

NIST Risk Category: Data Privacy & Security and Technology

Concerns:

- **Challenge allocating responsibility between AI vendor and deployer for personal data safeguards.**
- **Protection of training and other proprietary data utilized by the AI tool**
- **Risk of cyberattacks using or penetrating an AI tool**

Questions:

1. What privacy implications are raised with the tool or data used (i.e.: includes PII, overseas storage access, etc.)?
    a. Does the data include PII?
    b. Are there overseas access or storage concerns?
    c. Are there permission controls for internal access and safeguards for external access?
    d. Are data inputs and well as outputs protected?
2. What coordination is needed with existing privacy policies and processes and do those need to be updated to cover this AI use?
3. Does the vendor have robust measures in place to protect sensitive information including industry specific concerns (healthcare, financial services, governmental security, legal)?
4. To what cybersecurity and other data protection practices, measures, certifications, or standards (i.e.: SOC II, ISO, etc.) does the AI Vendor comply or adhere and do they meet our organizations requirements? (Check for industry certifications that demonstrate the tool's reliability and compliance.)

NIST Risk Category: Regulatory

Concerns:

- **Burden of keeping pace with complex and rapidly changing laws and regulations across multiple jurisdictions**
- **Challenging allocating compliance liabilities between parties**

Questions:

1. Is our outside counsel keeping us informed of relevant regulatory changes impacting the use of AI in general and for our industry specifically?
2. What measures are our outside counsel using to assure its own compliance with AI regulation?
3. What steps are our in-house team members taking to research and stay current on the evolving legal landscape and to educate their business clients?
4. Do your output validations check for violations of copyright/patent?
5. Is output screened for violent/hateful/obscene/offensive content before usage?
6. How does the AI tool and Vendor ensure compliance with relevant laws and regulations? This includes privacy & data protection laws and industry-specific or environmental regulations.
7. What Reps/Warranties are provided by AI Vendor? What are the liability limits?
8. What indemnities are provided by AI Vendor and are they financially viable if invoked?

NIST Risk Category: Ethical and Environmental

Concerns:

a. **Risk of bias in output**
b. **Risk of inappropriate use of the AI in analysis or for high-risk circumstances**
c. **Social and environmental impact of the AI**

Questions:

1. What creation, screening, and testing measures are in place to detect and mitigate biases in the AI system? (Bias detection and mitigation are critical for ensuring fairness and avoiding discriminatory outcomes.)
    a. Are the algorithms built to detect bias?
    b. Is the training data diverse?
    c. Is the AI training/validation team diverse?
2. Are there regular and on-going reviews of the AI output for issues and concerns?
3. What are the vendor's policies on AI ethics and responsible AI use and sustainable development? (Ensure the vendor aligns with your organization's values and ethical standards.)
4. Are we assessing our AI business needs to assure the benefits are proportional and necessary to all risks, impacts, and mitigations required for a proposed use case, particularly for high-risk use cases?

NIST Risk Category: Operational

Concerns:

a. **Risk of Downtime**
b. **Unstable system performance**
c. **Resourcing involved for system maintenance**
d. **Use by the uneducated**
e. **Vendor concerns**

Questions:

1. How does the vendor handle updates and maintenance of the AI tool? Understand the frequency, notice, and process for updates to ensure the tool remains available, effective, and secure.
2. What are the vendor's policies on incident response and handling AI-related issues?
    a. How are incidents classified and severity levels differentiated?
    b. What mitigation steps are available for implementation and how is their use determined?
    c. What are the recovery steps and rollback options for restoration in the event of an incident and how are they selected for implementation?
    d. What notice and communication protocols will be followed?
3. What user support and training does the vendor provide? (Assess the level of support and training available to help your team effectively use the AI tool.)
4. How long is the vendor in business and what is their financial stability/# of customers, growth rate, funding, history/any publicly known issues with AI?
5. For what industries is the AI/application designed?
    a. How long has it been deployed?
    b. Are there references?

APPENDIX 2

Example High Risk and Low Risk Use Cases for Generative AI applications:

While not exhaustive, the list below provides use cases that illustrate some of the unique concerns around generative AI. Declaring a use case as a High or Low Risk is not definitive however as any of the lower risk examples can have high risk implications dependent upon the circumstances or industry of a particular organization.

Higher Risk Use Cases

- **Biometrics**: AI systems used for biometric identification and categorization, such as facial recognition, can pose significant privacy and security risks.
- **Critical Infrastructure**: AI applications in critical infrastructure, like energy grids or water supply systems, can have severe consequences if they fail or are compromised.
- **Education and Vocational Training**: AI systems used in educational settings for grading or admissions can perpetuate biases and affect students' futures.
- **Employment and Workers Management**: AI tools used for hiring, performance evaluation, or workforce management can lead to discriminatory practices.
- **Access to Essential Services**: AI systems that determine access to essential services like healthcare, insurance, or financial services can have life-altering impacts.
- **Law Enforcement**: AI applications in law enforcement, such as predictive policing or surveillance, can lead to privacy violations and biased outcomes.
- **Migration and Border Control**: AI systems used in migration and border control can affect individuals' rights and freedoms.
- **Justice**: AI tools used in the judicial system, such as for sentencing or parole decisions, can have profound implications for fairness and justice.
- **Medical Algorithms**: AI systems used in healthcare for diagnosis or treatment recommendations can exhibit biases that impact disadvantaged populations.
- **AI-enabled Recruiting Tools**: These tools can perpetuate biases in hiring processes, leading to unfair treatment of candidates.
- **AI Facial Recognition**: This technology has led to wrongful arrests and privacy violations.
- **Agentic AI –** This technology uses AI to adjust processes independently and proactively based on real-time contextual learning. This autonomous decision making could increase bias and yield unexpected and/or undesirable behaviors or outcomes.

Lower Risk Use Cases

- **Legal Writing**: Generative AI can assist in drafting legal documents such as position papers; regulatory impact analyses, and memoranda of understand. This helps legal professionals gather insights from large sets of data and focus on the information that matters most, enabling them to be more efficient and strategic.
- **Legal Research**: AI tools can conduct legal research, generating summaries and identifying relevant case law, which saves time and reduces costs.
- **Litigation Support**: Generative AI can summarize depositions, suggest interrogatories, and conduct privilege reviews. Courts have mandated human validation for use in identifying relevant case law.
- **Document Analysis**: Generative AI can automate the sorting and prioritization of legal documents, enhancing the accuracy and speed of legal document analysis.
- **Due Diligence**: AI can review document estates at speed and offer actionable recommendations, proactively identifying risks, saving time, and conserving legal resources.
- **Contract Drafting**: Generative AI can help draft contracts or leases by analyzing language patterns and clauses associated with legal risks, pinpointing areas that require closer examination.
- **Contract Redlining**: Generative AI can suggest redline edits to align a contract with the preferred positions reflected in a template or playbook.
- **Summarizing Documents**: AI can summarize large volumes of documents, which is particularly useful for due diligence and understanding new legal concepts.
- **Client Communications**: AI can generate legal communications, making it easier for legal professionals to maintain consistent and accurate client interactions.