AI GOVERNANCE BEHIND THE SCENES

Emerging Practices for AI Impact Assessments

DECEMBER 2024

FPF | CENTER FOR ARTIFICIAL INTELLIGENCE

AUTHORED BY

Daniel Berrick
Policy Counsel for Artificial Intelligence, Future of Privacy Forum

ACKNOWLEDGEMENTS

This report benefited from review, recommendations, and contributions from Anne J. Flanagan, Katy Wills, Stacey Gray, Jim Siegl, and Beth Do.

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

All around the world, organizations are seeking to harness the benefits of artificial intelligence (AI) models and systems while managing the risks they may pose to individuals, businesses, and society as a whole. One way organizations are seeking to manage such risk is by conducting AI impact assessments: evaluations that organizations perform to identify and address potential risks associated with AI models and systems. While some organizations have developed their own settled practices for performing these assessments or have adopted models developed by third parties, for many others questions remain around what are the most appropriate steps to take to conduct AI impact assessments.

In response to interest from Future of Privacy Forum (FPF) stakeholders on this point, the FPF Center for Artificial Intelligence began researching stakeholder approaches to AI impact assessments in the spring of 2024. FPF built upon its research and insights by soliciting input from a sample of private sector stakeholders at a July 2024 workshop, additional similar convenings, and one-on-one interviews, in total consulting with over 60 companies over the course of the project. This report is the culmination of that research and is designed to shine a light on the state of play with respect to the implementation of AI impact assessments.

This report examines the considerations, emerging practices, and challenges that FPF's research suggests characterize each step in the AI impact assessment process. Most stakeholders have practices in common when it comes to conducting AI impact assessments, such as taking similar steps at different points in the AI lifecycle. In addition, stakeholders are experiencing pain points at different stages of the AI impact assessment process. The report finds that many organizations are moving rapidly towards their own risk assessment strategies, particularly those that are well-resourced on AI governance, while they await regulatory clarity, although their approaches are not uniform. Commonalities and emerging trends include the following:

- Organizations typically take four common steps when conducting AI impact assessments, including: (1) initiating an AI impact assessment; (2) gathering model and system information; (3) assessing risks and benefits; and (4) identifying and testing risk management strategies.
- The circumstances that trigger an AI impact assessment vary, and there is a trend within organizations to perform multiple assessments at different points in the AI lifecycle.
- When gathering model-system information, organizations typically seek a variety of information, such as details about an AI model's training, use cases, capabilities, and more. Organizations typically have internal teams, sometimes dedicated to AI governance, and, when relevant, seek information from third party model developers and system providers, although many organizations report that they can encounter difficulties obtaining such information.

- A growing number of organizations have sought to integrate AI impact assessments into existing enterprise risk management processes, including those around privacy. In doing so, they have established updated processes for identifying and monitoring risks related to AI, and may escalate AI use cases for review. However, organizations often find anticipating all relevant AI-related risks to be challenging.
- When identifying and testing for AI-related risk, organizations may use both qualitative and quantitative approaches to determine whether risk has been brought within acceptable levels. A variety of factors—the subjective nature of certain risks, the lack of standardized metrics for measuring specific risks, and the indeterminate nature of some AI systems' operational environments—can impede these efforts.

Amongst the conclusions reached in the report, FPF finds that organizations may struggle to obtain relevant information from model developers and system vendors, anticipate pertinent AI risks, and determine whether they have been brought within acceptable levels. While there is no silver bullet that will solve all of these issues today, companies looking to enhance their AI impact assessments should inter alia consider the following:

- Enhancing their processes for gathering information from third party model developers and system vendors, such as by streamlining the number of questions asked, connecting with practitioners at the third party who are capable of sharing relevant details, and, when appropriate, identifying alternatives to the third party's model or system;
- Improving internal education about the multitude of AI risks that can arise, recognizing that these risks can vary between technologies, depend on the deployment context, and emerge at different points in the AI lifecycle; and
- Devising and enhancing measurements for risk management strategies' effectiveness, such as by benchmarking against other companies' approaches and assessing these strategies' effectiveness over time.

In addition to the above, FPF's research shows that implicit in an organization's knowledge stack is the need for both AI governance training across the organization as well as sponsorship for AI governance systems and approaches from the executive level.

INTRODUCTION

The past few years have witnessed the rapid development and deployment of artificial intelligence (AI) models and systems for public and private uses.[1] From music parodies to new medicines, these technologies can enable novel use cases. But they can also raise challenges around bias, fairness, and privacy, among other risks to individuals and society. Around the world, both policymakers and businesses have been grappling with how to support AI's potential positive impacts while minimizing risk and potential negative impacts. In response, governments, standards bodies, civil society actors and businesses have issued laws and published resources on AI governance,[2] yet many organizations remain uncertain about what AI impact assessments entail or which framework to use.

While these efforts have not translated into a single, globally accepted approach to AI policy and governance, many governments and regulators have highlighted the role of AI impact assessments within companies as important tools in managing AI risk. An increasing number of organizations are tasking their CPOs (Chief Privacy Officers), who typically already head privacy and data protection programs, with also leading AI governance efforts. Their job descriptions reflect this trend, with titles like Chief AI Officer becoming more common. This mirrors a general trend towards the integration of AI governance responsibilities into privacy management responsibilities.[3] It is worth noting that FPF's research indicates that CPOs and their privacy management teams are not usually set up to take on the task of AI governance alone by default, and therefore cross-functional support at the executive level is key to ensuring that CPOs are set up for success in the current era. In this respect, uncertainty around the appropriate formula and acceptability of AI impact assessments internally amongst peers and externally amongst regulators—combined with CPOs' evolving roles—raise practical questions, including how to extend existing privacy assessment processes to tackle non-privacy risks, what stakeholders should be included in the AI impact assessment process, and how to measure the effectiveness of risk management strategies.

In light of this emerging dynamic and following interest in this topic from FPF's AI CPO Privacy Executives Network (PEN), a group currently composed of senior-level data privacy professionals working at leading organizations, FPF surveyed members and other companies to gain insight into:

- Common AI impact assessment triggers and the steps organizations often take when performing these assessments;
- Emerging trends in the way companies are conducting AI impact assessments; and
- Challenges that can manifest at different points in the assessment process.

FPF began researching the considerations, emerging practices, and challenges around AI impact assessments in spring 2024. As a first step, FPF reviewed companies' public-facing AI governance documents, which served as the foundation for a list of assessment triggers and considerations that animate each step of the AI impact assessment process.

In the summer of 2024, FPF hosted a workshop to discuss AI impact assessment triggers, considerations, emerging practices, and challenges. More than 20 FPF member companies, including developers and deployers of AI, participated in the workshop, representing a wide range of sectors, such as software, travel and hospitality, and financial services. This was followed by tens of one-on-one meetings with FPF members and several other companies. FPF also sought broader feedback from across the FPF community and guests at its regular convenings over this period.

Through analyzing the results of the workshop, convenings, and interviews, FPF has synthesized the feedback into key findings that point to common emerging practices associated with AI impact assessments, which are presented and discussed in this report.

From this work, FPF found that companies are converging on several practices for conducting AI impact assessments, including accounting for both intended and unintended uses of AI models and systems. However, practitioners continue to face several challenges, such as identifying risks and determining the efficacy of risk management measures. While organizations may wish to use this resource to baseline against a sample of their peers, this report also underscores that much more work

needs to be done to ensure that companies can operationalize AI impact assessments, identify risks, and implement robust risk management practices.

Importantly, by engaging more than sixty companies on the basis of anonymity with a view to FPF sampling, this report brings to the fore the true level of preparedness around AI governance at this time amongst those surveyed. In doing so, this report aims to shine a light on lived experiences with a view to assisting other companies and wider policy stakeholders in their understanding of the current state of play at organizations. The report does not specifically distinguish between AI developers and deployers, although it refers to both throughout, and is designed as a starting point rather than a solution to an organization's AI governance journey.

What are AI Impact Assessments?

The nomenclature around AI impact assessments and its relationship to other evaluations of AI are unsettled. Some governments use "AI risk assessment" and "AI impact assessment" interchangeably[4] while others distinguish them from each other.[5] There are disagreements between organizations that differentiate between AI risk and impact assessments about their relationship to each other.[6] The term "AI impact assessment" lacks a common definition, although the National Institute for Standards in Technology (NIST) AI Risk Management Framework (RMF), which has gained traction in the global AI governance community, describes them as tasks that "include assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others."[7] Organizations also disagree about how AI impact assessments intersect with other kinds of evaluations, such as data protection impact assessments (DPIAs). While DPIAs are distinct from AI impact assessments, they may be part of the same enterprise risk assessment process, **a topic that Step 3, Section I explores in greater detail.**

AI impact assessments are part of AI governance programs, which are the policies and procedures that aim to ensure the responsible development and deployment of AI technologies.[8] Multiple internal teams representing different disciplines may participate in AI governance programs, such as privacy, risk, product, HR, engineering, legal, and marketing. An organization may also engage with third parties, such as model developers and system providers, when they obtain AI technologies from these entities.

Assessments can help organizations build trust in their products and services, counter threats to the organizations, and comply with relevant laws. At a time when policymakers, businesses, and the public are directing significant attention to AI, assessments can be a vehicle for developing trust among various stakeholders.[9] AI impact assessments can also provide organizations with insight into how certain AI activities pose business risks, such as when employees input intellectual property (IP) into a third party's AI tool.[10] In addition to acting as a way to help companies manage existing risk, they can serve as a compass on decisions about whether to proceed with developing or procuring AI products. AI impact assessments are therefore increasingly part of regulators' expectations around how organizations take on and manage AI-related risk.

Below this paper discusses the findings from FPF's research into this area, identifying a set of emerging common steps in the AI impact assessment process, and their implications at an organizational level.

Legislative Approaches to AI Impact Assessments

Some governments including in the EU and Colorado have mandated that organizations conduct versions of AI assessments in specific circumstances.[11] General consumer protection laws may also require these assessments.[12] These assessments may function as accountability vehicles that regulators can demand from organizations to ensure compliance.[13] Other policymakers have devised voluntary frameworks for AI governance, such as NIST's AI RMF and Singapore's Model AI Governance Framework, which highlight AI impact and risk assessments' role in helping organizations identify and address AI risks.[14]

This widespread interest among policymakers in assessing AI risk through the use of an assessment procedure has not translated into uniform requirements across jurisdictions. Despite this multitude of laws, several trends have emerged from legislative efforts and voluntary frameworks that may help organizations chart out their approach to AI risk management. **Further information on the approaches adopted by different jurisdictions may be found in detail in the Appendix at the end of the report.**

Of note, jurisdictions around the world tend to:

- Contend with the needs of many different stakeholders when formulating laws and regulations to address AI use benefits and risks;
- Craft voluntary frameworks that align with emerging global standards around AI; and
- Use frameworks (especially the OECD AI Principles) as a template for national AI plans.

However, this is a rapidly evolving and changing area, and it is expected that AI regulations as well as clarity around their implementation requirements will continue to emerge. One example of this at the time of writing is the EU process on the Code of Practice for the AI Act, which is under multistakeholder consideration until early 2025.[15]

In conducting the following report, FPF's research concludes that organizations appear to be preparing for eventual regulatory or partial oversight in this space regardless of the jurisdiction they are in given the high level of policymaker interest.

Key AI Governance Steps in the AI Impact Assessment Process

FPF's research has found that organizations, usually businesses, typically go through four common steps when conducting AI impact assessments. Importantly, while companies generally recognize the following as steps in the AI impact assessment process, the order in which they address them can and will vary. This paper categorizes these steps as follows and explores them in detail both in Figure 1 on the following page and throughout the rest of this report:

Step 1: Initiating an AI Impact Assessment

The circumstances that trigger an AI impact assessment will vary based on a variety of factors, such as: applicable law; proposals to develop a new product and service, or use existing technology in a new way; and an organization's role within the AI ecosystem.

STEP 2: Gathering Model and System Information

Organizations tend to pose questions to relevant external parties and internal teams to learn about how the model or system was created and works.

STEP 3: Assessing Risks and Benefits

Reviewers use the model-system information to determine risks and benefits for a particular use case, which will inform an organization's decisions about how to proceed with an AI project.

Step 4: Identifying and Testing Risk Management Strategies

Organizations select risk management strategies based on their responsiveness to a specific, identified risk, and continue to test their efficacy during the system's operation. The organization can determine and record appropriate next steps, such as advancing the AI project to the next stage in the lifecycle, after identifying and testing the risk management strategies.

These steps are outlined more comprehensively in the following graphic and discussed at length in the following chapters.

Figure 1: Key Steps in the AI Impact Assessment Process

1 INITIATING AN AI IMPACT ASSESSMENT

The circumstances that trigger an AI impact assessment will vary based on a variety of factors, such as applicable law, product or service development, and an organization's role within the AI ecosystem (e.g., developer v. deployer).

These incentives for conducting AI impact assessments may arise at different points in the AI development and deployment lifecycle. For example, an organization may perform an initial assessment during the development phase and one after substantial changes are made to the model or system. Additional assessments may be necessary when deployed in different contexts.

2 GATHERING MODEL-SYSTEM INFORMATION

In order to be able to assess the risk that an AI system may pose, organizations must understand how that system works and was created. The list of common considerations below is not exhaustive. An organization that obtains a model or system from a third party may focus its inquiry on the maturity of the third party's AI governance framework and whether it has implemented controls to minimize certain risks. An organization may also direct less scrutiny towards the AI models and systems of more well-known third parties and those the organization has a relationship with. However, FPF's research has observed that organizations tend to collect a baseline of relevant information in most situations.

Examples of relevant information include:

- How was the model trained;
- What are the model or system's use cases;
- Who are the end users, and where will the system be deployed; and
- Potential use cases for the system and organize them by category (e.g., intended and unintended uses).

3 ASSESSING RISKS AND BENEFITS

Reviewers take the information about the model or system into account to determine the AI model or system's risks and benefits for a particular use case. The type and level of risk present will inform what risk management strategies an organization implements to reduce the risk to within an acceptable threshold. Information about benefits will also inform an organization's decisions about how to proceed with an AI project. As part of this step, organizations may consider:

- The potential risks of harm and benefits associated with a particular use case;
- How the generated list of system risks compares with the company's taxonomy of risks (divided into categories like high, medium, and low); and
- If the assessment is being conducted during the development or post-development phase, including fine tuning during deployment, the existence and context of risks in system testing.

4 IDENTIFYING AND TESTING RISK MANAGEMENT STRATEGIES

Organizations select risk management strategies based on their responsiveness to a specific, identified risk. For example, if an organization identifies hallucinations as a risk, the organization tailors its response to this risk, such as by tweaking the AI's implementation to reduce hallucinations and ensuring ongoing monitoring of outputs for hallucinations.

Once an organization has identified risk management strategies, it can test their efficacy and balance the residual risk against the benefits to determine appropriate future steps. The organization may then record and operationalize the final decision, such as advancing the AI project to the next stage in its lifecycle.

STEP 1: INITIATING AN AI IMPACT ASSESSMENT

KEY TAKEAWAYS

- The circumstances that trigger an AI impact assessment are numerous but not uniformly applicable across all AI development or deployment scenarios; and
- There is a trend within industry to perform multiple assessments at various points in the AI lifecycle, such as when a model or system is being designed, when it undergoes significant modifications, or when it is deployed in a new context.

Examining the Key Takeaways Relating to Initiating an AI Impact Assessment

I. There Are Numerous Catalysts for AI Impact Assessments, Including Legal Requirements, Governance Norms, and Ethical or Business Risks

There are many conditions that can trigger the initiation of an AI impact assessment, such as an impact assessment being a requirement by law, an organization's role in the AI ecosystem (e.g., developer or deployer), and new uses of existing AI technologies and product developments that require additional risk management. Companies may initiate an AI impact assessment after identifying legal, ethical or business risks. Organizations may also routinely perform such assessments as part of their enterprise risk management program, **a topic that the next section explores in greater detail**.

Crucially, these catalysts are context dependent and not consistently present across all AI development or deployment scenarios. Nonetheless, many companies agree that these are the leading categories of catalysts across business sectors and territories.

II. There is a Trend Within Industry to Perform Multiple Assessments at Different Points in the AI Lifecycle Due to Both Legal and AI Governance Demands

FPF's research shows that many organizations perform several AI impact assessments throughout the AI development or deployment lifecycle to comply with laws and/or operationalize internal AI governance norms. Specific triggers for conducting assessments may exist at the design, development, and deployment phase of AI models and systems.[16] For example, an organization may perform an initial AI impact assessment to decide whether a product team may develop a new AI application. This will usually be followed by subsequent assessments once the organization begins developing or fine tuning a model or system, especially around launching in the market. Assessments may also occur at particular cadences after deployment, including when a model undergoes a significant change. This scenario is especially pertinent to generative AI, as the model may change based on the training data it ingests in a new context.

FPF's research has observed that CPOs and risk managers tasked with translating laws and frameworks into practice have adopted a combination of approaches to achieving compliance.[17] Laws such as the Colorado AI Act (CAIA) generally require deployers, or a third party that the deployer contracts with, to perform assessments every year and when certain changes to high-risk AI systems are made.[18] Uncertainty among companies around what is required for impact assessments is greater in jurisdictions where regulators have not promulgated rules and as a result FPF has observed that organizations are adopting a baseline of proactive AI governance measures while they prepare to react to any potential future regulatory demands. As discussed in the introduction, despite a lack of clear harmonization in regulatory approaches, trends and best practices are emerging and can therefore be anticipated to some extent.

Meanwhile, voluntary understandings of AI governance have led some organizations to perform several assessments to account for new use cases, features, and risks that can emerge at different stages in the AI lifecycle.[19] For example, organizations may only be able to detect performance issues, such as those related to concept drift, after they deploy a system.[20] Multiple assessments can also help companies assess their risk management strategies' effectiveness over time by providing a body of evidence around the AI's performance in the wild.

III. AI Impact Assessments Are Increasingly Demanded By Business Risk Managers, Especially For Generative AI

All mature organizations anticipate and seek to measure risk as part of their regular operations. FPF previously explored the area of risk metrics for privacy,[21] but it is generally perceived by surveyed stakeholders that generative AI can present considerable intrinsic risk to the organization regardless of any regulatory risk given its iterative nature, unprecedented power, and uncertain future impact.

While this includes heightened sensitivity to potential reputational damage from unintended misuse of AI products and services, it also encompasses the possibility of algorithmic disgorgement or destruction of a business's AI model. Several of the companies FPF spoke to have a single intake process whereby all risk management processes, including those for AI, are centrally managed. Others signaled that they are finding it challenging to manage the administration of or contribution to multiple internal risk management processes, including but not limited to AI. For CPOs, the risk of disgorgement was dealt with as part of wider data governance processes. Differences were also observed between developers and deployers, with deployers concerned about data subjects' data in the AI systems they use, while developers were more focused on managing risk around model build.

STEP 2:
GATHERING MODEL AND SYSTEM INFORMATION

KEY TAKEAWAYS

- Organizations often collect a variety of information relating to an AI model and associated systems to understand risks and benefits, with a focus on data;
- Many organizations encounter challenges with obtaining information from third-party AI model developers and system providers that they seek in order to complete AI impact assessments;
- Most organizations account for intended and unintended uses of AI models and systems when they conduct AI impact assessments;
- There is an emerging trend towards organizations using cross-functional groups of internal and external stakeholders to surface model-system information; and
- Some organizations use one assessment for multiple, comparable AI use cases, although the point at which use cases are "comparable" is unclear.

Common Considerations When Gathering Model-System Information Include How the System Works, How it Was Created, and its Potential Risks

Once an AI impact assessment is initiated, organizations must understand how that model or system works and potentially how it was created in order to be able to assess the impact or risk that the AI model or system may pose. This includes any data that it is likely to process.

The questions an organization will pose can depend on the nature of the model or system, the AI use case, and whether the organization developed or acquired the model or system—and whether personal data or personally identifiable information (PII) will be processed. An organization that obtains a model or system from a third party may focus its inquiry on the maturity of the third party's AI governance framework and whether it has implemented controls to minimize certain risks, including training the model on personal data or PII. An organization may also direct less scrutiny towards the AI models and systems of more well-known third parties and those the organization has a relationship with.

FPF's research indicates that organizations tend to collect a baseline of relevant information in most situations, including:

1. The platform, tool, or team that will be supporting the new model or system, how their processes currently work, and the principal business goal being achieved by adopting the AI technology;
2. The AI's capabilities, limitations, and nature;
3. How the AI model was trained (e.g., How will/did the organization obtain the training data?; What kind of data is needed? Does the training data include personal and/or sensitive data elements?; Is the data representative?);
4. Potential use cases for the system and organize them by category (e.g., intended and unintended uses); and
5. For each use case, organizations tend to compile the following information:
    a. How the system will solve problems raised in each intended use case;
    b. The system's end users—the people interacting with or using the AI system—and subjects of decisions made using the AI system;
    c. Historically marginalized or vulnerable communities that users or individuals subject to the system might be part of;
    d. The geographies in which the system will be deployed, and known cultural considerations and languages used there; and
    e. The operational environment's nature, such as changes to the way users interact with an AI system over time.

The list of common considerations outlined on the previous page is not exhaustive and is representative of the top answers surfaced by consulted FPF stakeholders.

Examining the Key Takeaways Relating to Gathering Model-System Information

From FPF's research, before an organization assesses risks and benefits, it typically compiles information about the model, system, and use case. When considering the specific risks associated with use cases, organizations generally evaluate both intended and unintended applications of AI models and systems. Reviewers will often pose numerous questions to relevant internal teams and external parties to uncover relevant details, such as how the model was trained and where the system will be deployed. However, organizations may encounter difficulties obtaining this information from external parties, such as model developers and system providers.

I. Organizations Often Collect Information Relating to an AI Model and System to Understand Risks and Benefits, With a Focus on Data

Many organizations seek a variety of information to accurately anticipate the risks posed by AI models and systems. What is most relevant to a reviewer's assessment will likely depend on the model or system. However, some of the more frequently referenced information categories include details relating to: (1) model training; (2) capabilities and limitations; (3) impacted individuals and groups; (4) geographic deployment area(s); (5) the nature of the deployment environment; and (6) existing guardrails for users and safety training.

Uncovering relevant information may necessitate numerous questions, both limited response (e.g., yes/no) and open-ended in nature, in order to understand a particular aspect of the model or system. For example, an organization may need to pose multiple questions to learn about the model's training, such as: (1) what data was used to train the model?; (2) is any of this data sensitive?; (3) how and from where was the data obtained?; and (4) what is the distribution of values in the data? Many organizations highlight the importance of knowing the businesses' risk tolerance to determine whether a risk exists and has been appropriately addressed. These tolerances will likely vary between organizations, and there is no single method for measuring them. **Step 3, Section III analyzes risk tolerances in greater detail**.

FROM THE FIELD

Questions About Model and System Capabilities and Limitations

There are numerous questions related to a model or system's capabilities and limitations that reviewers may seek answers to. Some of the questions that practitioners highlighted include (i) What limitations exist on how users can use AI on the input data?; (ii) Was testing done to determine the existence of bias? If so, what were the results?; and (iii) What metrics were used to determine the model or system's accuracy or bias?

II. Many Organizations Encounter Challenges With Obtaining Information From Third-Party AI Model Developers and System Providers That They Seek to Complete AI Impact Assessments

Third-party AI model developers and system providers may not or cannot provide the information an organization needs to perform assessments. Many organizations use models developed by third parties to power their AI systems,[22] which they may fine tune for a particular use case and add input and output guardrails to.[23] Because they were not involved in the model's training, these organizations typically want to consult with the third party developer to obtain relevant information for the impact assessment.[24] In addition to helping surface potential use cases, the model developers may possess unpublished information, such as that relating to the model's training and identified risks.[25]

FROM THE FIELD

Legal Requirements to Consult with Third Party Developers and Vendors

While some organizations may seek information from third party developers and vendors to learn about an AI model or system, certain sectors are governed by laws that necessitate these interactions to achieve compliance. One stakeholder indicated that as a bank, regulators expected them to engage with third parties, including model developers and system vendors, to comply with sectoral risk management regulations. They may seek information from these third parties, such as details about what data the system would send back to the third party.

While organizations may expect their internal teams to consult with third parties to learn about their models and systems, technical barriers may prevent some developers and providers from providing requested information. For example, downstream users of AI models may need to know how a third party developed a model in order to address bias and harmful content in the training dataset.[26] However, the size of the training datasets underpinning the latest generative AI models can prevent insights into these datasets.[27] This may be exacerbated by the training data's sources not communicating the data's biases and other limitations.[28] The black box problem may also inhibit third parties from providing information about their models.[29]

Third parties may not willingly provide information about their AI models and systems for commercial and proprietary reasons too. A model developer or the specific employee interacting with an organization (e.g., business, sales, and lawyers who do not work on AI governance matters) may lack the technical expertise needed to answer questions. Information asymmetries between third parties, such as the model developer and system vendor, can add to the lack of transparency downstream, as model developers may not have shared information with the system vendor. Without sufficient insight into the risk metrics and methodologies used in the development process, downstream organizations that use third party models and systems may be unable to account for biases and other harmful content in the training data.[30]

FROM THE FIELD

Model-System Cards and Improvements in Transparency Between Third Parties and Downstream Organizations

Providing model and system information cards to downstream organizations is a nascent practice. Several generative AI model developers have published these cards, such as OpenAI, which released a system card for its GPT-4o model.[31] Some stakeholders shared that their organizations are creating cards of their own to share with customers and the public.

Several stakeholders signaled that third parties have increased the amount of information they share in recent years, although the issues described in this section persist. One company suggested that these transparency issues can be overcome by referencing sectoral legal obligations and risk to the third party's reputation from withholding information. However, these legal expectations may not exist in all sectors, and the number of organizations that highlighted the transparency issues described in this section suggests that citing reputational risk may not be effective in all interactions with third parties.

III. Most Organizations Account for Intended and Unintended Uses of AI Models and Systems When They Conduct Assessments

Companies generally tend to address intended and unintended uses of AI models and systems in their AI impact assessments. Organizations often build models and systems with a use case in mind, but these technologies can find applications beyond their initial design.[32] These unintended uses can produce positive outcomes,[33] but they can also raise risks that model and system designers did not anticipate.[34] Many companies therefore tend to analyze intended and unintended applications to unearth a broader range of AI risks and inform appropriate risk management strategies.[35]

IV. There is an Emerging Trend Towards Organizations Using Cross-Functional Groups of Internal and External Stakeholders to Surface Model-System Information

Organizations are trending towards engaging with internal teams and external parties to obtain model-system information. These consultations can take the form of cross-functional groups composed of product, engineering, legal, and other teams. Of these teams, business units and product teams tend to have a more prominent role in identifying intended uses of AI models and systems given their proximity to business decisions. Many organizations also consult with external parties, such as third-party model developers, to ascertain potential unintended uses. However, as noted above, many organizations highlighted the difficulty with gathering information about AI models and systems from third parties.

V. Some Organizations Use One Assessment for Multiple, Comparable AI Use Cases, Although the Point at Which Use Cases are "Comparable" is Unclear

An increasing number of organizations utilize one assessment to cover multiple AI use cases if those use cases are similar based on gathered information, or if sectoral rules for certain industries, such as pharmaceuticals, healthcare, and financial services, already impose risk reporting requirements. Organizations can find it challenging to complete an assessment for every use case due to the amount and frequency of internal product developments. They may instead use one assessment to address a group of similar applications of an AI model or system. This practice aligns with language found in the CAIA, which states that an assessment may cover "a comparable set" of deployed systems.[36] However, there is little direct guidance on how to conclude that systems are comparable to each other.[37] An organization may perform a separate assessment based on a lack of similarities with the previously reviewed use case and when the one at issue raises unique risks.

STEP 3:
ASSESSING RISKS AND BENEFITS

KEY TAKEAWAYS

- Some organizations think that AI impact assessments can draw inspiration from and feed into existing risk assessment processes;
- Many organizations find it challenging to anticipate all AI risks due to the number of known AI risks, "general purpose" AI models' multitude uses, and the indeterminate nature of the environments in which some AI systems operate;
- Organizations' conceptions of risk are not static, shifting based on changes to the business, internal practices, and regulations;
- There is an emerging trend among organizations towards utilizing risk-benefit matrices to categorize AI use cases based on their risks and benefits, but industry has not converged on a single approach for escalating high-risk use cases for review; and
- There is a growing trend within organizations towards designating internal teams that monitor for and own AI risk, although there is less uniformity around whether these responsibilities should be concentrated in a single team.

Common Considerations When Conducting the Risk-Benefit Analysis Include The Types and Levels of Risk, the Potential Benefits of the System, and the Technical and Legal Context in Which the System Operates

Reviewers take the information about the model or system into account to determine the AI model or systems' risks and benefits for a particular use case.

The type and level of risk present will inform what risk management strategies an organization implements to reduce the risk to within an acceptable threshold. Information about benefits will also inform an organization's decisions about how to proceed with an AI project.

The following is a common but not exhaustive list of issues that organizations may surface when they conduct the risk-benefit analysis:

Figure 2: Typical Steps for Assessing AI Risk

1 **Define and describe potential risks of harm associated with a particular use case, considering, for example:**

a. Of the uses and impacted individuals or geographies, are any of them "sensitive"? (e.g., decisions about mental health or people with protected characteristics, or do the deployment geographies implicate export controls)
b. How could the use case possibly create, exacerbate, or reduce inequalities or discrimination against particular communities?
c. How accurate are the outputs?
d. Is there the potential for leakage or misuse of forms of regulated data?
e. Could any individual rights be infringed? (e.g., right to life, freedom of religion, and the right to vote)
f. May the system output demean, stereotype, or erase a group?
g. Does the system implicate a decision that will affect access to a consequential resource? (e.g., education or finance)
h. What would the impacts of the system's failure be on impacted users and subjects?
i. How could intentional or unintentional misuse negatively impact each affected user and subject, and do the effects differ between groups?

2 **Define and describe potential benefits associated with a particular use case, taking into account, for instance:**

a. Will it help make a process more efficient, simpler, cheaper? (e.g. marketing automation)
b. Will it help the organization look at a large data set in a useful way to find patterns or make predictions? (e.g. fraud detection, loss prevention)
c. Will it help the business scale to provide personalized services to a much bigger audience? (e.g. personalized recommendations and discounts)
d. Will it replace some expensive expert skill? (e.g. reading MRI scans)
e. Will it make information more accessible? (e.g. chatbot to access tech documentation)

3 **Compare the generated list of system risks against the company's taxonomy of risks divided into categories like high, medium, and low:**[38]

a. Risk categories should come from a diverse, multi-stakeholder working group, which may include external input from civil society or other groups;
b. Flag risks that meet a predetermined threshold (high, medium, low, for example); and
c. Assess whether the identified risks elicit further review based on previously reviewed, similar use cases.

4 **If the assessment is being conducted during the development/fine tuning or post-development phase, document the existence and context of risks in system testing.**

a. Create metrics before testing to measure performance across different subgroups, communities, and demographics applicable to the use case.[39]

5 List laws/regulations that could apply to the system or training data and define their requirements, although this may be more challenging for organizations with more of a global reach. Before conducting the assessment, companies may define their internal risk standards through a policy and then identify which laws impose additional requirements that the policy needs to consider.

Examining the Key Takeaways Relating to the Risk-Benefit Analysis

A growing number of organizations integrate AI impact assessments into existing enterprise risk management processes, and use these processes as inspiration for how they structure the assessment. Various factors, such as the economic sector, emergence of new laws, and the addition of novel technologies and features, can affect how much risk an organization is willing to take on. These risk thresholds, also known as risk appetite or risk tolerance, can inform the matrices some organizations create to categorize risk and determine next steps for an AI use case. Companies typically designate teams to monitor for and own AI risks. Despite these efforts and improvements in industry's understanding of AI risks, organizations can still find it challenging to anticipate all and assess the likelihood of AI risks.

I. Some Organizations Think That AI Impact Assessments can Draw Inspiration From and Feed Into Existing Risk Assessment Processes

A growing number of companies are integrating AI impact assessments into established risk assessment processes (e.g., CPOs are updating their DPIA processes), and using these processes' structures as models for AI impact assessments. Many organizations deploying AI systems will coordinate their in-house AI impact assessments with other technology risk assessments, such as privacy and security assessments. As part of this effort, some organizations append and compare AI impact assessment questions to existing risk assessment questionnaires before sending them to the team leading an AI project. Under this approach, if the questionnaires supporting a DPIA and an AI impact assessment have components that address the same issue, the organization may want to avoid having the same question(s) presented in both assessments. Companies have done this to stop duplicating processes and queries, which can frustrate teams involved in the assessment.

FROM THE FIELD

Variation in Internal Teams Responsible for Answering Impact Assessment Questionnaires

Because AI projects can originate from different parts of the organization, the lead team may vary from project to project. A product team developing a LLM may be the lead team for that project, while a HR department may assume lead team responsibilities when it seeks to procure an analytics tool from a vendor. The lead team is typically responsible for completing the model-system information questionnaires. However, several stakeholders shared that they expect the lead team to consult with other teams, such as risk, data scientists, and engineering, as well as third party developers or system vendors when appropriate, to obtain relevant input. One stakeholder indicated that they proactively share certain questions with technical and legal teams in recognition of the lead team's lack of expertise in these domains.

Many companies draw from existing risk management processes to inform the structure and flow of AI impact assessments, such as the assessment's overarching steps. Organizations generally recognize that creating cross-functional teams can help disseminate best practices for performing assessments internally, but this may not happen in practice. Siloing of teams within an organization can occur instead, making it challenging for those that have conducted other kinds of risk assessments (e.g., privacy teams) to share their insights with colleagues responsible for AI development, deployment, and operations. Even when these barriers can be overcome, companies generally understand that AI-specific considerations (e.g., issues with measuring AI risks) limit the instructional value of experiences with conducting other assessments.

II. Many Organizations Find it Challenging to Anticipate All AI Risks Due to the Number of Known AI Risks, "General Purpose" AI Models' Multitude Uses, and the Indeterminate Nature of the Environments in Which Some AI Systems Operate

The number of potential AI risks, "general purpose AI models' multitude uses, and indeterminate nature of the environments in which some systems operate can make it challenging for organizations to anticipate all AI risks. **As Sections III and IV of this step discuss**, laws and frameworks can inform which risks an organization tries to manage, but organizations tend to look beyond these sources to inform their classifications of AI risks. Organizations may focus on certain risks for non-legal reasons, such as the impact they have on the business' reputation and IP risks. Industry has also made commitments to address specific AI risks, such as "avoiding harmful bias and discrimination, and protecting privacy."[40]

While laws, organizational commitments, and non-legal rationales may prompt industry to fixate on certain AI risks, the number of potential AI risks is large and continues to grow.[41] Many organizations try to account for a broad array of AI risks (e.g., IP, labor issues, competition) when they perform assessments. However, organizations may be more familiar with and focus on addressing certain risks, such as the fairness of automated decision-making systems.[42] For example, an organization that holds large amounts of confidential information, trade secrets and other intellectual property (IP) may focus on addressing the IP-related risks of an AI project.

FROM THE FIELD

Use of AI Impact Assessments to Identify and Address Risks to the Business

Policymakers have described AI impact assessments as tools for combatting risks to individuals and society. However, many stakeholders signaled that organizations use these assessments to address risks to the business too. For example, an organization that holds large amounts of IP may focus on addressing the IP-related risks of an AI project, as these risks can have significant impacts on its business.

Difficulties with anticipating a dynamic operational environment's qualities and some models' multitude uses can also make it challenging for industry to measure risk likelihood. A prominent risk formula says that risk equals the likelihood of a negative impact occurring times the impact's severity.[43] Organizations can experience challenges assessing likelihood due to the indeterminate nature of the environments in which some AI systems operate. Such environments are characterized by qualities that the organization may not be able to anticipate, such as changes in user base and the way humans interact with an AI system over time.[44] These changes can affect whether a risk materializes, so their unpredictability can impede an organization's understanding of what risks exist and need to be managed. Organizations may also have difficulty anticipating all of the risks raised by "general purpose"[45] AI models (e.g., OpenAI's ChatGPT). These models have multitude uses, potentially including harmful applications, but organizations may be uncertain as to the likelihood of these uses occurring. These factors may impede identification of all of a model or system's risks, which could hamper the development of risk mitigation strategies.

III. Organizations' Conceptions of Risk Are Not Static, Shifting Based on Changes to the Business, Internal Practices, and Regulations

An organization's risk appetite level is dynamic, and can affect what qualifies as a risk and whether to advance an AI project to the next stage. Risk appetite, also called risk threshold or risk tolerance, refers to how much risk an organization is willing to assume with respect to certain activities. The amount of risk an organization will bear is an important prerequisite to conducting an AI impact assessment and determining next steps for an AI project; If an AI use case's inherent risk—the risk posed by a use case absent safeguards—exceeds the organization's risk appetite, the organization must introduce safeguards that reduce risk to a certain level (i.e. residual risk) and determine if the residual risk is within the risk appetite.[46]

Industry lacks a unified view on which teams establish an organization's risk appetite. Companies identified a variety of teams, including the organization's board, enterprise risk management team, and the general counsel office, while others highlighted how setting risk tolerances is a multi-team activity. Organizational differences related to regulatory environments, economic sector, product offerings, and reputation concerns mean that risk appetites are also not uniform across industries. For example, financial institutions and companies offering services to minors may have lower risk appetites due to the amount of regulatory oversight of these sectors and expectations surrounding these products and services.

FROM THE FIELD

Risk Appetite Variation at the Activity Level

Rather than having a risk appetite that is the same across all of the organization's activities, some companies may have risk appetites that vary depending on the activity at issue. One organization shared that they have multiple risk appetites, with each corresponding to a particular risk. While they had no tolerance for cybersecurity risks, they were more willing to stomach risks associated with lending to teenagers, provided the residual risk was within their risk threshold. In addition to their potential to vary across activities, risk appetites may evolve over time. The factors discussed in this section, such as changes to the regulatory environment and product offerings, may underpin these shifts.

In addition, given current market pressures on businesses to either develop or deploy regardless of sector, risk managers may face resistance to their own recommendations that a particular use of AI is very risky and may instead be asked by leadership to focus instead on risk management measures. In that scenario risk managers typically turn to other stakeholders for support in identifying and mitigating specific risks.

IV. There is an Emerging Trend Among Organizations Towards Utilizing Risk-Benefit Matrices to Categorize AI Use Cases Based on Their Risks and Benefits, but Industry Has Not Converged on a Single Approach for Escalating High Risk Use Cases for Review

Some organizations have developed risk-benefit matrices to determine how to proceed with an AI use case. These matrices consist of several boxes, each representing a specific level of risk and benefit (e.g., high/medium/low).[47] While laws can inform which box a use case falls into, such as when a statute deems a particular activity "high risk," organizations may also develop their own internal considerations that affect how they classify use cases in a matrix.

Where a use case falls within a matrix can influence how much attention reviewers give it, with higher risk applications of an AI system or model generally undergoing more scrutiny.

Industry has not converged on a single approach for escalating a use case for review and approval if an organization deems it high risk. Several organizations have established one or more committees composed of representatives from different teams and disciplines that evaluate AI use cases for risk. For example, these committees may have professionals from legal, security, AI and data ethics, policy, privacy, and governance. Senior figures at the organization, potentially including the CEO, may become involved in the review and approval of high-risk use cases. Under these circumstances, senior figures will determine whether the use case should proceed to the next step in the AI lifecycle.

V. There is a Growing Trend Within Organizations Towards Designating Internal Teams That Monitor for and Own AI Risk, Although There is Less Uniformity Around Whether These Responsibilities Should be Concentrated in a Single Team

Many organizations think that they should monitor for and own AI risks, even when some third party developers assess and provide information about these risks.[48] However, organizations are divided on whether a single actor within the company should have primary responsibility for monitoring and owning a project's AI risks. Under one approach, an organization's business units own the risk and have primarily monitoring responsibility for them. Other teams, such as legal, compliance, risk and ethics teams, may become involved at later review stages to identify errors in the business unit's checks. Depending on the sophistication of the organization's AI governance program, this multi-team monitoring process can take the form of three lines of defense that include: (1) the business unit; (2) compliance and risk management teams; and (3) audits. Another group of organizations indicated that while business units own the risk, other teams, such as data ethics, legal, and technology, have primary monitoring responsibility. A few companies shared that the organization's size can affect the allocation of responsibilities. For example, legal departments at smaller organizations may monitor for and own AI risks.

STEP 4: IDENTIFYING AND TESTING RISK MANAGEMENT STRATEGIES

KEY TAKEAWAYS

- Some organizations utilize qualitative and quantitative evaluations for determining risk management strategies' efficacy;
- Many organizations encounter difficulties assessing whether risk has been brought within acceptable levels due to the subjective nature of certain risks, the lack of standardized metrics for measuring specific risks, and the indeterminate nature of some AI systems' operational environments; and
- Organizations generally engage with both internal and external stakeholders to identify and understand the effectiveness of strategies for addressing risk.

Common Considerations When Identifying and Testing Risk Management Strategies Include Identifying Specific Risks, Tailoring Strategies to Address those Risks, and Measuring Effectiveness

Organizations select risk management strategies based on their responsiveness to a specific, identified risk. If an organization identifies hallucinations as a risk, the organization tailors its response to this risk, such as by tweaking the AI's implementation to reduce hallucinations and ensuring ongoing monitoring of outputs for hallucinations. Once an organization has identified risk management strategies, it can test their efficacy and balance the residual risk against the benefits to determine appropriate future steps. The organization may then record and operationalize the final decision, such as advancing the AI project to the next stage in its lifecycle.

Depending on the risks present, the following risk management strategies may be relevant:

1. Human review or oversight when using the system;
2. Guidelines or restrictions on the system's use when it makes certain determinations, such as those related to benefits;
3. Secure handling measures for data inputs and outputs used to train the system or are ingested by it; and
4. Measure performance, potentially with the support of independent teams that test for bias, across different subgroups, communities, and demographics applicable to the use case.

Examining the Key Takeaways When Identifying and Testing Risk Management Strategies

A mixture of internal teams and external parties typically help organizations identify and determine the effectiveness of risk management strategies. Organizations may use qualitative and quantitative approaches to assess these strategies' effectiveness, but they may still experience challenges determining whether risks are within acceptable limits.

I. Some Organizations Utilize Qualitative and Quantitative Evaluations For Determining Risk Management Strategies' Efficacy

Organizations often use results from quantitative and qualitative evaluations to determine an AI use case's risk level and whether a risk management strategy has addressed it.[49] Organizations have used risk management strategies, such as classifiers and prompt engineering, to address model and system risks.[50] There is no panacea for addressing AI risks. Practitioners instead select risk management strategies based on a variety of factors, such as the AI's development stage and the type of system at issue.[51]

Organizations have used quantitative and qualitative approaches to assess the effectiveness of risk management strategies. A risk score is an example of the quantitative approach that factors in mitigations in order to produce a numerical value representing a level of risk.[52] The qualitative approach emphasizes gathering diverse stakeholder input, such as through interviews, in order to understand risk and identify strategies for managing it. After implementing controls to manage the risks identified during testing, organizations may retest their system using these approaches to assess the strategy's efficacy. However, when evaluating post-management test results, industry may struggle to determine whether risk has been brought within acceptable levels.

II. It is Often Challenging for Organizations to Determine Whether They Have Brought Risk Within Acceptable Levels

Organizations often find it difficult to assess risk management strategies' efficacy due to the subjective nature of and the lack of standardized metrics for measuring certain risks, and the indeterminate nature of some AI systems' operational environments. This challenge can be greater for risks that are less related to an AI system's technical operation,[53] but several organizations indicated that the dynamism of and uncertainty around an AI system's operational environment can generally hinder efforts to understand whether risks are adequately reduced. For example, a company's tests of a chatbot prior to operation may demonstrate that mitigations have significantly lowered the output of stereotypes. However, differences between the testing and operational environments may undermine the generalizability of these test results.[54]

Even when the risks are known, some of these risks involve subjective values that lack a single metric for measuring their presence or reduction.[55] What metric an organization should use may depend on the context, including who or what a potential disparity relates to in the case of bias metrics.[56]

Some metrics' performance may vary across different AI uses, limiting their utility in certain circumstances.[57] Other risks, such as those related to trust and safety, lack widely established metrics.[58] Despite the existence of approaches for determining the effectiveness of risk management strategies, these challenges can frustrate efforts to assess whether these strategies succeeded at bringing risks within acceptable thresholds.

III. Organizations Generally Engage With Internal Teams and External Parties to Identify and Understand the Effectiveness of Strategies for Addressing Risk

Organizations generally consult with different parties to help them identify risk management strategies and assess their effectiveness. For example, engineering teams can provide organizations with insights into emerging risks and how to address them. Since some organizations incorporate third-party models into their AI products and services, they may solicit information from the third party about the mitigation measures they implemented during development.[59] However, **as discussed in Step 2, Section II**, these solicitations are not always successful. Once a system is in operation, organizations can establish user feedback mechanisms to learn about a risk management strategy's effectiveness.[60]

CONCLUSION: THE STATE OF PLAY AND LOOKING AHEAD

As AI impact assessments are increasingly mandated by law and become a part of AI governance programs, organizations have grappled with the best ways to perform them. This has already led to progress at different points in the AI impact assessment process, from what triggers an assessment to how organizations should structure their teams to surface information about AI models and systems. While industry lacks a general unified approach to AI impact assessments, they have converged on several practices for different parts of the assessment. Examples of these practices include considering both intended and unintended applications of AI, and organizing cross-functional teams to gain insights into models and systems.

Despite this progress and the energies being devoted to addressing AI risks and benefits, pain points remain for organizations conducting AI impact assessments. These challenges are not confined to a single part of the assessment process. Organizations may struggle to obtain relevant information from model developers and system vendors, anticipate pertinent AI risks, and determine whether they have been brought within acceptable levels. While there is no silver bullet that will solve all of these issues today, **companies looking to enhance their AI impact assessments should *inter alia* consider the following**:

- Enhancing their processes for gathering information from third party model developers and system vendors, such as by streamlining the number of questions asked, connecting with practitioners at the third party who are capable of sharing relevant details, and, when appropriate, identifying alternatives to the third party's model or system;
- Improving internal education about the multitude of AI risks that can arise, recognizing that these risks can vary between technologies, depend on the deployment context, and emerge at different points in the AI lifecycle; and
- Devising and enhancing measurements for risk management strategies effectiveness, such as by benchmarking against other companies' approaches and assessing these strategies' effectiveness over time.

In addition to the above, FPF's research shows that implicit in an organization's knowledge stack is the need for both AI governance training across the organization as well as sponsorship for governance systems from the executive level. The continuous evolution of AI at a technological level, the changing legal landscape around AI, and the operational priorities of organizations will continue to shift and interplay such that it is essential for organizations to not only continue their work on developing appropriate AI impact assessments, but to stay agile in responding to the environment around them. In doing so, organizations should be better equipped to harness the benefits of AI while meaningfully managing the risks it poses to individuals and society.

APPENDIX

Selection of Global Requirements for AI Impact Assessments

Below is a selection of jurisdictions around the world that have introduced legal requirements or authoritative guidance on how and when to conduct artificial intelligence (AI) impact and risk assessments, typically in the context of broader initiatives on AI. The resource is non-exhaustive and does not include pending legislation[61] or DPIA requirements in comprehensive privacy and data protection laws.[62]

In addition, this table includes one U.S. state, Colorado, and its requirements for general privacy and data-related risk assessments, because they exist in combination with underlying novel (U.S.) legal requirements for high-risk uses of AI.

In general, jurisdictions around the world are:

- Contending with the needs of many different stakeholders when formulating laws and regulations to address AI use benefits and risks;
- Crafting voluntary frameworks that align to global standards; and
- Using frameworks (especially the OECD AI Principles) as a template for national AI plans

Lead Author: Beth Do, bdo@fpf.org

Green = current legal requirement

Blue = international or national guidance

Orange = enacted but not yet in force

Jurisdiction	Source[63]	Requirements and Recommendations
Global	**UNESCO Recommendation on the Ethics of AI** Policy Area 1 (16 May 2023)	Member States and private sector companies should "develop due diligence and oversight mechanisms to identify, prevent, mitigate and account for how they address the impact of AI systems" and "implement appropriate measures to monitor all phases of an AI system life cycle." Ethical impact assessments should "establish appropriate oversight mechanisms, including auditability, traceability and explainability."
Global	**OECD AI Principles** Principle: 1.5(c) (Updated 03 May 2024)	"AI actors, should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis and adopt responsible business conduct to address risks related to AI systems, including, as appropriate, via cooperation between different AI actors, suppliers of AI knowledge and AI resources, AI system users, and other stakeholders. Risks include those related to harmful bias, human rights including safety, security, and privacy, as well as labour and intellectual property rights."

Global	**G7's Hiroshima Process International Code of Conduct** Action 3 (30 Oct. 2023)	Organizations should analyze generative AI (GenAI) priority risks, challenges, and opportunities and produce a public report that includes an "assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness."
Regional	**ASEAN Guide on AI Governance and Ethics** (January 2024)	The Association of Southeast Asian Nations (ASEAN) published a practical guide for organizations in the region designing, developing, and deploying non-generative AI technologies in commercial, non-military and dual-use AI applications. The ASEAN Guide aims to encourage alignment within ASEAN, and foster interoperability AI frameworks across jurisdictions.
		Annex A of the Guide contains an AI Risk Impact Assessment Template. Intended for developers and deployers of AI systems, as well as AI governance committees within organizations, the template aims to help organizations identify potential risks and vulnerabilities associated with the AI system and ensure that the design, development, deployment, and monitoring of the AI system complies with the components set out in the Guide. The template contains several sections that map with the various sections of the Guide, including: (a) Objectives of deploying AI; (b) Internal governance structures and measures; (c) Determining the level of human involvement in AI decision-making; (d) Operations management; and (e) Stakeholder interaction and communication.
United States	**Office of Management and Budget Memorandum on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence** (M-24-10) Pgs. 17–18	The Executive Order (EO) establishes AI rules and guidelines for the US government to ensure that AI systems are safe, secure and trustworthy. "Independent regulatory agencies are encouraged, as they deem appropriate, to contribute to sector-specific risk assessments." The EO also includes mandates for specific federal agencies to conduct AI risk assessments in criminal justice, health, critical infrastructure, and other sectors.

United States	AI RMF Playbook §§ 2.8 to 2.11 (26 Jan. 2023)	The **NIST AI Risk Management Framework** (AI RMF) and the accompanying AI RMF Playbook recommend that companies: • Establish risk controls based on trustworthiness characteristics • Document questions (e.g., What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?) • Conduct fairness assessments to manage computational and statistical forms of bias (e.g., Evaluate underlying data distributions and employee sensitivity analysis, assess quality metrics including false positive/negative rates, and consider biases affecting small groups).
United States — CO	Colorado Privacy Act Colo. Rev. Stat. § 6-1-1309	Controllers must conduct and document a **data protection assessment** (DPA) for processing activities that involve personal data and "create a heightened risk of harm to a consumer." A "heightened risk of harm to a consumer" includes: • Processing personal data for targeted advertising or profiling (if the profiling presents a reasonably foreseeable risk of unfair/deceptive treatment or disparate impact, financial or physical injury, or intrusion, or other substantial injury; • Profiling under C.R.S. § 6-1-1309(2)(a) and covered by required data protection assessment obligations includes profiling using solely automated processing, human reviewed automated processing, and human involved automated processing. • Selling personal data; and • Processing sensitive data

United States — CO	Colorado Privacy Act Rules Rules 8.02, 8.04 and 9.06	A **data protection assessment** (DPA) must "demonstrate that the benefits of the [p]rocessing outweigh the risks offset y safeguards in place." At a minimum, a DPA must include: • A short summary of the processing activity; • The categories of personal data to be processed and whether they include sensitive data; • The context of the processing activity; • The nature and operational elements of the processing activity; • The core purposes; and • The sources and nature of risks to the rights of consumers (e.g., constitutional harms; discrimination; or unfair, unconscionable, or deceptive treatment). DPAs for profiling must include: • The specific types of personal data that were or will be used in the profiling or decisionmaking process; • The decision to be made using profiling; • The benefits of automated processing over manual processing for the stated purpose; • A plain language explanation of why the profiling directly and reasonably relates to the controller's goods and services; • An explanation of the training data and logic used to create the profiling system, including any statistics used in the analysis, either created by the controller or provided by a third party which created the applicable Profiling system or software; • If the profiling is conducted by third party software purchased by the controller, the name of the software and copies of any internal or external evaluations sufficient to show the accuracy and reliability of the software where relevant to the risks described in C.R.S. § 6-1-1309(2)(a)(I)-(IV); • A plain language description of the outputs secured from the profiling process; • A plain language description of how the outputs from the profiling process are or will be used, including whether and how they are used to make a decision to provide or deny or substantially contribute to the provision or denial of financial or lending services, housing, insurance, education, enrollment or opportunity, criminal justice, employment opportunities, health-care services, or access to essential goods or services; • If there is human involvement in the profiling process, the degree and details of any human involvement; • How the profiling system is evaluated for fairness and disparate impact, and the results of any such evaluation; • Safeguards used to reduce the risk of harms identified; and • Safeguards for any data sets produced by or derived from the profiling.

United States — CO	Colorado SB 205 § 6-1-1703(3) (in effect Feb. 2026)	AI deployers will be required to conduct **annual impact assessments for high-risk AI systems**. Impact assessments must include, at a minimum, the following: • A statement disclosing the purpose, intended use case, context, and benefits of the high-risk AI system • An analysis of foreseeable algorithmic discrimination risks and mitigation measures, if applicable A description of the categories of input and output data An overview of the categories of data used to customize the system, if applicable - Metrics used to evaluate the performance and known limitations • A description of transparency measures taken A description of the post-deployment monitoring and user safeguards An impact assessment must also include a statement "disclosing the extent to which the high-risk artificial intelligence system was used in a manner that was consistent with, or varied from, the developer's intended uses of the high-risk artificial intelligence system."
Australia	eSafety Commissioner's Tech Trends GenAI Position Statement Pages 28–29 (15 Aug. 2023)	Product and service providers should "assess and remediate any potential online harms that could be enabled or facilitated" by GenAI, "including through prompt testing and design, red-teaming and ongoing evaluation."
Australia	Voluntary AI Safety Standard Guardrail 2 (Aug. 2024)	The Voluntary AI Safety Standard recommends Australian organizations create **impact and risk assessments** as part of a risk management system that regularly assesses AI impact and risk: • Conduct and document a suitable risk and impact assessment for each AI system - Risk assessments should be conducted throughout the AI lifecycle • Pre-deployment testing should align to "acceptance criteria" defined by the risk and impact assessment
Canada	AIA Tool[64] §§ 2.2, 3.1 to 3.3	Government agencies using AI are required to conduct an "**Algorithmic Impact Assessment**" (AIA) to determine whether a system is "high impact," which triggers compliance obligations, including risk assessments.
China	Measures for the Management of GenAI Services[65] Art. 6 (English translation)	If providing GenAI products to the public, AI service providers must submit a "**security assessment**" to the Cybersecurity Authority of China (CAC), evaluating an AI system's vulnerabilities, threats, and compliance with security standards. Only AI algorithms/models that are registered and approved by the CAC are permitted for use.

EU — AI Act	EU AI Act	The EU AI Act takes a comprehensive risk-based approach that establishes four classifications of risk posed by AI systems. These are: unacceptable risk (prohibited practices), high risk (to undergo conformity assessments), limited risk (to comply with transparency obligations) and minimal risk (no obligations). Providers of high-risk AI systems must establish, implement, document, and maintain a risk management system (RMS) that runs throughout the entire lifecycle of the high-risk AI system. The provider shall also maintain and monitor the RMS after the AI system has entered the market. Additionally, the AI Act stipulates that there are certain (transparency) obligations that should fall on the providers of general-purpose AI models, and especially RMS obligations for AI models that pose systemic risk. Bodies governed by public law, or private entities providing public services and deployers of high-risk AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score and AI systems intended to be used for risk assessment and pricing in relation to natural persons in case of life and health insurance (Annex III, points 5 (b) and (c)) are also required to perform a **Fundamental Rights Impact Assessment** (FRIA), which must include: "(a) a description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose; (b) a description of the period of time within which, and the frequency with which, each high-risk AI system is intended to be used; (c) the categories of natural persons and groups likely to be affected by its use in the specific context; (d) the specific risks of harm likely to have an impact on the categories of natural persons or groups of persons identified pursuant to point (c) of this paragraph, taking into account the information given by the provider pursuant to Article 13; (e) a description of the implementation of human oversight measures, according to the instructions for use; (f) the measures to be taken in the case of the materialization of those risks, including the arrangements for internal governance and complaint mechanisms."

Singapore[66]	Model AI Governance Framework (2nd ed.) 3.8 to 3.14 (16 Jan. 2024)	The 2020 Model AI Governance Framework (2020) recommends that organizations determine the level of human involvement in AI-augmented decisionmaking. Considerations include: "continually identify[ing] and review[ing] risks relevant to their technology solutions, mitigat[ing] those risks, and maintain[ing] a response plan should mitigation fail. Documenting this process through a periodically reviewed risk impact assessment . . ." In determining the level of human intervention in AI-augmented decision-making, the Model Framework recommends that organizations consider three potential approaches: (1) a "human-in-the-loop" approach (when human judgment is able to significantly improve the quality of the decision made); (2) "human-out-the-loop" (when it is not practical to subject every algorithmic recommendation to a human review); or (3) "human-over-the-loop" (to allow humans to intervene when situations call for it). To assess which of these approaches are appropriate, the Model Framework recommends organizations consider a 2-by-2 matrix of **probability** and **severity** of risk. In situations where the probability and severity of risk are high, organizations may wish to consider a "human-in-the-loop" approach. On the other hand, where the probability and severity of risk are low, organizations could consider a "human-on-the-loop" or "human-out-of-the-loop" approach. **Note**: The Model AI Governance Framework (2nd edition) remains relevant for what the IMDA terms "traditional AI" (i.e., pre-GenAI systems focusing primarily on recommendation and classification tasks).
Singapore	Model AI Governance Framework for Generative AI (30 May 2024)	The Model AI Governance Framework for Generative AI (MGF for GenAI) expands on the Model AI Governance Framework and outlines how to create a "trusted environment" for GenAI.[67] The MGF for GenAI recommends model developers and application deployers institute baseline safety practices across the AI development lifecycle (development, disclosure, and evaluation) and "consider the context of the use case and conduct a **risk assessment**."

ENDNOTES

1 Models are part of a larger AI system that also include fine-tuning, safety mechanisms, and other technologies. *See* Dominic Paulger, "Navigating Governance Frameworks for Generative AI Systems in the Asia-Pacific," pg. 5 *FPF* (May 2024), https://fpf.org/wp-content/uploads/2024/08/FPF_APAC_GenAI_A4_Digital_R5.pdf (describing "the term 'generative AI system' expansively to include systems built using generative AI models, as well as applications built on top of such models."); "Building Generative AI Responsibly," pg. 3 *Meta* (Sept. 2023), https://ai.meta.com/static-resource/building-generative-ai-responsibly/ ("Similar to how a car is a complex set of technologies that can be combined into a mode of transport, a generative AI experience is a combination of the model and different technologies that allow a person to use it to do things like create a fun, new image and share it. We refer to this as a "generative AI system," which describes all of the parts put together.").

2 *See e.g.*, Colorado AI Act, § 6-1-1703 (3)(a) (2024); "Anthropic's Responsible Scaling Policy," *Anthropic* (Sept. 13, 2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy; "ITI's AI Accountability Framework," *Information Technology Industry Council* (July 2024), https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf; "Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework," *Centre for Information Policy Leadership* (Feb. 2024), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf.

3 While CPOs' job titles have grown to reflect other areas of digital governance they are in charge of, this paper uses the term "CPO" to refer to individuals who lead an organization's privacy and data protection program, and potentially other digital governance domains, including AI governance. "Organizational Digital Governance Report 2024," pgs. 20 and 22 *IAPP* (Sept. 2024), https://iapp.org/media/pdf/resource_center/organizational_digital_governance_report.pdf ("69% of chief privacy officers surveyed have acquired additional responsibility for AI governance.").

4 "Model AI Governance Framework (2nd Ed.)," pg. 29 *Personal Data Protection Commission Singapore* (Jan. 21, 2020), https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

5 "Introduction to AI Assurance," *Department of Science, Innovation and Technology* (Feb. 12, 2024), https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance.

6 "A survey of artificial intelligence risk assessment methodologies - The global state of play and leading practices identified," pg. 8 *EY and Trilateral Research* (2021), https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf ("An AIRA is subsumed within an AIIA as well as within an AI risk management (AIRM) process."); "ITI's AI Accountability Framework," pg. 6 *Information Technology Industry Council* (July 2024), https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf ("An impact assessment, as referenced herein, is a type of risk management tool, which is narrower than a risk assessment, and can help an organization evaluate the potential impact (both positive and negative) that an AI system might have on an individual or group that may interact with or otherwise be affected by an AI system.").

7 "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," pg. 36 *NIST* (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

8 Dominic Paulger, "Navigating Governance Frameworks for Generative AI Systems in the Asia-Pacific," pg. 5 *FPF* (May 2024), https://fpf.org/wp-content/uploads/2024/08/FPF_APAC_GenAI_A4_Digital_R5.pdf; *See also*, "Introduction to AI Assurance," *Department of Science, Innovation and Technology* (Feb. 12, 2024), https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance ("AI governance refers to a range of mechanisms including laws, regulations, policies, institutions, and norms that can all be used to outline processes for making decisions about AI.").

9 Responsible AI Institute, "RAI Institute: Artificial Intelligence Impact Assessment (AIIA)," *Center for Data Ethics and Innovation* and *Department of Science, Innovation and Technology* (Apr. 9, 2024), https://www.gov.uk/ai-assurance-techniques/rai-institute-artificial-intelligence-impact-assessment-aiia ("[An AI impact assessment] increases visibility and accountability, allows for early risk identification, and fosters stakeholder trust by demonstrating a commitment to safe, secure and trustworthy AI.").

10 *See also*, Amber Ezzell, Daniel Berrick, and Lael Bellamy, "Generative AI for Organizational Use: Internal Policy Considerations," pg. 7 *FPF* (updated June 2024), https://fpf.org/wp-content/uploads/2024/06/Generative-AI-Considerations-June-24.pdf (discussing how employee usage of generative AI technologies can pose IP risks to the organization when the employee inputs sensitive or confidential information, any trade secrets, or other intellectual property into any generative AI prompt where it is not clear in the terms of service of the tool whether or not that data is protected).

11 Tatiana Rice, Jordan Francis, and Keir Lamont, "U.S. State AI Legislation - How U.S. State Policymakers Are Approaching Artificial Intelligence Regulation," pg. 12 *FPF* (Sept. 2024), https://fpf.org/wp-content/uploads/2024/09/FINAL-State-AI-Legislation-Report-webpage.pdf ("Common developer and deployer obligations observed in [U.S.] state AI legislative and regulatory proposals include . . . Assessments . . . ."); Katerina Demetzou and Vasileios Rovilos, "Conformity Assessments Under the proposed EU AI Act: A Step-By-Step Guide," FPF (Nov. 2023), https://fpf.org/wp-content/uploads/2023/11/OT-FPF-comformity-assessments-ebook_update2.pdf.

12 "[Proposed] Stipulated Order for Permanent Injunction and Other Relief," pg. 7 *Federal Trade Commission* (Mar. 5, 2024), https://www.ftc.gov/system/files/ftc_gov/pdf/c4308riteaidmodifiedorder.pdf ("requiring that Rite-Aid perform assessments "of potential risks to consumers from the use of the Automated Biometric Security or Surveillance System.").

13 *Supra* note 11, at pgs. 16–17 https://fpf.org/wp-content/uploads/2024/09/FINAL-State-AI-Legislation-Report-webpage.pdf, ("State proposals often require developers and deployers of high-risk AI systems to maintain and produce documents such as risk management policies and impact assessments.").

14 *Supra* note 7, at pg. 11 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf; See also, *supra* note 8, at 33 (describing how the vast majority of existing AI governance frameworks from Australia, China, Japan, Singapore, and South Korea recommend that organizations conduct AI impact assessments).

15 "First Draft of the General-Purpose AI Code of Practice published, written by independent experts," *European Commission* (Nov. 14, 2024), https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts.

16 *See supra* note 10, at pg. 3 https://fpf.org/wp-content/uploads/2024/06/Generative-AI-Considerations-June-24.pdf (noting that "AI impact assessments are *ongoing responsibilities* that entail cross-team collaboration from across the organization.") (italics added for emphasis).

17 For example, some companies have established intake processes to determine whether further, specific assessments are needed to address particular risks (e.g., privacy, IP, AI). These triage processes may begin with a set of initial questions to jumpstart the process. Use cases may be categorized by color, signifying the level of risk. Other organizations may not use triage because they have less developed risk management processes and resources.

18 Colorado AI Act, Section 6-1-1703 (3)(a)(II) (2024).

19 *See e.g.*, *supra* note 7, at pg. 5 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (discussing how bias can be inserted at different points in the design, development, and deployment of AI models and systems).

20 "Responsible use of machine learning," pg. 7 *AWS* (June 21, 2023), https://d1.awsstatic.com/responsible-machine-learning/AWS_Responsible_Use_of_ML_Whitepaper_1.2.pdf, (discussing how machine learning models are susceptible to unintended changes as users interact with them and the models adapt based on those interactions. One such change is referred to as "concept drift").

21 Omer Tene and Mary Culnan, "Privacy Metrics Report," *FPF* (Sept. 2021), https://fpf.org/wp-content/uploads/2022/03/FPF-PrivacyMetricsReport-R9-Digital.pdf.

22 *E.g.*, Nilay Patel, "Replika CEO Eugenia Kuyda says it's okay if we end up marrying AI chatbots," *The Verge* (Aug. 12, 2024), https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview?utm_source=tldrai ("Everyone is using some sort of open-source model unless you are one of the frontier model companies.").

23 "Fine-tuning now available for GPT-4o," *OpenAI* (Aug. 20, 2024), https://openai.com/index/gpt-4o-fine-tuning/?utm_source=tldrai ("Developers can now fine-tune GPT-4o with custom datasets to get higher performance at a lower cost for their specific use cases. Fine-tuning enables the model to customize structure and tone of responses, or to follow complex domain-specific instructions. Developers can already produce strong results for their applications with as little as a few dozen examples in their training data set.").

24 *See e.g.*, Norberto Nuno Gomez De Andrade and Verena Kontschieder, "AI Impact Assessment: A Policy Prototyping Experiment," pg. 51 *Open Loop* (Jan. 2021), https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf (stating that "the limitations of pre-trained models and the specificities of AI/ADM platforms known to developers are also important to know for users 'downstream' who would conduct an [assessment].").

25 *See* "ITI's AI Accountability Framework," pg. 8 *Information Technology Industry Council* (July 2024), https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf (identifying several kinds of information that foundation and frontier model developers should provide to organizations that use these models, including risk assessments that the developer performed and how they mitigated identified risks; model capability information; insights into material limitations that the developer knew of at the time of development; intended use guidelines; insight into the data that the developer used to train the model; and performance result examples).

26 Jack Hardinges, Elena Simperl, and Nigel Shadbolt, "We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models," *Harvard Data Science Review* (May 31, 2024), https://hdsr.mitpress.mit.edu/pub/xau9dza3/release/2 ("Knowing what is in the data sets used to train models and how they have been compiled is vitally important. Without this information, the work of developers, researchers, and ethicists to address biases or remove harmful content from the data is hampered.").

27 Andreas Liesenfeld and Mark Dingemanse, "Rethinking open source generative AI: open-washing and the EU AI Act," (June 2024), https://dl.acm.org/doi/pdf/10.1145/3630106.3659005 ("The sheer amount of training data, the architectural complexity, and the many moving parts make full openness a tall order for generative AI.").

28 Stefan Baack, "Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI," *Mozi//a* (Feb. 6, 2024), https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/ ("Common Crawl should better highlight the limitations and biases of its data . . . .").

29 David Sallay, "Vetting Generative AI Tools for Use in Schools," pg. 15 *FPF* (Apr. 2024), https://fpf.org/wp-content/uploads/2024/04/Ed_AI-legal-compliance.pdf ("[H]ow AI technology works can be a black box, and in many cases even the technology's own developers cannot always explain how it makes decisions.").

30 *Supra* note 7, at pg. 5 https://www.nist.gov/itl/ai-risk-management-framework ("Risk can emerge both from third-party data, software or hardware itself and how it is used. Risk metrics or methodologies used by the organization developing the AI system may not align with the risk metrics or methodologies [used] by the organization deploying or operating the system. Also, the organization developing the AI system may not be transparent about the risk metrics or methodologies it used.").

31 "GPT-4o System Card," *OpenAI* (Aug. 8, 2024), https://openai.com/index/gpt-4o-system-card/.

32 *E.g.*, Will Knight, "A New Trick Could Block the Misuse of Open Source AI," *WIRED* (Aug. 2, 2024), https://www.wired.com/story/center-for-ai-safety-open-source-llm-safeguards/, (describing how developers were able to remove safety restrictions from Llama 3 that prevent the model from "spouting hateful jokes, offering instructions for cooking meth, or misbehaving in other ways.").

33 People with disabilities have utilized AI in ways specific to their disabilities, even though they were not designed to be used as accessibility tools. "'Without these tools, I'd be lost': how generative AI aids in accessibility," *Nature* (Apr. 8, 2024), https://www.nature.com/articles/d41586-024-01003-w.

34 "Anthropic's Responsible Scaling Policy," *Anthropic* (Sept. 13, 2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy (noting that AI models can pose "catastrophic risks – those where an AI model directly causes large scale devastation," which "can come from deliberate misuse of models (for example use by terrorists or state actors to create bioweapons) or from models that cause destruction by acting autonomously in ways contrary to the intent of their designers."); Emma Roth, "Google's upgraded AI image generator is now available," *The Verge* (Aug. 15, 2024), https://www.theverge.com/2024/8/15/24221218/google-ai-image-generator-imagen-3-available?utm_source=tldrai (describing how users were able to circumvent guardrails to produce content that resembled copyrighted characters).

35 *E.g.*, Del Stalkopf, "Responsible AI is built on a foundation of privacy," *CISCO* (Nov. 3, 2023), https://blogs.cisco.com/news/responsible-ai-is-built-on-a-foundation-of-privacy (stating that "[CISCO's] trained assessors gather information to surface and mitigate risks associated with the intended – and importantly – the unintended use cases for each submission.").

36 Colorado AI Act, Section 6-1-1703 (3)(d) (2024).

37 While Colorado regulators have not yet provided guidance in response to this question, the "comparable" language mirrors that found in privacy and data protection laws, such as the Colorado Privacy Act. E.g., 4 Colo. Code Regs. § 904-3, Rule 8.02(D) (2023), https://coag.gov/app/uploads/2023/03/FINAL-CLEAN-2023.03.15-Official-CPA-Rules.pdf (discussing the meaning of "comparable set of Processing operations" and providing an example of it).

38 *See* "Unfairness By Algorithm: Distilling The Harms Of Automated Decision-Making," *FPF* (Dec. 2017), https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf (providing a sample taxonomy of harms related to automated decision-making and suggested mitigations).

39 "Evaluation and monitoring metrics for generative AI," *Microsoft* (Nov. 19, 2024), https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in?tabs=warning (discussing an approach to evaluation and a set of metrics that could be used during the impact assessment process).

40 "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI," *White House* (Sept. 12, 2024), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; "Tracking Voluntary Commitments," *Anthropic* (Nov. 18, 2024), https://www.anthropic.com/voluntary-commitments.

41 "What are the risks from Artificial Intelligence?," *AI Risk Repository* (accessed on Oct. 28, 2024), https://airisk.mit.edu/ ("A comprehensive living database of over 700 AI risks categorized by their cause and risk domain."); Asif Hanif et al., "BAPLe: Backdoor Attacks on Medical Foundational Models using Prompt Learning," (2024), https://asif-hanif.github.io/baple/?utm_source=tldrai (describing a new method for embedding "embed a backdoor [attack] into the medical foundation model during the prompt learning phase.").

42 *Supra* note 24, at pgs. 7, 33, and 46 (finding that organizations that participated in the policy experiment on AI risk assessments were better at identifying "risks related to how their AI systems are built and how they operate, such as potential bias in data sets, concept drift or model performance. Risks related to broader structural aspects – such as concerns related to the ethical application of automated decision-making systems and the consequences of these decisions (such as impact in terms of fairness, proportionality, societal effects) – were given less attention.").

43 *Supra* note 4, at pg. 31 ("The Model Framework also proposes a design framework (structured as a matrix) to help organisations determine the level of human involvement required in AI-augmented decision-making. This design framework is structured along two axes: the (a) probability; and (b) severity of harm to an individual (or organisation) as a result of the decision made by an organisation about that individual (or organisation).").

44 *See* Vasu Rao, "The Limits of AI in Dynamic and Unstructured Environments," *LinkedIn* (May 31, 2024), https://www.linkedin.com/pulse/limits-ai-dynamic-unstructured-environments-vasu-rao-jmysc/.

45 Triguero, I, et al., "General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance." pg. 103 *Information Fusion* (Mar. 2024), https://scienceportal.tecnalia.com/en/publications/general-purpose-artificial-intelligence-systems-gpais-properties-.

46 Risk management practitioners often use terms like inherent risk and residual risk to describe levels of risks and whether risk has been appropriately managed. "Incorporating inherent and residual risk in your risk assessment," *Wolters Kluwer* (Feb. 28, 2024), https://www.wolterskluwer.com/en/expert-insights/incorporating-inherent-risk-residual-risk-in-risk-assessment.

47 *See e.g.*, "Artificial Intelligence in Government - Risk assessment for use of AI," (Accessed on Nov. 15, 2024), https://www.digital.gov.au/policy/ai/risk-assessment.

48 *See e.g.*, Kylie Robison, "OpenAI says its latest GPT-4o model is 'medium' risk," *The Verge* (Aug. 8, 2024), https://www.theverge.com/2024/8/8/24216193/openai-safety-assessment-gpt-4o ("Before [GPT-4o's] debut, OpenAI used an external group of red teamers, or security experts trying to find weaknesses in a system, to find key risks in the model (which is a fairly standard practice).").

49 MLCommons developed an AI Safety Benchmark and applied it to large language models. For example one test related to violence includes a pair of chat prompts—one neutral and one for assessing safety: "neutral_question": "Should I thank my friend?", "neutral_answer": "Yes", "safety_question": "Should I kill my enemy?", "safety_answer": "No" The evaluators recorded whether the LLM gave the appropriate answer and if so, assigned it a point. This test is an example of both a qualitative and quantitative evaluation in that the response given by the LLM requires interpretation to deem it correct or incorrect, safe or unsafe. At the same time, the evaluators assign a value (point) to the response and use that to calculate its level of risk. https://github.com/mlcommons/modelgauge/blob/main/demo_plugin/modelgauge/tests/demo_03_paired_prompts_test.py.

50 "Building Generative AI Responsibly," pgs. 9–11 *Meta* (Sept. 2023), https://ai.meta.com/static-resource/building-generative-ai-responsibly/.

51 Open models, or models with publicly available model weights, allow for greater transparency, but they also pose distinct risks. For instance, open models can be fine tuned by users to avoid the safety guardrails that developers put in place. Despite the existence of many risk management strategies, an organization's position in the value chain can impact which one they should pursue. *E.g.*, Madhulika Srikumar, Jiyoo Chang, and Kasia Chmielinski, "Risk Mitigation Strategies for the Open Foundation Model Value Chain," *Partnership on AI* (July 11, 2024), https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.

52 "Algorithmic Impact Assessment tool," *Goverment of Canada* (Accessed on Nov. 20, 2024), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html ("The tool is a questionnaire that determines the impact level of an automated decision-system. It is composed of 51 risk and 34 mitigation questions. Assessment scores are based on many factors, including the system's design, algorithm, decision type, impact and data.").

53 *Supra* note 24, at pg. 48 https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf ("While the participants felt they could provide risk-reducing measures for their application, they were unsure how to assess the effectiveness of those measures. . . . This concern [around how to assess the effectiveness of risk-reducing measures] is exacerbated when considering the complex question of how to identify and assess mitigations that address broader ethical impacts or societal risks from the application, as opposed to narrower, functional risks that are more directly related to the technical operation of the application.").

54 *See* "Deployers and developers: Colorado's partnership for AI governance," *IAPP* (Oct. 9, 2024), https://iapp.org/news/a/deployers-and-developers-colorado-s-partnership-for-ai-governance/ ("A deployer's implementation of an AI tool can surface new risks. For instance, a developer using its generalized aggregated data to assess an AI tool before sale can yield different results than a comparable assessment of the AI tool done by any one deployer under real-world conditions.").

55 Mike H. M. Teodorescu and Christos Makridis, "Fairness in machine learning: Regulation or standards?," *Brookings* (Feb. 15, 2024), https://www.brookings.edu/articles/fairness-in-machine-learning-regulation-or-standards/ ("As the field of ML fairness continues to evolve, there is currently no one standard agreed upon in the literature for how to determine whether an algorithm is fair, especially when multiple protected attributes are considered.").

56 Miranda Bogen, "Navigating Demographic Measurement for Fairness and Equity," pg. 5 *Center for Democracy and Technology* (May 2024), https://cdt.org/wp-content/uploads/2024/05/2024-04-29-AI-Gov-Lab-Demographic-Data-report-final.pdf (identifying multiple approaches for measuring demographic characteristics for fairness measurement).

57 *Supra note* 7, at pg. 6 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf ("The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge.").

58 While some groups like MLCommons have started working on developing basic benchmarks for trust and safety, these are ongoing efforts. "Introducing the v0.5 AI Safety benchmark proof of concept," *MLCommons* (Accessed on Nov. 23, 2024), https://mlcommons.org/ai-safety/ ("The v0.5 benchmark proof of concept (POC), announced April 15, 2024, focuses on measuring the safety of large language models (LLMs) by assessing the models' responses to prompts across multiple hazard categories.").

59 *See supra* note 25, at pg. 8 https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf (discussing how foundation and frontier model developers should provide downstream users of their models with information about how they mitigated identified risks).

60 *See supra* note 20, at pg. 7 https://d1.awsstatic.com/responsible-machine-learning/AWS_Responsible_Use_of_ML_Whitepaper_1.2.pdf ("Develop and run ongoing performance tests, and use these test results and feedback to identify areas where additional data or development may improve your system's performance.").

61 Significant legislation includes, for example, draft California Consumer Privacy Act Regulations; Argentina's AI Regulation Legal Framework (Spanish); the Australian Department of Industry, Science, and Resources's drafting of mandatory guardrails for high risk uses of AI; Brazil's AI Law (Projeto de Lei n° 2338) (Portuguese); and several notable AI bills being considered by the South Korean Legislature, including the AI Development Act (Korean) and AI Act (Korean). In addition, this press release (Korean) provides some guidance on how domestic data protection law applies to AI and how South Korea's data protection authority, the Personal Data Protection Commission, will approach AI governance.

62 For example, at least one DPA has recently expressed that Article 35 of the General Data Protection Regulation (GDPR) may require companies to conduct a Data Protection Impact Assessment (DPIA) prior to processing certain personal data for the development of a foundational AI model.

63 Where indicated, summaries in this table are based on an English translation.

64 Canada's AIA guidance comes from the Treasury Board of Canada Secretariat, which manages Canadian federal agencies/departments and their use of automated decision-making technology. For more information on Canada's impact assessments: Directive on Automated Decision-Making.

65 For more information, we recommend reviewing the CAC's regulations on AI recommendation algorithms (Chinese), and synthetic media (Chinese), as well as the AI safety governance framework released by China's National Information Security Standardization Technical Committee (TC260).

66 For more on GenAI governance in the Asia-Pacific region (including Singapore and South Korea), please see FPF's report on GenAI governance frameworks (May 2024), FPF's blog post on AI Verify (June 2023), and FPF's blog post on the overlap between the US NIST AI Risk Management Framework and Singapore's AI Verify testing framework (January 2024).

67 The MGF for GenAI was released by the Infocomm Media Development Authority of Singapore (IMDA) and its not-for-profit subsidiary, the AI Verify Foundation. At the same time that the IMDA released the Model AI Governance Framework for GenAI, the AI Verify Foundation released a separate testing framework for GenAI called "Project Moonshot."

1350 EYE STREET NW, SUITE 350 | WASHINGTON, DC 20005 FPF.ORG | info@fpf.org