iapp | FTI CONSULTING | TECHNOLOGY

AI Governance in Practice Report 2024

Table of contents

What's inside?

Executive summary

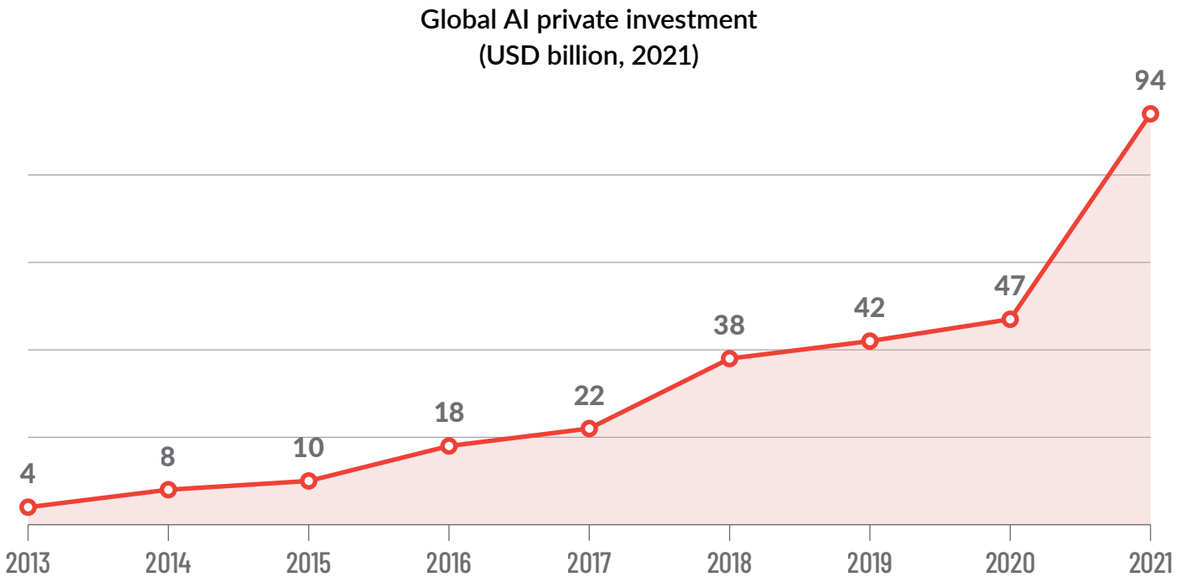
Recent and rapidly advancing breakthroughs in machine learning technology have forever transformed the landscape of AI.

AI systems have become powerful engines capable of autonomous learning across vast swaths of information and generating entirely new data. As a result, society is in the midst of significant disruption with the surge in AI sophistication and the emergence of a new era of technological innovation.

As businesses grapple with a future in which the boundaries of AI only continue to expand, their leaders face the responsibility of managing the various risks and harms of AI, so its benefits can be realized in a safe and responsible manner.

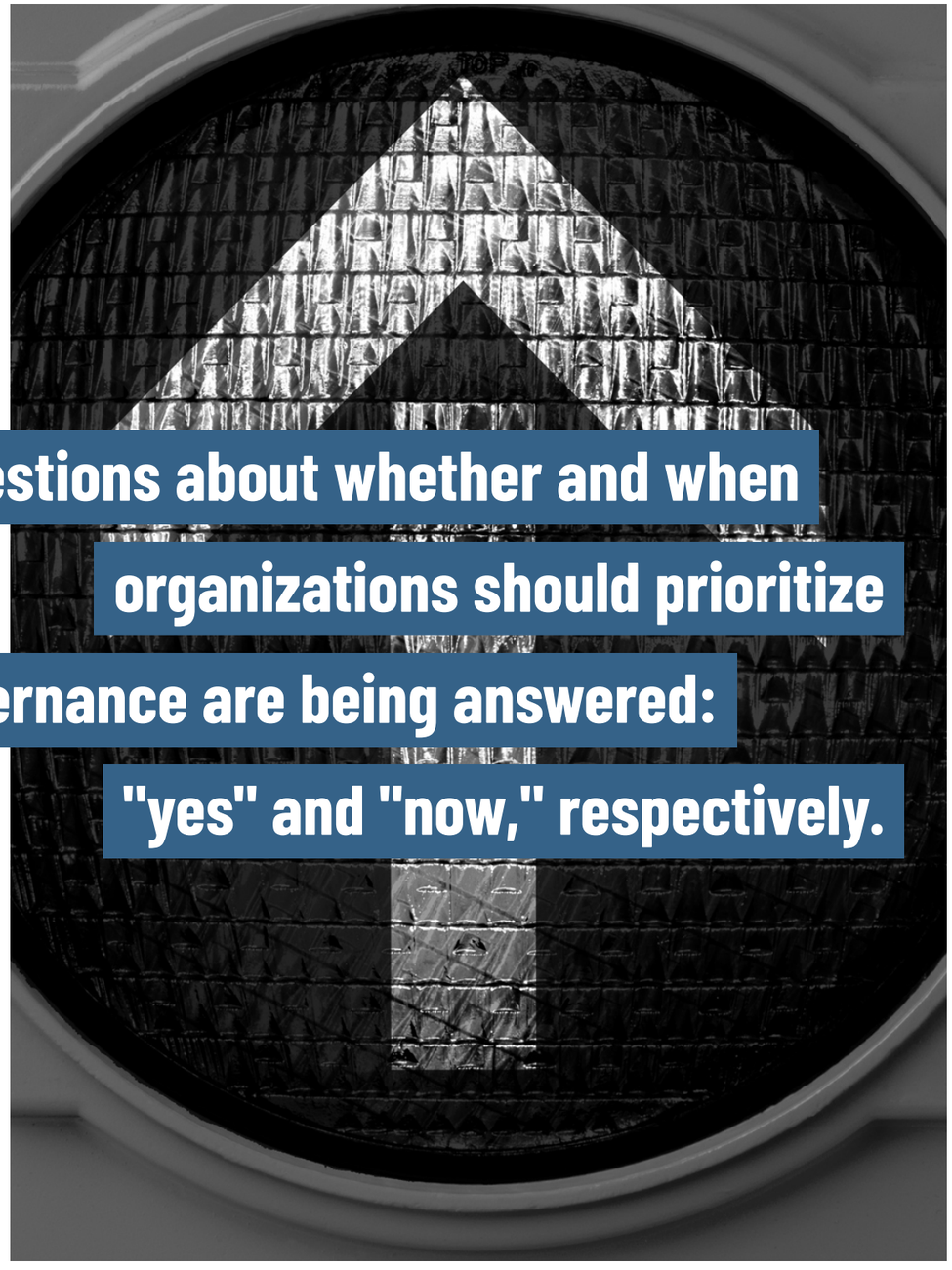
Critically, these benefits are accompanied by serious considerations and concerns about the safety of this technology and the potential for it to disrupt the world and negatively impact individuals when left unchecked. Confusion about how the technology works, the introduction and proliferation of bias in algorithms, dissemination of misinformation, and privacy rights violations represent only a sliver of the potential risks.

The practice of AI governance is designed to tackle these issues. It encompasses the growing combination of principles, laws, policies, processes, standards, frameworks, industry best practices and other tools incorporated across the design, development, deployment and use of AI.



While relatively new, the field of AI governance is maturing, with government authorities around the world beginning to develop targeted regulatory requirements and governance experts supporting the creation of accepted principles, such as the Organisation for Economic Co-Operation and Development's AI Principles, emerging best practices and tools for various uses of AI in different domains.

There are many challenges and potential solutions for AI governance, each with unique proximity and significance based on an organization's role, footprint, broader risk-governance profile and maturity. This report aims to inform the growing, increasingly empowered and increasingly important community of AI governance professionals about the most common and significant challenges to be aware of when building and maturing an AI governance program. It offers actionable, real-world insights into applicable law and policy, a variety of governance approaches, and tools used to manage risk. Indeed, some of the challenges to AI governance overlap and run through a range of themes. Therefore, an emerging solution for one thematic challenge may also be leveraged for another. Conversely, in certain circumstances, specific challenges and associated solutions may conflict and require reconciliation with other approaches. Some of these potential overlaps and conflicts have been identified throughout the report.



Questions about whether and when organizations should prioritize AI governance are being answered: "yes" and "now," respectively. This report is, therefore, focused on how organizations can approach, build and leverage AI governance in the context of the increasingly voluminous and complex applicable landscape.

Uzma Chaudhry
IAPP AI Governance Center Research Fellow

Joe Jones
IAPP Director of Research and Insights

Ashley Casovan
IAPP AI Governance Center Managing Director

Nina Bryant
FTI Technology Senior Managing Director

Luisa Resmerita
FTI Technology Senior Director

Michael Spadea
FTI Technology Senior Managing Director

Part I. Understanding AI and governance

Components of an AI system and their governance

To understand how to govern an AI system, it is important to first understand what an AI system is. The EU AI Act, for example, defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

As indicated in the OECD's Framework for the Classification of AI systems, AI systems are comprised of data used to train and operate a system, model, output and context. While a model is a fundamental building block of an AI system, a single model seldom operates in isolation. Instead, multiple AI models come together and interact with each other to form complex AI systems. Additionally, AI systems are often designed to interact with other systems for sharing data, facilitating seamless integration into real-world environments. This results in a network of AI systems, each with its specialized models, working together to achieve a larger goal.

> **With AI poised to revolutionise many aspects of our lives, fresh cooperative governance approaches are essential. Effective collaboration between regulatory portfolios, within nations as well as across borders, is crucial: both to safeguard people from harm and to foster innovation and growth.**
>
> Kate Jones
> U.K. Digital Regulation Cooperation Forum CEO

Navigating AI governance sources

Given the complexity and transformative nature of AI, significant work has been done by law and policymakers on what is now a vast and growing body of principles, laws, policies, frameworks, declarations, voluntary commitments, standards and emerging best practices that can be challenging to navigate. Many of these various sources interact with each other, either directly or by virtue of the issues covered.

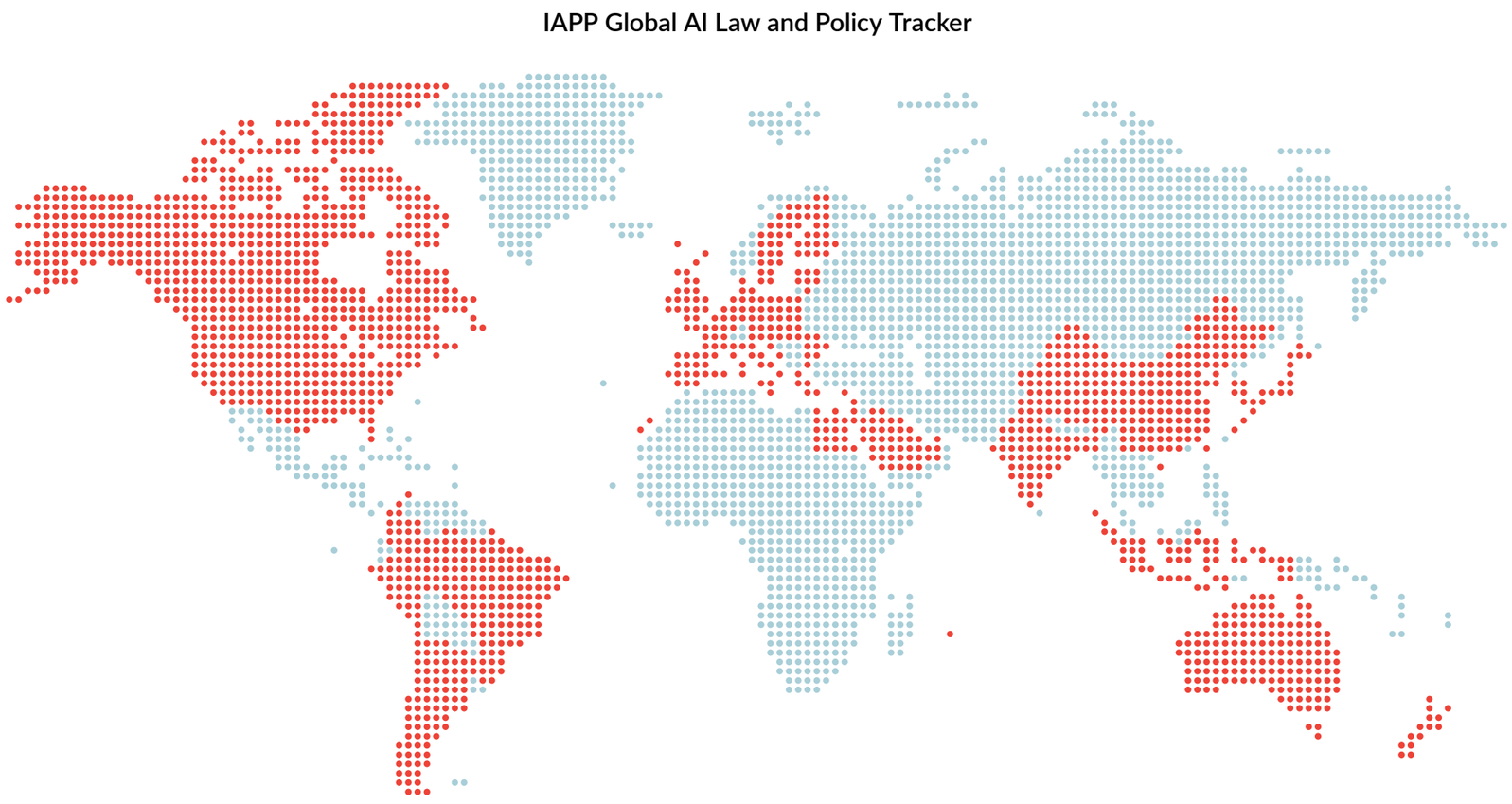
AI principles, such as the OECD's AI Principles or UNESCO's Recommendation on the Ethics of AI, can shape global standards, especially when national governments pledge to voluntarily incorporate such guidance into their domestic AI governance initiatives. They provide a nonbinding, principled approach to guide legal, policy and industry efforts toward tackling thematic challenges. Algorithm Watch created an inventory of these principles, identifying 167 reports.

Laws and regulations include existing legislation that is not specific but is nonetheless applicable to AI, as well as emerging legislation that more specifically addresses the governance of AI systems, such as the EU AI Act. The EU AI Act is the world's first comprehensive AI regulation. Although jurisdictional variations can be observed across the emerging global AI regulatory landscape, many draft regulations adopt a risk-based approach similar to the EU AI Act.

The EU AI Act mandates AI governance standards based on the risk classification of AI systems and the organization's role as an AI actor. Certain AI systems are deemed to pose unacceptable risk and are prohibited by law, subject to very narrow exceptions. The bulk of the requirements imposed by the act apply to providers of high-risk AI systems, although deployers and resellers, namely distributers and importers, are are also subject to direct obligations.

The act imposes regulatory obligations at enterprise, product and operational levels, such as establishing appropriate accountability structures, assessing system impact, providing technical documentation, establishing risk management protocols and monitoring performance, among other key requirements. In the context of the growing variety of generative AI use cases and adoption of solutions embedding generative AI such as MS Copilot, general purpose AI-specific provisions are another crucial component of the EU AI Act. Depending on their capabilities, reach and computing power, certain GPAI systems are considered to present systemic risk and attract broadly similar obligations to those applicable to high-risk AI systems.

> **“**
>
> **AI governance is about to get a lot harder. The internal complexity of governing AI is growing as more internal teams adopt AI, new AI features are built, and the systems get complex, but at the same time, the external complexity is also set to grow rapidly with new regulations, customer demands, and safety research evolving.**
>
> **The organizations who have invested in structured AI governance already have a leg up and will continue to have a competitive advantage.**
>
> Andrew Gamino-Cheong
> Trustible AI Co-founder and Chief Technology Officer



This map shows the jurisdictions in focus and covered by the IAPP Global AI Law and Policy Tracker. It does not represent the extent to which jurisdictions around the world are active on AI governance legislation. Tracker last updated January 2024.

In addition to binding legislation, voluntary AI frameworks, such as the National Institute of Standards and Technology's AI Risk Management Framework and the International Organization for Standardization's AI Standards, offer structured and actionable guidance stakeholders can elect to use to support their work on implementing AI governance. Voluntary commitments are often developed to bring different stakeholders closer to a shared understanding of identifying, assessing and managing risks. Standards serve as benchmarks that can demonstrate compliance with regulatory requirements.

International declarations and commitments memorialize shared commitments, often between governments, to specific aspects or broad swathes of AI governance. While not binding, such commitments can, at a minimum, indicate a country's support for and intention to advance AI governance in particular or general ways, even at the highest of levels.

Navigating a growing body of draft AI laws, regulations, standards and frameworks can be challenging for organizations pioneering with AI. By understanding their unique AI risk profile and adopting a risk-based approach, organizations can build a robust and scalable AI governance framework that can be deployed across jurisdictions.

The following are examples of some of the most prominent and consequential AI governance efforts:

Principles	→ OECD AI Principles → European Commission's Ethics Guidelines for Trustworthy AI → UNESCO Recommendation on the Ethics of AI → The White House Blueprint for an AI Bill of Rights → G7 Hiroshima Principles
Laws and regulations	→ EU AI Act → EU Product Liability Directive, proposed → EU General Data Protection Regulation → Canada – AI and Data Act, proposed → U.S. AI Executive Order 14110 → Sectoral U.S. legislation for employment, housing and consumer finance → U.S. state laws, such as Colorado AI Act, Senate Bill 24-205 → China's Interim Measures for the Management of Generative AI Services → The United Arab Emirates Amendment to Regulation 10 to include new rules on Processing Personal Data through Autonomous and Semi-autonomous Systems → Digital India Act
AI frameworks	→ OECD Framework for the classification of AI Systems → NIST AI RMF → NIST Special Publication 1270: Towards a Standard for Identifying and Managing Bias in AI → Singapore AI Verify → The Council of Europe's Human Rights, Democracy, and the Rule of Law Assurance Framework for AI systems
Declarations and voluntary commitments	→ Bletchley Declaration → The Biden-Harris Administration's voluntary commitments from leading AI companies → Canada's guide on the use of generative AI
Standards efforts	→ ISO/IEC JTC 1 SC 42 → The Institute of Electrical and Electronics Engineers Standards Association P7000 → The European Committee for Electrotechnical Standardization AI standards for EU AI Act → The VDE Association's AI Quality and Testing Hub → The British Standards Institution and Alan Turing Institute AI Standards Hub → Canada's AI and Data Standards Collaborative

> **It is important to take an ecosystem approach to AI governance. Policy makers and industry need to work together at platforms such as the AI Verify Foundation to make sense of the opportunities and risks that this technology brings. The aim is to find common guardrails to manage key risks in order to create a trusted ecosystem that promotes maximal innovation.**
>
> **Denise Wong**
> Singapore Infocomm Media Development Authority Assistant Chief Executive, Data Innovation & Protection Group

AI Risks

The AI governance imperative

With private investment, global adoption rates and regulatory activity on the rise, as well as the growing maturity of the technology, AI is increasingly becoming a strategic priority for organizations and governments worldwide. Organizations of all sizes and industries are increasingly engaging with AI systems at various stages of the technology product supply chain.

The exceptional dependence on high volumes of data and endless practical applicability that make AI technology a disruptive opportunity can also generate uniquely multifaceted risks for businesses and individuals. These include legal, regulatory, reputational and/or financial risks to organizations, but also risks to individuals and the wider society.

Enterprise governance

AI governance starts with defining the corporate strategy for AI by documenting:

- → Target operating models to set out clear roles and responsibilities for AI risk.
- → Compliance assessments to establish program maturity and remediation priorities.
- → Accountability processes to record and demonstrate compliance.
- → Policies and procedures to formulate policy standards and operational procedures.
- → Horizon scanning to enhance and align the program with ongoing regulatory developments.

Product governance

AI governance also requires enterprise policy standards to be applied at the product level. Organizations can ensure their AI products match their enterprise strategy by using:

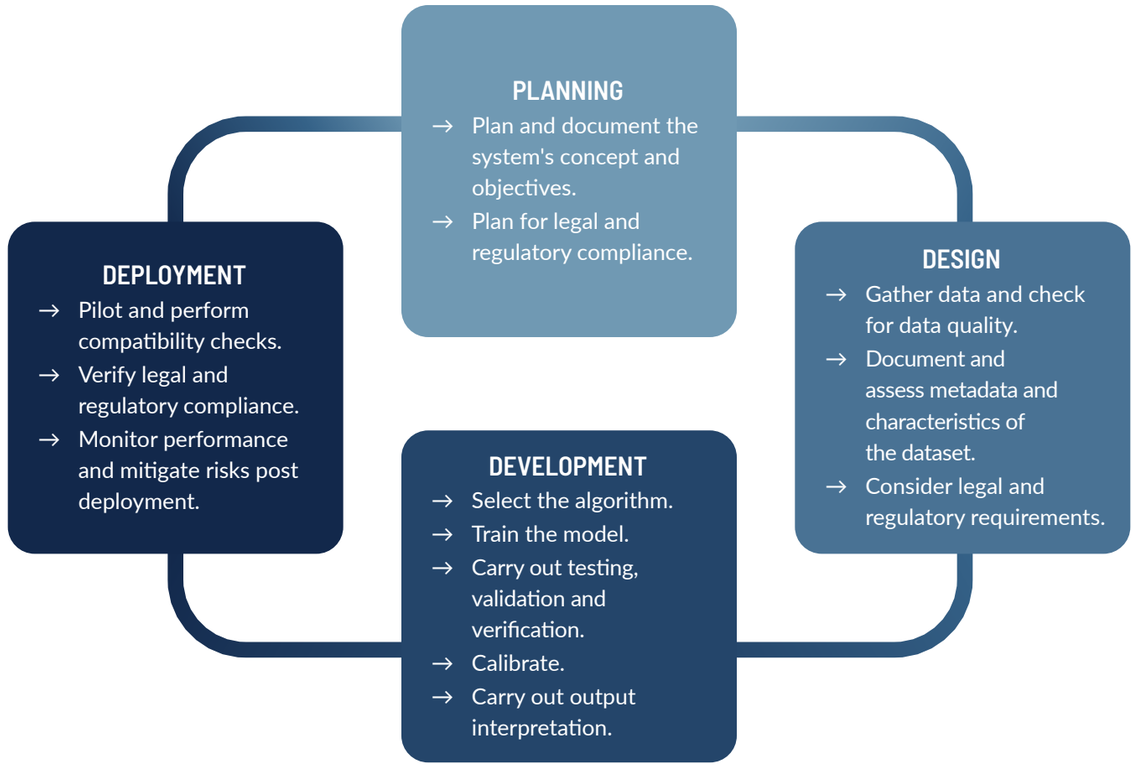
- → System impact assessments to identify and address risk prior to product development or deployment.
- → Quality management procedures tailored to the software development life cycle to address risk by design.
- → Risk and controls frameworks to define AI risk and treatment based on widely recognised standards such as ISO and NIST.
- → Conformity assessments and declarations to demonstrate their products are compliant.
- → Technical documentation including standardized instructions of use and technical product specifications.
- → Post-market monitoring plans to monitor product compliance following market launch.
- → Third-party due diligence assessments to identify possible external risk and inform selection.

Operational governance

The organization's AI strategy must ultimately be operationalized throughout the business through the development of:

- → Performance monitoring protocols to ensure systems perform adequately for their intended purposes.
- → Transparency and human oversight initiatives to ensure individuals are aware and can make informed choices when they interact with AI systems or when AI-powered decisions are made.
- → Incident management plans to identify, escalate and respond to serious incidents, malfunctions and national risks impacting AI systems and their operation.
- → Communication strategies to ensure transparency toward internal and external stakeholders in relation to the organization's AI practices.
- → Training and awareness programs to enable staff with roles and responsibilities for AI governance to help them understand and perform their respective roles.
- → Skills and capabilities development to assess human resources capabilities and review or design job requirements.

The AI life cycle



Understanding that AI systems, like all products, follow a life cycle is important as there are governance considerations across the life cycle. The NIST AI RMF sets out a comprehensive articulation of the AI system life cycle and includes considerations for testing, evaluation, validation, verification and key stakeholders for each phase. A more simplified sample life cycle is included above, along with some top-level considerations.

> **An effective AI governance model is about collective responsibility and collective business responsibility, which should encompass oversight mechanisms such as privacy, accountability, compliance, among others. This responsibility should be shared by every stakeholder who is part of the AI governance chain.**
>
> **Vishal Parmar**
> British Airways Global Lead Privacy Counsel and Data Protection Officer

→ HOW TO
Navigate developers from deployers

Various contractual and regulatory obligations may arise depending on whether an organization is a vendor or buyer, or if it sources external services such as hardware, cloud or data collection, for the development and operations of its AI system.

Prior IAPP research found more than 70% of organizations rely at least somewhat on third-party AI, so the responsibility for ensuring the AI system is safe and responsible may be spread across multiple roles.

In both current legislation and proposed legislation we are starting to see different obligations for those who provide and supply AI versus those who deploy AI. Understanding whether you are a developer and/or deployer is important to ensuring you meet compliance obligations. Once this is understood, it is possible to establish AI-governance processes for procurement, including evaluations and contracts to avoid taking on additional liabilities.

→ The World Economic Forum put together a useful toolkit to help those who are procuring AI systems.

Part II. The data challenge

Data is an integral part of training and operating an AI system.

Most AI requires sizeable amounts of high-quality data, especially during the training phase to maximize the model's performance, as well as to ensure the desired and accurate output. With the advancement of new AI technologies, models are requiring increasingly more data, which may come from a variety of sources. Given the importance of the data used to fuel the AI system, it is important to understand what data is being used; how, where and by whom it was collected; from whom it was collected; if it is the right data for the desired outcome; and how it will be managed throughout the life cycle.

Accessing data and identifying data sources

Understanding where data comes from and how it is collected is not only necessary for AI systems, but also for building trust in AI by ensuring the lawfulness of data collection and processing. Such documentation can assist with data transparency and improve the AI system's auditability as well.

Although data may originate from multiple sources, it can be broadly categorized into three types: first-party data, public data and third-party data.

First-party data

This refers to data collected directly from individuals by an organization through their own interactions and transactions. Such data may originate from sources such as website visits, customer feedback and surveys, subscriptions, and customer relationship management systems, among others. This data is extremely valuable for organizations as it provides direct and firsthand insights into individuals' behavior.

First-party data can be collected from various sources. Identifying the data channels and documenting the source will not only help the organization determine what types of data, e.g., text, numerical, image or audio, will be collected from each source, but also alert the legal team about where legal compliance will be required on the organization's part.

Public data

This refers to data that is available to the wider public and encompasses a range of sources, such as publicly available government records, publications, and open source and web-scraped data. Public data is a valuable resource for researchers and innovators as it provides readily available information. Public data can come from multiple sources.

While it is arduous and cumbersome to maintain data lineage for public datasets, it is important for upholding organizational reputation and fostering user trust, legal compliance and AI safety overall. A lack of understanding of where data comes from eventually leads to a lack of understanding of the training dataset and model performance, which can reinforce the black-box problem. Therefore, in the interest of transparency, tracking and documenting public-data sources as much as possible may prove beneficial for the organization, as it can later support other transparency efforts, such as drawing up data, model or system cards.

Moreover, without knowledge of public-data sources, the organization may inadvertently train the AI system on personal, sensitive or proprietary data. From the privacy standpoint, this can be problematic in cases of data leakage, where personally identifiable data may be exposed. AI security challenges may also be amplified if data was procured from unsafe public sources, as that carries the risk of introducing malicious bugs into the system. It may also lead to biases in the AI system.

> **Ethical development practices start with responsible data acquisition and management systems, as well as review processes that track the lineage of sourced data.**
>
> Christina Montgomery
> IBM Vice President and
> Chief Privacy and Trust Officer

> **Without understanding the quality of the data being ingested into an AI model, you may not know the quality of the output. Companies must establish and define what 'data quality' involves and consists of, as this determination is highly contextual for any organization, and can depend on business goals, use cases, focus areas and fitness for purpose.**
>
> **Regardless of context, there are minimum baseline attributes which can and should be established: accuracy, completeness, consistency and validity. Timeliness and uniqueness may also be important to establishing fitness for purpose.**
>
> **Dera Nevin**
> FTI Technology Managing Director

An organization can begin by establishing a clear understanding of how and why public data is being collected, how it aligns with the purposes the AI system will fulfil, if and how system accuracy will be affected by using public data, what the trustworthy sources for gathering public data are, if the organization has rights to use the public data, and other legal considerations that may have to be taken into account, particularly given that public data is treated differently across jurisdictions.

Third-party data

This refers to data obtained or licensed by the organization from external entities that collect and sell data, such as data brokers. Datasets purchased from brokers are webbed together from a wide range of sources. While this may have the benefit of providing insights into a wider user base, the insights may not be accurate or may be missing key data. It may lack direct insights into customer behavior, as brokers do not interact with the organization's customer base.

Third-party data can also include open-source data, available through open-source data catalogues. Sometimes these databases are provided by government or academic institutions with a clear understanding of how the data was collected and how it can be used, including a clear use license. Open-source data collected through other community efforts may not follow the same collection and distribution practices. As when using all data, it is important to know where the data came from, how it was collected, in which context it is meant to be used and what rights you have to use it.

Data quality

The quality of data that AI is trained and tested on directly impacts the quality of the outputs and performance, so ensuring the data is high quality can help lay the initial foundations for a safe and responsible AI system. Measuring data quality often includes a few baseline considerations.

Accuracy confirms the correctness of data. That is, whether the data collected is based on real-world insights. Completeness refers to checking for missing values, determining the usability of the data, and looking for any over or underrepresentation in the data sample. Validity ensures data is in a format that is compatible with intended use. This may include valid data types, metadata, ranges and patterns. Consistency refers to the relationships between data from multiple sources and includes checking if the data shows consistent trends and values it represents. Ideally, this process of ensuring data quality is documented to support transparency, explainability, data fairness, auditability, understanding of the data phase of the life cycle and system performance.

Appropriate use

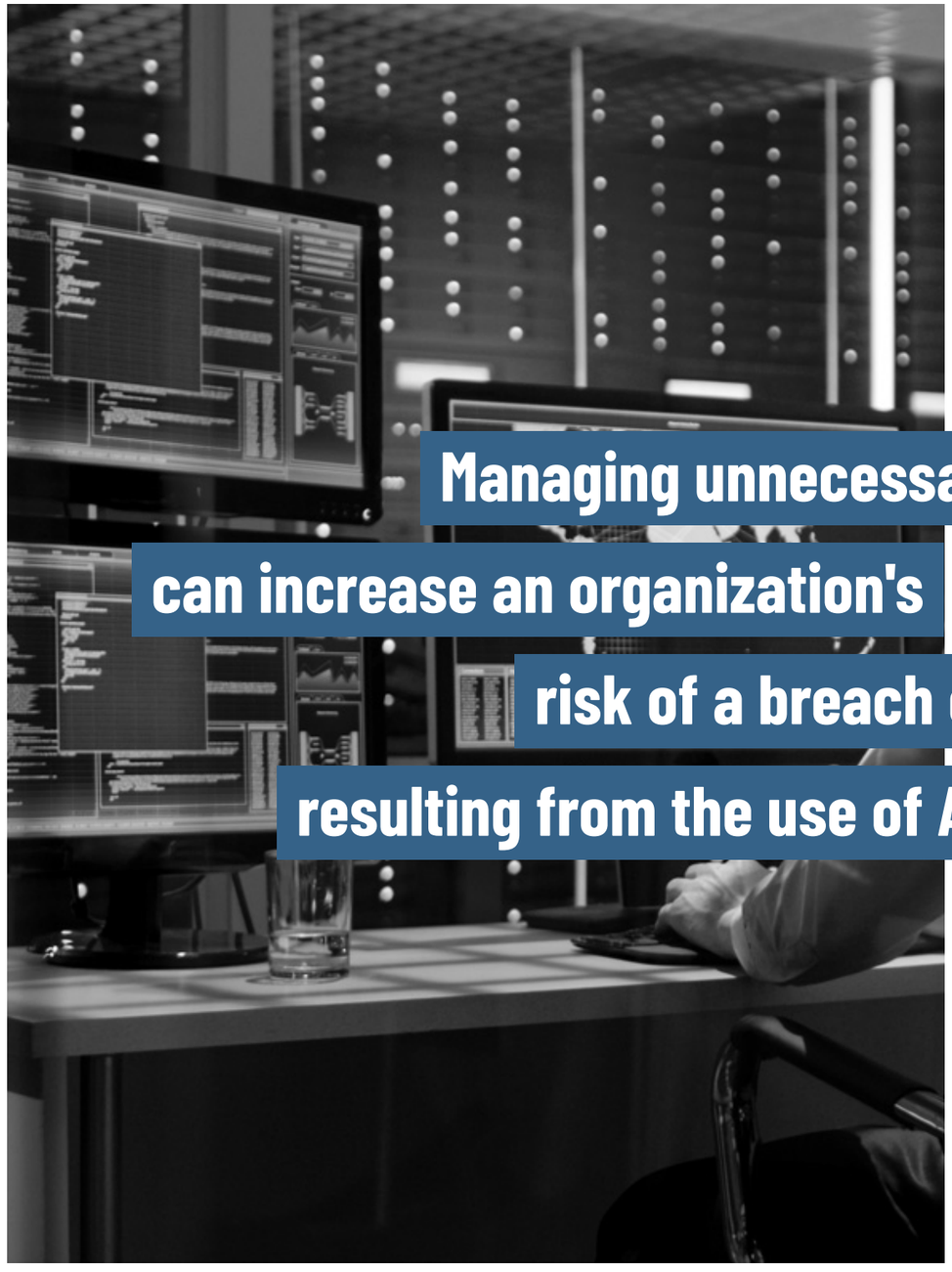
One of the most significant challenges when designing and developing AI systems is ensuring the data used is appropriate for the intended purpose. Often data is collected with one intention in mind or within a specific demographic area, and, while it might appear to be a useful dataset, upon further analysis it might include data that does not match the industry or geographic area of operation. When data is not fit for purpose, it can skew the AI system's predictions or outcomes.

When thinking about appropriate use, consider the proportionality of data required for the desired outcome. Often, there are occurrences of collecting or acquiring more data than necessary to achieve the outcome. It is important to understand if it is even necessary to collect and use certain data in your AI system.

Managing unnecessary data, especially data that may contain sensitive attributes, can increase an organization's risk of a breach or harm resulting from the use of AI.

Law and policy considerations

Approaches can be categorized according to how the data was collected.



First-party data

Where first-party data amounts to personal or sensitive data, relevant provisions may be triggered under the data protection and privacy legislation of the jurisdictions where the organization carries out its business, where the processing takes place or where the individuals concerned are located.

The EU General Data Protection Regulation, for instance, has a default prohibition against processing of personal data, unless such processing falls under one of the six bases for lawful processing under Article 6(1): consent, contractual performance, vital interest, legal obligation, public task and legitimate interest pursued by a controller or third party.

Public data

Web scraping may involve compliance with the terms of service and privacy policies of websites. Otherwise, when an organization is aware the public dataset contains personal or sensitive information, lawfulness of use may require compliance with relevant data protection or privacy laws, such as by acquiring valid consent.

While web scraping, it is possible for copyrighted data to be collected to train AI systems.

Another type of public data is open-source data, which is publicly available software that may include both code and datasets. Although accessible to the public, open-source software is often made available by the organization through various open-source licensing schema. In addition to complying with the terms of the licenses, organizations using open-source data may also consider conducting their own due diligence to ensure the datasets were acquired lawfully, are safe to use and were assessed for bias mitigation.

Third-party data

As organizations have neither proximity to how third-party data was first collected nor direct control over the data governance practices of third parties, an organization can benefit from carrying out its own legal due diligence and third-party risk management. The extent and intensity of this exercise will largely depend on the organization's broader governance and risk-management approach and the relevant facts.

Legal due diligence may include verification of the personal data's lawful collection by the data broker, review of contractual obligations and licenses, and identification of protected intellectual property interests. When data is licensed, the organization will first have to lawfully procure rights to use data through a licensing agreement. This will help maintain data provenance and a clear understanding of data ownership. The lawful and informed use of such data at subsequent stages of the AI life cycle will also be governed by the license.

> **With growing public concerns and increased regulation aimed at developing trustworthy, transparent and performative AI systems, an internal data governance program is integral to understanding and documenting metadata prior to usage, and to identifying risks associated with lawful data use.**
>
> **Christina Montgomery**
> IBM Vice President and
> Chief Privacy and Trust Officer

→ SPOTLIGHT

Joint statement by international data protection and privacy authorities on web scraping

In August 2023, 12 international data protection and privacy authorities released a joint statement to address data scraping on social media platforms and other publicly accessible websites.

The joint statement outlined:

- → Key privacy risks associated with data scraping, such as targeted cyberattacks, identity fraud, monitoring and profiling individuals, unauthorized political or intelligence gathering, and unwanted direct marketing or spam.
- → How social media companies and other websites should protect individuals' personal information from unlawful data scraping, such as through data security measures and multilayered technical and procedural controls to mitigate the risk.
- → Steps individuals can take to minimize the privacy risks of scraping, including reading a website's privacy policy, limiting information posted online, and understanding and managing privacy settings.

Some key takeaways from the joint statement include:

- → Publicly accessible personal information is still subject to data protection and privacy laws in most jurisdictions.
- → Social media companies and other website operators hosting publicly accessible personal data have legal obligations to protect personal information on their platforms from unlawful data scraping.
- → Accessing personal information through mass data scraping can constitute reportable data breaches in many jurisdictions.
- → Individuals can take steps to prevent their personal information from being scraped, and social media companies have a role to play in empowering users to engage with social media services in a manner that upholds privacy.

> IBM believes it is essential for data management practices tied to AI development to include advanced filtering and curation techniques to identify untrustworthy, protected/sensitive, explicit, biased/nonrepresentative or otherwise unwanted data.
>
> **Christina Montgomery**
> IBM Vice President and
> Chief Privacy and Trust Officer

Implementing AI governance

Numerous strategies are being leveraged to manage data in the context of AI.

Data management plans

Alongside ensuring the lawfulness of data acquisition, there are numerous measures an organization can take to keep track of where the data used to train AI systems comes from. Such organizational practices are especially important with the advent of generative AI, where training data is merged from numerous sources.

Developing a comprehensive plan for how data is managed across an organization is a foundational element to managing all AI systems. Some considerations for data management plans include understanding what data is being used in which system; how it is collected, retained and disposed; if there is lawful consent to use the data; and who is responsible for ensuring the appropriate oversight.

It is likely your organization is already keeping track of the data used across the organization. While there are additional considerations involved when using data for AI systems as discussed above, it is possible to add to your existing data workflows or management practices. It is important to consider the use and management of data used for AI systems at every stage of the life cycle as there are different concerns and implications to consider during different stages. If your organization does not already have a data management practice, resources such as those from Harvard Biomedical Data Management can help you get started.

Additionally, the data management plan should identify relevant data standards, such as ISO 8000 for data quality, to set appropriate controls and targets for your organization to meet. Data standards for aspects of AI are under development through various initiatives at the NIST, ISO/IEC and other national standards bodies.

Data labels

Growing in importance, data labels are tools that can require organizations to provide information on how data was collected and used to train AI models. They are transparency artifacts of AI datasets that explain the processes and rationale for using certain data and explain how it was used in training, design, development and use. This will help explain if the data being used is fit for purpose, if it is representative of the demographics being served with the AI system and if the data meets relevant data quality standards.

Ideally data labels are requirements of a robust data management process, which includes data quality and data impact assessments. While data labels are intended to provide documentation and awareness of the data being used, they can also assist with the assessment and review process. These tools should be aligned where possible within the organization to avoid redundant efforts.

Data-source maintenance through documentation and inventories can help organizations keep track of where the data is acquired and carry out relevant legal due diligence at first-party or third-party levels.

Dedicated processes and functions

When third-party data is used, it is important to follow the terms of service and provide attribution where possible. This will also help inform users of the AI system where the data originated. Where possible, when data is being used from a third party, an appropriate data sharing agreement with clear terms of use for both parties is highly recommended. This helps to resolve any liability issues that may arise as a result of using the system.

INDUSTRY EXAMPLE

When third-party organizations use publicly available data, processes can be put into place. Meta's External Data Misuse team detects, blocks and deters web scraping. Some actions taken by the EDM team include disabling accounts, serving cease-and-desist notices, using CAPTCHAs for bot detection and blocking IP addresses where data scraping is identified. OpenAI has put in place an opt-out process for organizations that do not want GPTbot to access their websites for the purpose of web crawling.

Part III. The privacy and data protection challenge

Privacy and data protection governance practices are woven into the AI life cycle.

Given that AI is a data-dependent enterprise and that privacy law governs the processing of personal data, privacy laws have emerged as a prominent mechanism for managing the key AI governance challenges. After all, information privacy seeks to provide a framework "for making ethical choices about how we use new technologies."

Indeed, national data protection authorities have been among the first to intervene and bring enforcement actions when AI-based products were thought to harm consumers. For example, Italy's data protection authority, the Garante, imposed a temporary ban on ChatGPT after concluding the service was in violation of the GDPR for lacking a legal basis for processing and age-verification mechanism.

> **The enforcement landscape for AI governance is incredibly unsettled. Which regulators will lead on what and how they will collaborate or conflict is subject to heavy debate and will differ by country, creating heightened uncertainty for organizations. Whether or not privacy regulators have the lead remit, they will play a key role given the centrality of data to AI governance.**
>
> Caitlin Fennessy
> IAPP Vice President and Chief Knowledge Officer

Law and policy considerations

The OECD's Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data — developed in 1980 and revised in 2013 — enshrine eight principles that have served as the foundation for most global privacy and data protection laws written over the past several decades, including landmark legislation such as the GDPR. These eight principles include collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation and accountability.

Many DPAs around the world already put forth guidance on how AI systems can work to align themselves with these foundational principles of information privacy. Yet, as Australia's Office of the Victorian Information Commissioner noted in a resource on issues and challenges of AI and privacy, "AI presents challenges to the underlying principles upon which the (OECD Privacy) Guidelines are based." To better understand where these challenges currently exist, each of these principles is discussed below in the context of their applicability to — and potential conflict with — the development of AI.

Collection limitation

The principle of collection limitation states, "There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject." It most readily translates to the concept and practice of data minimization. GDPR Article 5(1)(c) emanates from this idea that data, at the collection stage, should have some predefined limit or upper bound. Specifically, data collection should be "... limited to what is necessary in relation to the purposes for which they are processed." As many observers have noted, this is one of the privacy principles for which there appears to be an "inherent conflict" with AI systems that rely on the collection and analysis of large datasets. Performing adequate AI bias testing, for example, requires collecting more data than might otherwise be collected.

> **At Mastercard, we are testing innovative tools and technologies to address some of the potential tensions between privacy and AI governance. For instance, we know that a lot of data is needed, including sometimes sensitive data, for AI to produce unbiased, accurate and fair outcomes.**
>
> **How do you reconcile this with the principle of data minimization and the need for individual's explicit consent? We are exploring how the creation of synthetic data can help, so as to achieve all desired objectives at the same time.**
>
> Caroline Louveaux
> Mastercard Chief Privacy and Data Responsibility Officer

Data quality

This is the principle that "Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date." Data quality is the privacy principle with which AI may be most in synchrony. The accuracy of AI model outputs depends significantly on the quality of their inputs. A breakdown in AI governance can lead to data becoming inconsistent and error-laden, underscoring the need for AI-based systems to orient themselves around the principle of data quality. Data brokers and other companies can become the target of enforcement actions for failing to ensure the accuracy of the data they collect and sell.

Purpose specification

The principle of purpose specification states, "The purposes for which personal data are collected should be specified … and the subsequent use limited to the fulfilment of those purposes …" Indeed, as the U.K. Information Commissioner's Office explained in the context of its consultation on purpose limitation in the generative AI life cycle, purposes of data processing "must be specified and explicit: organizations need to be clear about why they are processing personal data." This need for clarity applies not only to internal documentation and governance structures, but in communication with the people to whom the personal data relates. In sum, organizations should be able to explain what personal data they process at each stage and why it is needed to meet the specified purpose.

A conflict with the purpose specification principle can arise if and when a developer wants to use the same training dataset to train multiple models. The ICO advises developers reusing training data to consider whether the purpose of training a new model is compatible with the original purpose of collecting the training data. Considering the reasonable expectations of those whose data is being reused can help an organization make a compatibility assessment. Currently, the ICO considers collating repositories of web-scraped data, developing a generative AI model and developing an application based on such a model to constitute different purposes under data protection law.

Use limitation

Related to purpose specification, use limitation is the principle that states personal data "should not be disclosed, made available or otherwise used for purposes other than those specified," except with the consent of the data subject or by the authority of law. Purposes of use must be specified at or before the time of the collection, and subsequent uses must not be incompatible with the initial purposes of collection.

This is another principle that is challenged by AI systems, with potential regulatory gaps left by both the EU GDPR and EU AI Act. Proposals to address these gaps have included restricting the training of models only to stated purposes and requiring alignment between training data collection and the purpose of a model.

Security safeguards

Uniting the fields of privacy, data protection and cybersecurity for decades is the principle that "Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data." Ensuring the security of personal data collected and processed is a key to building and maintaining trust within the digital economy.

Remedying problems of security and safety is and will remain a critical challenge for AI. Ensuring the actions of an AI system align "with the values and preferences of humans" is central to keeping these systems safe. Yet, many AI systems remain susceptible to hacking and so-called "adversarial attacks," which are inputs designed to deceive an AI system, as well as data poisoning, evasion attacks and model extraction. Examples include forcing chatbots to provide answers to responses to harmful prompts or getting a self-driving vehicle's cameras to misclassify a stop sign as a speed-limit sign.

Openness

The right to be informed and the principle of transparency are touchstones of global privacy and data protection laws. Beginning at the collection stage and enduring throughout the life cycle of processing, these rights form the basis of organization's transparency obligations. They often require organizations to disclose various types of information, from the types of data collected and how it is used to the availability of data subjects' rights and how to exercise them to the logic involved and potential consequences of any automated decision-making or profiling the organization engages in. The "black-box" nature of many AI systems can make this principle challenging to navigate and adhere to.

> **“As all AI and machine learning models are 100% data dependent, the models must be fed high-quality, valid, verifiable data with the appropriate velocity. As obvious as that may be, the challenges around establishing the governance requirements that ensure the appropriate use of private data may be far more complex. Modelers should absolutely be applying the minimization principle of identifiable data as they train. Adding private data that could leak or cause bias needs to be thought through early in the design process.**
>
> Scott Margolis
> FTI Technology Managing Director

Individual rights

Individual rights in privacy law commonly include the rights to access, opt in/opt out, erasure, rectification and data portability, among others. Many privacy laws contain rights for individuals to opt-out of automated decision-making underpinned by AI systems.

Accountability

Accountability is arguably one of the most important principles when it comes to operationalizing organizational governance. Accountability is based on the idea that there should be a person and/or entity that is ultimately responsible for any harm resulting from the use of the data, algorithm and AI system's underlying processes.

Implementing AI governance

The practice and professionalization of AI governance is a highly specialized, stand-alone field requiring multidisciplinary expertise. A holistic approach to AI governance requires support from established subject-matter areas, including data protection and information governance practitioners. Data from past IAPP research shows 73% of organizations are leveraging their existing privacy expertise to manage AI governance. This is not surprising, as data is a critical component of AI. Good AI governance weaves privacy and data governance practices into the AI life cycle alongside AI-specific issues. This chapter demonstrates the overlapping nature of privacy and AI governance.

Approaching the implementation of AI governance by adapting existing governance structures and processes enables organizations to move forward quickly, responsibly and with minimal disruption to innovation and the wider business. Target processes that may already be established by organization's data protection program include: accountability, inventories, privacy by design and risk management.

Accountability

Privacy compliance programs are likely to have established roles and responsibilities for those with direct and indirect responsibility for privacy compliance. These are likely supported by policies and procedures to help individuals fulfil the expectations of their role. Senior management contributions are likely channeled through privacy committees, with mechanisms in place to support risk-based escalation, reporting on key metrics and decision-making.

Privacy leaders often have a direct line to CEOs and boards of directors, as well as a matrixed structure of privacy champions across the organization to enable a multidisciplinary approach to privacy governance and ensure data protection needs are considered by product and service teams. This structure is well-suited to, and can be leveraged for, AI governance given the need for leadership engagement and skills spanning legal, design, product and technical disciplines.

Where AI systems process personal data, those with accountability for privacy compliance will need to ensure their existing privacy compliance processes are set up to address the intersection between AI and privacy. This will include considering data inventory, training, privacy by design and other topics further outlined in this section.

Inventories

Personal data inventories have long been the foundation of establishing a successful privacy program and a key requirement of privacy regulations. Knowing your data, how it is collected and used, and being able to demonstrate this remains a core part of accountability. Organizations have also matured in their approaches, from lengthy spreadsheets to technology-enabled approaches.

Where AI systems use personal data, the data inventory can play a crucial role. Organizations that have captured additional privacy compliance metadata alongside the minimum regulatory requirements may find their personal data inventories particularly useful in the age of AI. Additional uses of this metadata could include a single source of truth for lawful basis to identify if additional use within AI models is permitted, accuracy metrics on personal data to support AI models to make accurate inferences based on the latest personal data and a top-down view on processes relying on automated decision-making that can be aligned with AI registries.

> **Legal professionals need to keep an open and flexible mind — technology brings new challenges but also new solutions. General counsel should position themselves as the center of a multidisciplinary team of stakeholders across their organizations, including product design, compliance, data and privacy, which can deploy to manage multifaceted data risks. Companies that strive for established best privacy practice will more easily be able to comply with the rising standards of global privacy laws.**
>
> Tim de Sousa
> FTI Technology, Managing Director, Australia

Effective AI governance is underpinned by AI inventories with similar functionalities to those of data inventories. AI registers can help organizations keep track of their AI development and deployment. Some functional requirements that overlap with data inventories include the ability to connect into the system-development life cycle, maintenance and regular updates by multiple users, and logging capability to ensure integrity.

Privacy by design

By embedding privacy at the outset, privacy by design continues to be a critical part of how organizations address privacy concerns. In implementing privacy by design, privacy functions may take steps to map and embed privacy into areas such as system-development life cycles, project initiation and development approaches within an organization, risk management and approval workflows, and stage gates.

Steps may include developing AI-specific risk-assessment workflows into existing risk-assessment processes, enhancing existing control catalogs with AI and privacy controls, or updating approval workflows to include stakeholders with AI accountabilities.

Additionally, the growing maturity of privacy enhancing technologies and their increasing traction as technical measures within organizations may have benefits for the development of AI. With some PETs potentially helping organizations reduce inherent risk of data use, an organization may be able to maximize the strategic use of its data. Examples include using differential privacy in training machine-learning models, federated learning and synthetic data.

Risk management

The risk-based approach often adopted by global privacy regulations has been distilled into organizational risk-management efforts, which put privacy impact assessments at the heart of deciding whether an organization can reduce harm from personal data processing through the implementation of organizational and technical measures. Privacy risk can also stem from wider privacy compliance activities and lessons learned in areas such as vendor risk, incident management and data subject requests management.

Privacy risk may already feed into wider enterprise risk-management programs, such as information technology and cybersecurity risk and control frameworks. These can be enhanced to accommodate the complex types and sources of AI risk into a unified risk-management framework at the enterprise level. This approach can also facilitate crucial visibility across different subject-matter practice areas across the business and enable a more effective analysis and treatment of AI risk.

As AI risk-management approaches mature, AI governance professionals face choices between embedding algorithmic impact assessments alongside or within PIAs. The need to align AI risk management with broader enterprise risk-management efforts is of equal importance. AI governance professionals will likely need to update enterprise risk-management strategies and frameworks to clearly factor in AI-related risks and document ongoing AI risks and remediations in a formal risk register.

Risk-assessments

A wide range of AI risk assessments are often talked about in the emerging global AI governance landscape.

Some of these assessments are required by existing data protection legislation, such as the GDPR, while others may emerge from AI-specific laws, policies and voluntary frameworks. For the latter, laws and policies often provide AI governance solutions with knowledge of the overlap.

→ SPOTLIGHT
AI governance assessments: A closer look at EU DPIAs and FRIAs

GDPR: DPIAs

Data protection impact assessments are required under GDPR Article 35. DPIAs are particularly important where systematic and extensive evaluation of personal or sensitive aspects of natural persons through automated systems or profiling leads to legal consequences for that person. Incorporating these assessments within the AI-governance life cycle can help organizations identify, analyze and minimize data-related risks and demonstrate accountability.

DPIAs at a minimum contain:

- → A systematic description of the anticipated processing, its purpose and pursued legitimate interest.
- → A necessity and proportionality assessment in relation to the intended purpose for processing.
- → An assessment of the risks to fundamental rights and freedoms.
- → Measures to be taken to safeguard security risks and protect personal data.

EU AI Act: FRIAs

Under the EU AI Act, FRIAS are required to be carried out in accordance with Article 27 by:

- → Law enforcement when they use real-time remote biometric identification AI systems, which are a prohibited AI practice under Article 5.
- → Deployers of high-risk AI systems that are governed by public law, private operators that provide public services and operators deploying certain high-risk AI systems referred to in Annex III, point 5 (b) and (c), such as banking or insurance entities.

FRIAs are required only for the first use of the high-risk AI system, and the act permits deployers to rely on previously conducted FRIAs, provided all information about the system is up to date. FRIAs must consist of:

- → Descriptions of the deployer's processes in line with intended use and purpose of the high-risk AI system.
- → Descriptions of the period and frequency of the high-risk AI system's use.
- → Categories of individuals or groups likely to be affected by the high-risk system.
- → Specific risks of harm that are likely to affect individuals or groups.
- → Descriptions of the human oversight measures in place according to instructions of use.
- → Measures to be taken when risk materializes into harm, including arrangements for internal governance and complaint mechanisms.

However, AI governance solutions often foresee the overlap with existing practices, and this is no different under the EU AI Act. FRIAs, for instance, do not need to be conducted for aspects covered under existing legislation. As such, if a DPIA and FRIA have an overlapping aspect, that aspect need only be covered under DPIA.

Part IV. The transparency, explainability and interpretability challenge

The black-box problem

One reason for the lack of trust associated with AI systems is the inability of users, and often creators, of AI systems to have a clear understanding of how AI works. How does it arrive at a decision? How do we know the prediction is accurate? This is often referred to as the "black-box problem" because the model is either too complex for human comprehension or it is closed and safeguarded by intellectual property.

AI techniques, such as deep learning, are becoming increasingly complex as they learn from terabytes of data, and the number of parameters has grown exponentially over the years. In July 2023, Meta released its Llama 2 model with a parameter count at 70 billion. Google's PaLM parameter count is reported to be as large as 540 billion. Due to the self-learning abilities of AI, including their size and complexity, the black-box problem is increasingly difficult to solve and often requires a trade-off to simplify aspects of the system.

Transparency is a term of broad scope, which can include the need for technical and nontechnical documentation across the life cycle. Having strong product documentation in place can also provide commercial benefits by supporting the product sales cycle and helping providers to navigate prospective clients' due diligence protocols.

In the open-source context, transparency can also refer to providing access to code or datasets in the open-source community to be used by AI systems. Transparency objectives can also include informing users when they are interacting with an AI system or identifying when content was AI generated. Independent of how the term is used, transparency is a key tenet of AI governance due to the desire to understand how AI systems are built, managed and maintained. It is crucial that clear and comprehensive documentation is available to those who design and use these systems to ensure trust and help identify where an error was made if an issue occurs.

Explainability refers to the understanding of how a black-box model, i.e., an incomprehensible or proprietary model, works. While useful, the difficulty with black-box models is that the explanation may not be entirely accurate or faithful to the underlying model, given its incomprehensibility. When full explainability is not possible due to the factors mentioned above, an alternative is interpretability.

Interpretability, on the other hand, refers to designing models that inherently make the reasoning process of the model understandable. It encourages designing models that are not black boxes, with decision or prediction processes that are comprehensible to domain experts. In other words, interpretability is applied ante hoc. While it does away with the problems of explainable models, interpretable models are often domain specific and require significant effort to develop in terms of domain expertise.

Law and policy considerations

One proposed solution to the black-box challenge has been codifying approaches to and requirements for transparency, explainability and interpretability in law or policy initiatives. Regulatory and voluntary governance tools that have established requirements for tackling the black-box problem through transparency and explainability include the EU GDPR and AI Act, NIST AI RMF, U.S. Executive Order 14110, China's Interim Measures for the Management of Generative AI Services, and Singapore's AI Verify.

> **The EU is first out of the gate with comprehensive AI legislation but the EU AI Act is just the tip of the regulatory iceberg. More guidance is coming and many laws enacted since the early 2000s, and under the recent European Data Strategy, will have to be considered in AI governance programs. The EU will continue to promote its approach to regulating AI on the global stage, furthering the Brussels effect on digital regulation.**
>
> Isabelle Roccia
> IAPP Managing Director, Europe

EU GDPR

Arguably one of the first legislative requirements for AI governance, GDPR Articles 13(2)(f), 14(2)(g) and 15(1)(h) refer to providing meaningful information about the logic underpinning automated decisions, as well as information about the significance and envisaged consequences of the automated decision-making for the individual. This is further supported by Article 22 and Recital 71, which state such decision-making should be subject to safeguards, such as through the right to obtain an explanation to challenge an assessment.

EU AI Act

The EU AI Act takes a risk-based approach to transparency, with documentary and disclosure requirements attaching to high-risk and general-purpose AI systems.

It mandates drawing up technical documentation for high-risk AI systems, and requires high-risk AI systems to come with instructions for use that disclose various information, including characteristics, capabilities and performance limitations. To make high-risk AI systems more traceable, it also requires AI systems to be able to automatically allow for the maintenance of logs throughout the AI life cycle.

Similarly, the AI Act places documentation obligations on providers of general-purpose AI systems with and without systemic risks. This includes maintenance of technical documentation, including results from training, testing and evaluation. It also requires up-to-date information and documentation to be maintained for providers of AI systems who intend to integrate GPAI into their system. Providers of GPAI systems with systemic risks must also publicly disclose sufficiently detailed summaries of the content used for training GPAI.

With certain exceptions, the EU AI Act provides individuals with the right to an explanation from deployers of individual decision-making "on the basis of the output from a high-risk AI system ... which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights."

In addition to the documentary and disclosure requirements, the AI Act seeks to foster transparency by mandating machine-readable watermarks. Article 50(2) requires machine-readable watermarks for certain AI systems and GPAI systems, so content can be detected as AI generated or to inform users when they are interacting with AI.

NIST AI RMF

The NIST AI RMF sees transparency, explainability and interpretability as distinct characteristics of AI systems that support each other. Under the RMF, transparency is meant to answer the "what," explainability the "how" and interpretability the "why" of a decision.

→ **Accountability and transparency:** The RMF defines transparency as the extent to which information about an AI system and its outputs are made available to individuals interacting with AI, regardless of whether they are aware of it. Meaningful transparency includes the disclosure of appropriate levels of information at different stages of the AI life cycle, tailored to the knowledge or role of the individual interacting with the system. This could include design decisions, the model's training data and structure, intended use-cases, and how and when deployment, post-deployment or end-user decisions were made and by whom. The RMF requires AI transparency to consider human-AI interaction, such as by notifying the human if a potential or actual adverse outcome is detected.

→ **Explainable and interpretable AI:** The RMF defines explainability as a representation of the underlying mechanisms of the AI system's operation, while it defines interpretability as the meanings assigned to the AI outputs in the context of their designed functional purpose.
Lack of explainability can be managed by describing how the system functions by tailoring such descriptions to the knowledge, roles and skills of the individual, whereas lack of interpretability can be managed by describing why the AI system gave a specific output.

> **It's important to align on a set of ethical AI principles that are operationalized through tangible responsible AI practices, rooted in regulations, e.g. EU AI Act, and best practice frameworks, e.g. NIST AI RMF, when developing AI features. At Workday, we take a risk-based approach to responsible AI governance.**
>
> **Our scalable risk evaluation dictates relevant guidelines such as requirements to map, measure, and manage unintended consequences including bias. Within the Workday AI Feature Fact Sheets, Workday provides transparency to customers on each feature such as, where relevant, how they were assessed for bias.**
>
> **These safeguards are intended to document our efforts to develop AI features that are safe and secure, human centered, and transparent and explainable.**
>
> Barbara Cosgrove
> Workday Vice President, Chief Privacy Officer

U.S. Executive Order 14110

U.S. Executive Order 14110 approaches transparency through an AI safety perspective. Under Section 4, the safety and security of AI technology is to be ensured through certain transparency measures, such as the requirements to share results of safety tests and other important information with the U.S. government, that have been imposed on developers of the most powerful AI systems. Watermarks to label AI-generated content are also required under the order, with the purpose of protecting Americans from AI-enabled fraud and deception.

China's Interim Measures for the Management of Generative AI Services

Article 10 of China's Interim Measures for the Management of Generative AI Services requires providers of AI services to clarify and disclose the uses of the services to user groups and to guide their scientific understanding and lawful use of generative AI. Watermarking AI-generated content is also a requirement under Article 11.

Singapore's AI Verify

Singapore's AI Verify is a voluntary testing framework on AI governance for organizational use comprised of two parts: a testing framework grounded in 11 internationally accepted principles grouped into five pillars and a toolkit to execute technical tests.

Transparency and explainability themes are among the 11 principles embedded in AI Verify. The framework addresses the transparency problem by providing impacted individuals with appropriate information about AI use in a technological system so they can make informed decisions on whether to use that AI enabled system. Explainability, on the other hand, is achieved through an understanding of how an AI model reaches a decision, so individuals are aware of the factors that contributed to a resulting output. Transparency is assessed through documentary evidence and explainability is assessed through technical tests.

Implementing AI governance

Organizations have been active in coming up with tools and techniques to address the black-box transparency and explainability challenge.

Model and system cards

Model cards are short documents that accompany an AI model to provide transparent model reporting by disclosing information about the model. Information may include explanations about intended use, performance metrics and benchmarked evaluation in various conditions such as across different cultures, demographics or race. In addition to providing transparency, model cards are also meant to discourage use of models outside their intended uses. At the industry level, use of model cards is becoming more prominent as evidenced by publicly accessible model cards for Meta and Microsoft's Llama 2, OpenAI's GPT-3 and Google's face-detection model.

It may not always be easy to explain a model in a short document. Model cards are to serve a broad audience and, therefore, standardizing explanations may prove either too simplistic for one audience or too complicated for another. Moreover, organizations should also be mindful of how much information they reveal in the cards to prevent adversarial attacks and mitigate security risks.

AI models are often part of a larger system comprised of a group of models and technologies that work together to give outputs. As a result, model cards can fall short of providing a more nuanced picture of how different models interact together within the system. That is where system cards can help achieve better insights.

INDUSTRY EXAMPLE

Meta released 22 system cards explaining how AI powers its Facebook and Instagram platforms. Each card has four sections that detail:

1. An overview of the AI system.
2. How the system works by summarizing the steps involved in creating experiences on Facebook and Instagram.
3. How the shown content can be customized.
4. How AI delivers content as part of the whole system.

System cards explain how a group of AI models and other AI and non-AI technologies work together as part of an AI system to achieve specific tasks. Meta released 22 system cards explaining how AI powers its Facebook and Instagram platforms. Each card has four sections that detail:

→ An overview of the AI system.

→ How the system works by summarizing the steps involved in creating experiences on Facebook and Instagram.

→ How the shown content can be customized.

→ How AI delivers content as part of the whole system.

AI systems learn from their environments and constantly evolve, so the way they work also changes over time, requiring updates to the system cards. Like with model cards, reducing technical concepts to a standardized language that serves all audiences can be challenging for system cards, and system cards can also attract security threats based on the amount and type of information shared.

The utility of model and system cards can go beyond meeting transparency challenges. Maintaining standardized records about the model itself can facilitate communication and collaboration between various stakeholders throughout the life cycle. This can also help with bias and security-risk mitigation. They are also useful for making comparisons with future versions of the models to track improvements. The cards provide a documented record of design, development and deployment, so they can facilitate attribution of responsibility for various decisions and outcomes related to the model or system. Auditors can use them not only to gain a holistic understanding of the system itself, but also to zoom in on the processes and decisions made during different phases of the life cycle.

> At IBM, we believe that open technology and collaboration are essential to further the responsible adoption of AI. An open approach can support efforts to develop and implement leading technical methods, such as those used during the testing and evaluation of data and AI systems.
>
> **Christina Montgomery**
> IBM Vice President and Chief Privacy and Trust Officer

Open-source AI

Another approach to addressing the black-box challenge is making AI open source. This requires making the source code public and allowing users to view, modify and distribute it freely. Open access can be especially useful for researchers and developers, as there is more potential for scrutiny by a wider, diverse and collaborative community of experts. In turn, that can lead to improvements to the transparency of algorithms, the detection of risks and the offering of solutions if things go wrong. Open-source AI can also improve technology access and drive collaborative innovation, which may otherwise be limited by proprietary algorithms.

Watermarking

With the rise of generative AI, it is becoming increasingly difficult to distinguish AI-generated content from human-created content. To ensure transparency, the watermarking or labeling of AI generated content has been legally mandated under the EU AI Act, U.S. Executive Order 14110 and state-level requirements, and China's Interim Measures for Management of Generative AI.

INDUSTRY EXAMPLE

Meta's Llama is open source, and it aims to power innovation through the open-source community. Given the safety and security concerns associated with generative AI models, Llama comes with a responsible-use guide and an acceptable-use policy. On the other hand, OpenAI has taken a closed approach toward its large language models, such as GPT-3 and GPT-4, in the interest of maintaining its competitive advantage, as well as to ensure AI safety.

INDUSTRY EXAMPLE

Google uses a technology called SynthID, which directly embeds watermarks into Google's text-to-image generator Imagen. Meta has moved toward labeling AI-generated images on Facebook, Instagram and Threads. Although Meta already adds the label "Imagined with AI" on images generated through its AI feature, it now also aims to work with industry partners on common standards to add multilingual labels on synthetic content generated with tools of other companies that users post on Meta's platforms. Specifically, Meta is relying on Partnership on AI's best practices, the Coalition for Content Provenance and Authenticity's Technical Specifications and the International Press Telecommunications Council's Technical Standards to add invisible markers at scale to label AI generated content by tools of companies such as Google, Microsoft and OpenAI.

Watermarking is gaining traction as a way for organizations to promote transparency and ensure safety against harmful content, such as misinformation and disinformation. Companies are embedding watermarks on AI-generated content. Watermarks are invisible to the human eye, but are machine readable and can be detected by computers as AI generated.

While watermarking is becoming a popular technique for transparency, it is still not possible to label all AI generated content. Moreover, techniques to break watermarks also exist.

> **Focusing on the building blocks of AI governance — like appropriate documentation of AI models and systems — is important because those foundations are necessary to enable risk management, impact assessment, and third-party auditing.**
>
> Miranda Bogen
> Center for Democracy and Technology
> AI Governance Lab Director

Part V. The bias, discrimination and fairness challenge

Hidden and harmful biases may lurk within an AI system.

Bias, discrimination and fairness are among the most important challenges of AI governance, given their potentially very significant real-world impacts on individuals and communities. Leaving this challenge unaddressed can lead to discriminatory outcomes and perpetuate inequalities at scale. Healthy AI governance must promote legal and ethical norms including human rights, professional responsibility, human-centered design and control of technology, community development and nondiscrimination.

While the automation of human tasks using AI has the advantages of scalability, efficiency and accuracy, it is accompanied by the challenge of algorithmic bias, whereby a systematic error manifests through an inaccuracy in the algorithm. It occurs when an algorithm systematically or repeatedly misses certain groups of people more than others. With transparency challenges around how or why an input turns into a particular output, biases in the algorithm can be difficult to trace and identify.

Instances of algorithmic bias have been well documented in policing, criminal sentencing and hiring. Algorithmic bias can impact even the most well-intentioned AI systems, and it can enter a model or system in numerous ways.

Ways biases may get into the AI system

Biases may get into the AI system in multiple ways during the input, training and output stages.

At the input stage

- **Historical data.** If historical data used to train algorithms is biased, then the algorithm may learn those biases and perpetuate them. For example, if an AI recruitment tool is trained on historical data containing gender or racial biases, those biases will be reflected in the tool's hiring decisions or predictions.

- **Representation bias.** Biases can also enter the algorithm through data that either overrepresents or underrepresents social groups. This can make the algorithmic decisions less accurate and create demographic or social disparities.

- **Inaccurate data.** The accuracy of data can be impaired if it is outdated or insufficient. Such data falls short of fully representing current realities, leading to inaccurate results, which may also lead to reinforcement of historical biases.

At the training stage

- **Model.** Biases can arise when they are an intrinsic part of the model itself. For example, models developed through traditional programming, i.e., those manually coded by human designers, can have intrinsic biases if they are not based on real-world insights. An algorithm assisting with university admissions may be biased if the human designer programmed it to give a higher preference to students from private schools over students from public schools. Intrinsic biases may be difficult to spot in AI models, as they are a result of self-learning and make correlations across billions of data points, which are often part of a black box.

- **Parameters.** The model adjusts its parameters, such as weights and biases in neural networks, during the training process based on the training data. Bias can manifest when the values assigned to these parameters inadvertently reinforce the bias present in the training data or the decisions made by the designers during architecture selection. In an algorithm for university admissions, for example, the attributes of leadership and competitiveness can reflect a gender stereotype present in the training data with the algorithm favoring male candidates over female ones. However, bias in parameters can also manifest more stealthily, such as through proxies. In absence of certain data, the algorithm will make correlations to make sense of the missing data. An algorithm for loan approval, for example, may disproportionately assign more weight to certain zip codes and the model may inadvertently perpetuate racial or ethnic bias by rejecting loan applications using zip codes as a proxy.

> “At Microsoft, we are steadfast in our commitment to developing AI technologies that are not only innovative but also trustworthy, safe, and secure. We believe that the true measure of our progress is not just in the capabilities we unlock, but in the assurance that the digital experiences we create will enhance rather than compromise the human experience.
>
> **Julie Brill,**
> Microsoft Chief Privacy Officer, Corporate Vice President

At the output stage

→ **Self-reinforcing biases.** A feedback loop is a process through which the AI system continues to learn based on the outputs it generates. The output goes back into the system as an input, which can influence the system's behavior or performance in some positive or negative way. While feedback loops can foster continuous learning and allow the system to adapt to its deployed environment, they can also lead to self-reinforcing biases if the outputs of the algorithm itself are biased. For example, if an algorithm consistently rejects loan applications for women and consistently approves them for men, there may be a gender bias at play, and the algorithm could fall into a loop where it learns from the biased outputs and continues to reinforce the biased pattern.

→ **Human oversight.** Although it is necessary to have humans in the loop throughout the life cycle of the AI system, there is a risk that human biases can reenter the algorithm. For example, human control over a system's final output is necessary, but bias can externally impact the output based on the human interpretation applied to that final output.

→ **Automation bias.** Automation bias refers to the human tendency to overly rely on automated outputs. This leads to people trusting the recommendations of algorithms without questioning or verifying their accuracy or being mindful of the system's limitations and errors. This can be especially dangerous when confirmation bias about protected characteristics is at play. That is, users are more likely to accept the outputs when they align with their preexisting beliefs.

Bias detection and mitigation is particularly challenging in the context of foundation models due to their size and complex architectures.

→ SPOTLIGHT

Joint statement by US Federal Agencies

In April 2023, a joint statement made by four federal agencies, namely the EEOC, Department of Justice, Federal Trade Commission and Consumer Financial Protection Bureau, reiterated the U.S.'s commitment to the principles of fairness, equality and justice, which are deeply embedded in federal laws. In April 2024, five additional cabinet-level agencies joined that pledge.

The joint statement now includes the Department of Education, Department of Health and Human Services, Department of Homeland Security, Department of Housing and Urban Development, and Department of Labor. The Consumer Protection Branch of the Justice Department's Civil Division also joined the pledge. Enforcement of discrimination through automated means is managed by these federal agencies.

Law and policy considerations

Many existing equalities and antidiscrimination laws apply to AI systems and many emerging initiatives specific to AI governance include provisions on bias.

Depending on the jurisdiction where the organization operates, liability could also fall under relevant civil rights, human rights or constitutional freedoms of that jurisdiction.

In the U.S., civil rights can be protected through private rights of action by individuals. For example, according to the guidance provided by the U.S. Equal Employment Opportunity Commission, private rights of action against discrimination through algorithms could occur under the Americans with Disability Act and Title VII of the Civil Rights Act. Under both the ADA and Title VII, employers can be exposed to liability even where their algorithmic decision-making tools are designed or administered by another entity. When individuals think their rights under either of those laws have been violated, they can file a charge of discrimination with EEOC.

OECD AI Principles

The principle of "human-centered values and fairness" in the OECD AI Principles requires respect for the rule of law, human rights and democratic values across the life cycle of the AI system, through respect for nondiscrimination and equality, diversity, fairness, and social justice. This is to be implemented through safeguards, like context-appropriate human determination that is consistent with the state of the art. The OECD AI Policy Observatory maintains a catalogue on tools and metrics for practically aligning AI with OECD's principles, including bias and fairness.

UNESCO Recommendations on the Ethics of AI

Principles 28, 29 and 30 of UNESCO's Recommendations on the Ethics of AI encourage AI actors to promote access to technology to diverse groups, minimize the reinforcement or perpetuation of discriminatory or biased outcomes throughout the life cycle of the AI systems, and reduce the global digital divide. Among the tools provided by UNESCO for the practical implementation of its recommendations is the ethical impact assessment, which is designed primarily for government officials involved in the procurement of AI systems but can also be used by companies to assess if an AI system aligns with UNESCO's Recommendations.

EU AI Act

The EU AI Act provides a relevant framework for data governance of high-risk AI systems under Article 10, which permits training, validation and testing of datasets to examine the possibility of biases that affect the health and safety of persons and negatively impact fundamental rights or lead to discrimination. To deal with the challenge of self-reinforcing biases, Article 15(4) also requires the elimination or reduction of biases emanating from feedback loops in high-risk AI systems after they have been put on the market or into service. The EU AI Act calls for consideration of the European Commission's Ethics Guidelines for Trustworthy AI, which are voluntary guidelines seeking to promote "diversity, non-discrimination and fairness."

Singapore

Singapore's AI Verify tackles bias via the principle of "ensuring fairness." This principle is made up of the pillars of data governance and fairness. While there are no specific tests for data governance in the toolkit, if the model is not giving biased outputs based on protected characteristics, fairness can be ensured by checking the model against ground truth. Process checks include the verification of documentary evidence that there is a strategy for fairness metrics and that the definition of sensitive attributes is consistent with the law.

→ SPOTLIGHT
US FTC Enforcement priorities and concerns

We have made no secret of our enforcement priorities and concerns.

1. There is no AI exemption from the laws on the books and businesses need to develop and deploy AI tools in ways that allow for an open and competitive market and protect consumers from potential harms.

2. We are scrutinizing existing and emerging bottlenecks across the AI design stack to ensure that businesses aren't using monopoly power to block innovation and competition.

3. We are acutely aware that behavioral advertising, brought on by web 2.0, fuels the endless collection of user data and recognize that model training is emerging as another feature that could further incentivize surveillance.

4. We are squarely focused on aligning liability with capability and control, looking upstream and across layers of the AI stack to pinpoint which actor is driving or enabling the lawbreaking.

5. We are focused on crafting effective remedies in cases that establish bright-line rules on the development, use and management of AI inputs, such as prohibiting the uses of inaccurate or highly-sensitive data when training models.

Samuel Levine
FTC Bureau of Consumer Protection Director

> As we navigate the transformative potential of AI, it is imperative that we anchor our journey in our collective ability to protect fundamental rights and the enduring values of safety and accountability.
>
> **Julie Brill**
> Microsoft Chief Privacy Officer, Corporate Vice President

The US

In the U.S., antidiscrimination laws that also extend to AI are scattered across various sectors, such as employment, housing and civil rights.

- **Employment.** Under the Americans with Disabilities Act, employers are prohibited from using algorithmic decision-making tools that could violate the act, such as in not providing reasonable accommodations, intentionally or unintentionally screening out an individual with a disability, or adopting disability-related inquiries and medical examinations.

- **Housing.** In 2023, to ensure fairness in housing, the Biden-Harris Administration issued a proposed rule against racial bias in algorithmic home valuations, empowering consumers to take action against appraisal bias, increasing transparency and leveraging federal data to inform policy and improve enforcement against appraisal bias.

- **Consumer finance.** The CFPB confirmed companies are not absolved of their legal responsibilities under existing legislation, such as the Equal Credit Opportunity Act, when they use AI models to make lending decisions. Remedies include compensating the victim, providing injunctive relief to stop unlawful conduct, or banning persons or companies from future participation in the marketplace.

- **Voluntary frameworks.** The NIST Special Publication 1270: Towards a Standard for Identifying and Managing Bias in AI lays down governance standards for managing AI bias. These include monitoring the system for biases, making feedback channels available so users can flag incorrect or harmful results for which they can seek recourse, putting policies and procedures in place for every stage of the life cycle, maintaining model documentation to ensure accountability, and embedding AI governance within the culture of the organization.

→ SPOTLIGHT
FTC enforcement action against Rite Aid

On 19 Dec. 2023, the FTC issued an enforcement action against Rite Aid's discriminatory use of facial recognition technology. Rite Aid deployed facial recognition surveillance systems for theft deterrence without assessing the accuracy or bias of the system.

Rite Aid recorded thousands of false-match alerts, and the FTC's gender-based analysis revealed Black, Asian, Latino and women consumers were more likely to be harmed by Rite Aid's surveillance technology.

The FTC placed a five-year moratorium on Rite Aid's use of facial recognition, and if after five years Rite Aid chooses to use this technology again, it will have to implement the FTC's governance plan detailed in the order. The enforcement decision also included an order for disgorgement, that is, to delete or destroy any photos and videos including any data, models or algorithms used for surveillance.

This case serves as good indication of the nature and intensity of liability that deployers and providers of AI in the U.S. may be exposed to for deploying discriminatory AI systems.

Implementing AI governance

One overarching practice used to mitigate biases is the promotion of diversity and inclusivity among teams working across the life cycle of the AI system. Personnel composition is often supported by organization-level principles for safe and responsible AI, many of which internal AI ethics policies, e.g., Google, IBM and Microsoft.

Bias testing

One way to minimize bias in AI systems is by testing the systems. While there are numerous ways to test for bias in AI systems, it is important to understand what is being evaluated. Demographic parity may be different than equality objectives. Establish goals based on the desired system outcomes to start, and then establish an appropriate technique for testing bias within the system. For example, does fairness mean an equal number of males and females will be screened for a new position on your team, or that candidates with the most distinguished resumes are identified as ideal applicants independent of their gender, race, experience, etc.

It is important to note that often testing for bias will require the use of personal information to determine if fairness objectives are being met. As such, there may be a privacy-bias trade-off, as safeguarding privacy through data minimization creates challenges for mitigating biases in AI systems. Some considerations when balancing privacy while mitigating bias include:

- Intentionally collecting sensitive data directly in the design phase so it is ready for the testing phase. This can be done by procuring consent from data subjects and disclosing the purpose for the collection and processing of their data.
- Creating intentional proxies to test how the system makes correlations without sensitive data, such as for demographic features.
- Buying missing data from data brokers, public data or other datasets in compliance with privacy and data governance policies.

> **Fairness tests and debiasing methods are not created equally — as an AI deployer or governance professional, it is critically important to use tools and methods that fundamentally align with equality and nondiscrimination law in your jurisdiction.**
>
> **Brent Mittelstadt**
> University of Oxford Internet Institute Director of Research, Associate Professor and Senior Research Fellow

Part VI. The security and robustness challenge

The pace, scale, and reach of AI development and integration demands strong security.

Compromises to the security of AI could result in manipulation of outputs, stolen sensitive information or interference with system operations. Unsecured AI can result in financial losses, reputational damage and even physical harm. For example, exploiting the vulnerabilities of medical AI could lead to a misdiagnosis, and adversarial attacks on autonomous vehicles could lead to road traffic accidents.

Although AI security overlaps with and suffers from traditional cybersecurity risks, cybersecurity is often about protecting computer systems and networks from attacks, whereas AI security is about guarding the AI system's components, namely the data, model and outputs. When it comes to AI security, malicious actors can enable adversarial attacks by exploiting the inherent limitations of AI algorithms.

Adversarial attacks

Adversarial attacks are a deliberate attempt to manipulate models in a way that leads to incorrect or harmful outputs. The intention behind the attack could be to lead the model toward misclassification or cause harm, and all it may take to trick the model is a slight switching of pixels or adding a bit of noise. Some types of adversarial attacks include:

- **Evasion attacks.** The aim of evasion attacks is to deceive the model into misclassifying data, such as by adding a small perturbation to the input image, as in the MIT example, leading to an incorrect output with high confidence.

- **Data poisoning.** This can happen in various ways, such as by switching the labels of labeled data or injecting entirely new data into the dataset. However, for this to work, the adversary will have to first gain access to training data. Data poisoning can also help attackers create backdoors so they can manipulate model behavior in the future.

- **Model extraction.** The aim of model extraction is model theft by reverse engineering to reveal the hidden mechanism of the model or sensitive information, or to make the model vulnerable to further attacks. This is done by feeding carefully crafted queries to a black-box model to analyze its outputs and steal its functionality. This can help the adversary copy the model and make financial gains.

AI vulnerabilities can also be exploited through open-source software and third-party risks.

- Open-source software can be manipulated in many ways, such as through supply-chain attacks, in which open-source AI libraries are targeted by malicious code that is planted as a legitimate update or functionality. Although open-source software suggests everything has been made publicly available, the original developers can restrict access to some parts of the software in the license agreement. In such cases, hackers may resort to model extraction. Even if an AI system is not open source, the project may rely on a complex ecosystem of open-source tools, exposing itself to a potential attack surface that malicious actors can exploit.

- A lack of control and visibility over third-party governance practices makes risk mitigation more difficult, including with respect to security. Third-party vendors may have weaker security standards and practices, making them more vulnerable to data breaches, supply chain attacks and system hacks, among other security risks.

Law and policy considerations

Regulatory and voluntary governance tools that have established requirements for tackling AI security issues include the NIS2 Directive, U.S. Executive Order 14110, the NIST AI RMF and the NIST Cybersecurity Framework.

NIS2

The NIS2 Directive replaces the EU Network and Information Security Directive from 2016. It aims to boost resilience and incident-response capacities in public and private sectors through risk management and reporting obligations. Some cybersecurity requirements under Article 21 include policies on risk analysis and system security, incident handling, supply-chain security, policies and procedures to assess cybersecurity risk-management effectiveness, cyber hygiene practices, policies on the use of cryptography, and encryption.

EU AI Act

As with most AI themes, under the EU AI Act, security takes a risk-based approach. As such, security and robustness requirements vary based on if the system is high risk or if it is a GPAI system with systemic risks.

→ **High-risk AI systems.** The EU AI Act lays down detailed security obligations for accuracy, security and robustness of high-risk AI systems. Technical and organizational measures are to be placed to ensure high-risk systems are resilient toward errors, faults and inconsistencies. Possible solutions include back-up or fail-safe plans.

The act also foresees risks emerging at the third-party level, requiring resilience against unauthorized third-party attempts to alter use, outputs or performance by exploiting the system vulnerabilities. Technical solutions to handle such security risks must be appropriate to circumstances and risk. These can include measures to prevent, detect, respond to, resolve and control data poisoning, model poisoning, adversarial examples and model evasion, confidentiality attacks, or model flaws.

Additionally, the EU AI Act obliges providers of high-risk AI systems to ensure they undergo conformity assessments that demonstrate compliance with requirements for high-risk systems.

→ **Obligations for providers of GPAI systems with systemic risks.** The EU AI Act lists the security requirements for high-impact AI systems. Requirements include:

- Evaluating models in accordance with standardized protocols, such as conducting and documenting adversarial testing to identify and mitigate systemic risks.
- Monitoring, documenting and reporting serious incidents to the AI Office.
- Ensuring GPAI models with systemic risks and their physical infrastructures have adequate levels of cybersecurity.

> There is an urgent need to respond to the complex challenges of AI governance by professionalizing the field. A professionalized workforce can take AI governance from theory to practice, spread trustworthy and standardized practices across industries and borders, and remain adaptable to swiftly changing technologies and risks.
>
> J. Trevor Hughes
> IAPP President and CEO

US Executive Order 14110

The U.S. AI Executive Order 14110 calls on developers of the most powerful AI systems to share their safety results and other critical information with the U.S. government. It also calls on the NIST to develop rigorous standards for extensive red-team testing to ensure safety before public release.

NIST AI RMF

The NIST AI RMF identifies common security concerns such as data poisoning and exfiltration of models, training data or other intellectual property through AI system endpoints. Under the AI RMF, a system is said to be secure when it can maintain confidentiality, integrity and availability through protection mechanisms that prevent unauthorized access and use. Practical implementation can be achieved through the NIST Cybersecurity Framework and RMF.

NIST Cybersecurity Framework

The NIST Cybersecurity Framework is a voluntary framework that provides standards, guidelines and best practices for organizations to mitigate cybersecurity risks. The framework is organized under five key functions: identify, protect, detect, respond and recover.

INDUSTRY EXAMPLE

OpenAI's latest text-to-video model, Sora, was red teamed. In preparation for the U.K. AI Safety Summit, Meta released a document detailing the safety of its Llama 2 model, as well as the benchmarks and potential attack vectors it was red teamed for.

Implementing AI governance

Due diligence in the identification of security risks throughout the life cycle of the system is an important activity, especially when a third-party vendor is involved. Due diligence can only ever inform. With appropriate information, an organization can seek contract terms with third-party vendors that mandate:

- Making the vendor's security practices compatible with the organization's own standards.
- Monitoring system robustness regularly through security assessments or audits to identify third-party risks and ensure the vendor is complying with the organization's security standards.
- Limiting access to third-party vendors only for the services they need to perform.

Red teaming

Red teaming is the process of testing the security of an AI system through an adversarial lens by removing defender bias. It involves the simulation of adversarial attacks on the model to evaluate it against certain benchmarks, "jailbreak" it and make it behave in unintended ways. Red teaming reveals security risks, model flaws, biases, misinformation and other harms, and the results of such testing are passed along to the model developers for remediation. Developers use red teaming to bolster and secure their product before releasing it to the public.

Secure data sharing practices

Differential privacy is primarily a privacy-enhancing technique that also has security benefits, it analyzes group data while preserving individual privacy by adding controlled noise to the data and blurring individual details. So, even if an attacker were to steal this data, they would not be able to link it back to specific individuals, minimizing harm. As such, differential privacy can limit the utility of stolen data. However, that impact to the utility of the data can also impact organizations with lawful and legitimate interests in processing the data. Moreover, differential privacy can also be a costly technique to implement, especially where large datasets are concerned.

HITL

Human in the loop refers to incorporating human expertise and oversight into the algorithmic decision-making process. Although HITL may provide a gateway for human biases to reenter the algorithm when making judgements about final outputs, in the context of AI security, HITL can make incident detection and response more efficient. This is especially true where subtle manipulations or attacks that the model may not have been trained to identify are involved. HITL allows for continuous monitoring and verification, however, optimal use of this approach rests on balancing the contradictions that may arise to address bias or safety and security.

Part VII. AI safety

AI safety is a cornerstone but somewhat mercurial principle for realizing safe and responsible AI.

Various themes, particularly value alignment, transparency and AI security, eventually culminate into the broader theme of AI safety. Given that safety is an all-encompassing theme, it has no settled global definition. It may include preventing so-called existential risks posed by artificial general intelligence. For some, such as the Center for AI Safety, AI risk is categorized based on malicious use, the AI race, rogue behavior and organizational risks. For others, such as the country signatories to the Bletchley Declaration, and most recently, for parties to the Seoul Declaration for Safe, Innovative and Inclusive AI, it is about managing risks and being prepared for unexpected risks that may arise from frontier AI. AI safety can also be the term used to describe minimizing AI harms from misinformation, disinformation and deepfakes, and the unintended behavior of an AI system, especially advanced AI systems.

Law and policy considerations

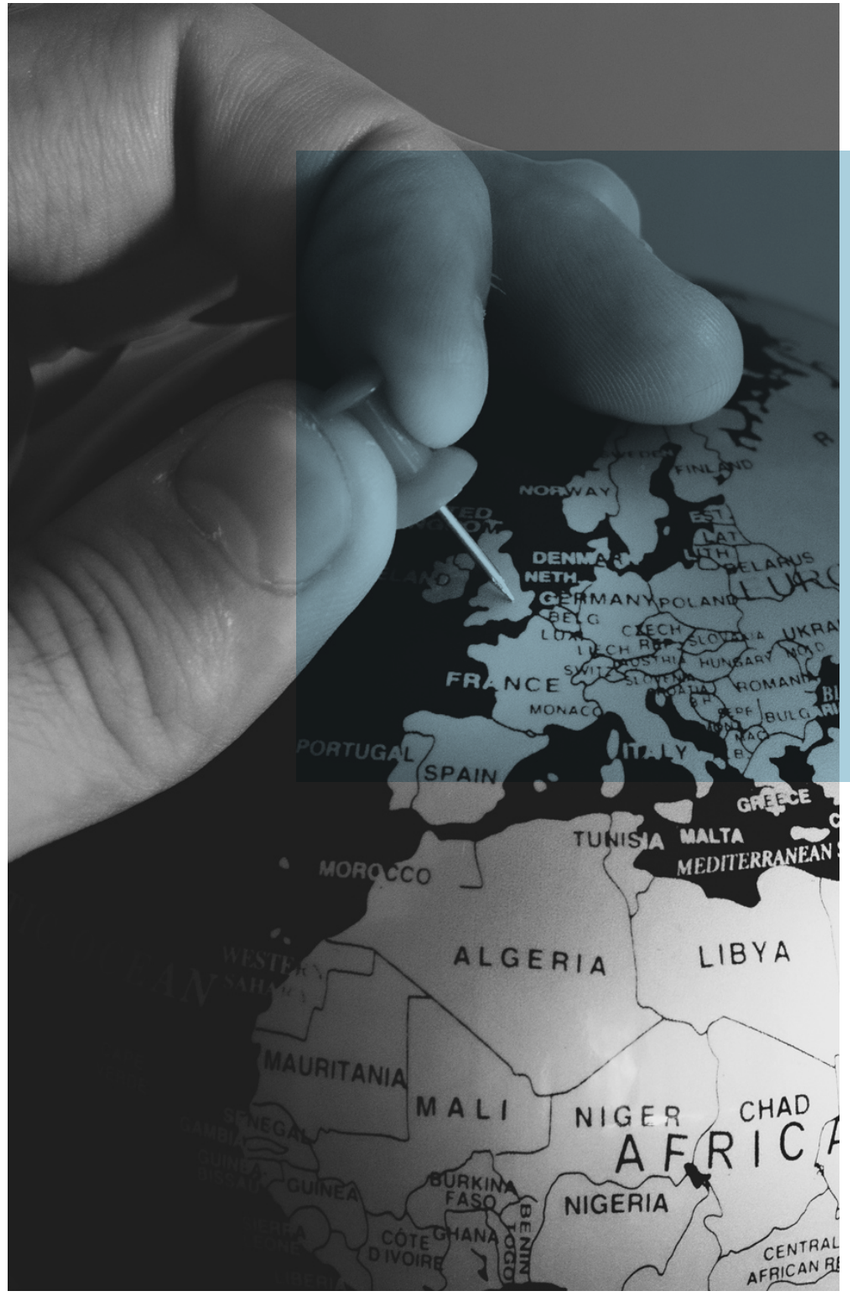
The importance of AI safety is reflected in the fact that, for some jurisdictions, it has been embedded as a main theme in national strategies toward AI. The Biden-Harris Administration's Executive Order 14110 focuses on developing "Safe, Secure and Trustworthy" AI. In 2023, the U.K. brought world leaders together for first AI Safety Summit, and the country's approach toward AI is focused on the safety of advanced AI systems, or "frontier AI." Safety is also an important factor under the EU AI Act, which is reflected in the security and robustness requirements for high-impact GPAI systems and high-risk AI systems.

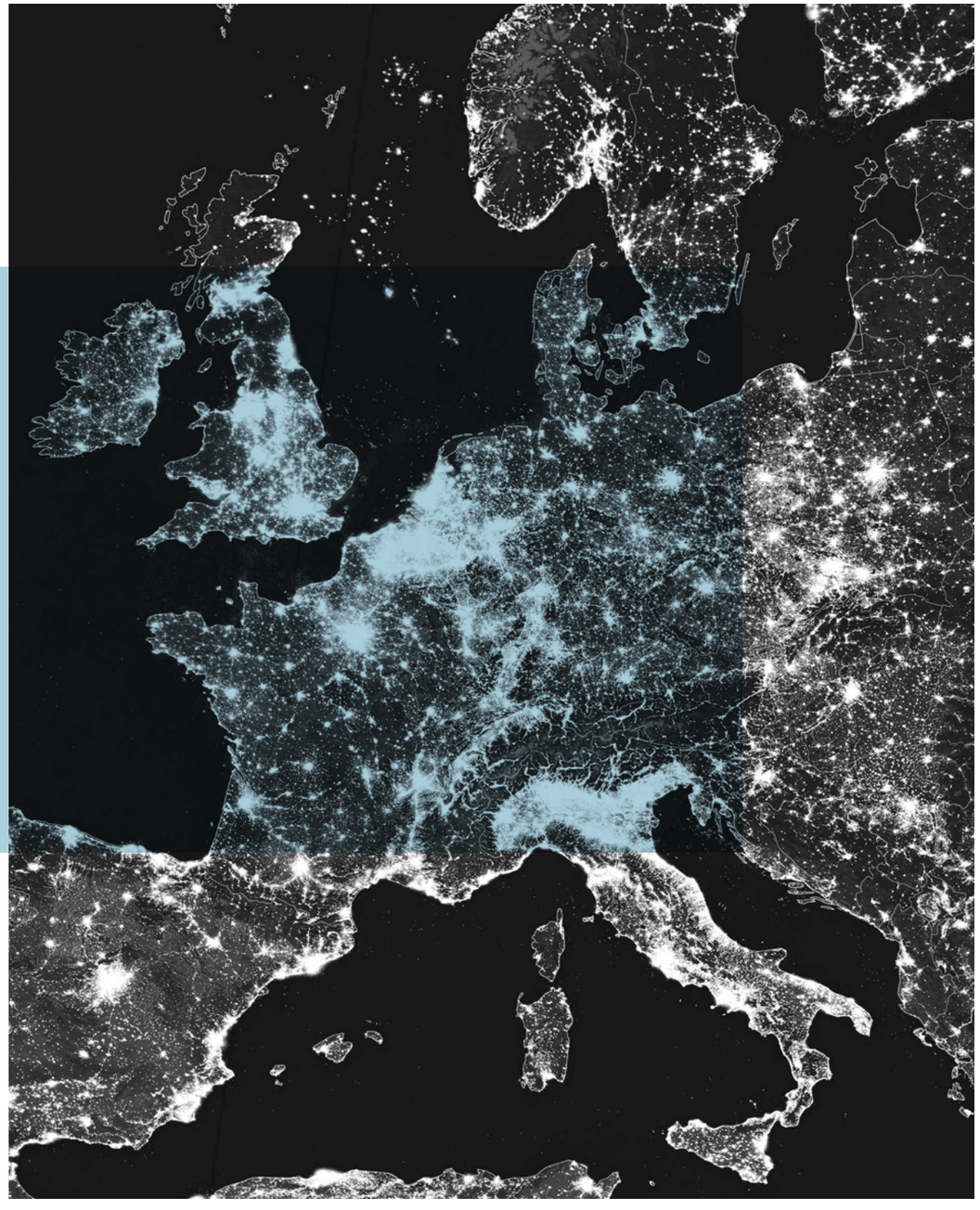
AI safety institutes

Recently, the NIST announced it would establish the U.S. AI Safety Institute. To support this institute, the NIST also created an AI Safety Institute Consortium, which brought more than 200 organizations together to develop guidelines and standards for AI measurement and policy that can lay the foundation for AI safety globally.

Among many security- and safety-related initiatives, the AISIC is tasked with enabling collaborative and interdisciplinary research and establishing a knowledge and data sharing space for AI stakeholders. More specifically, the AISIC will develop new guidelines, tools, methods, protocols and best practices to facilitate the evolution of industry standards for AI safety. The AISIC will also develop benchmarks for evaluating AI capabilities, especially harmful ones.

The U.K. government established an AI Safety Institute to build a sociotechnical infrastructure that can minimize risks emerging from unexpected advancements in AI technology. The institute has been entrusted with three main functions: developing and conducting evaluations on advanced AI systems, driving foundational AI research, and facilitating the exchange of information.





Bletchley Declaration

The 2023 U.K. AI Safety Summit brought together international governments, leading AI companies and civil society groups to discuss frontier AI risks and ways to promote AI safety. As a demonstration of their commitments to AI safety, participating nations also signed the Bletchley Declaration, which makes various affirmations to cooperate globally on innovation, sustainable development, economic growth, protection of human rights and fundamental freedoms, and building public trust and confidence in AI technology.

EU AI Act

The security requirements for general-purpose AI systems under the AI Act are also focused on regulating "systemic risks." The EU AI act defines this risk as one emerging from high-impact general purpose models that "significantly impact the internal market, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain."

AI Safety Standards

ISO/IEC Guide 51:2014 provides requirements and recommendations for drafters of standards to include safety aspects in those standards. It applies to safety aspects pertaining to people, environments or both.

> We are generating 2.5 quintillion bytes of data globally per day. Much of this is flowing into our internet. Therefore, generative AI models are dynamic and the applications that are built on top of them will move. It is up to the organizations to ensure that the movement meets their standards.
>
> **Dominique Shelton Leipzig**
> Mayer Brown Partner, Cybersecurity & Data Privacy and Leader, Global Data Innovation & AdTech

→ SPOTLIGHT
Compute governance

On a broader level, AI safety also refers to regulating compute, i.e., the power source of AI systems, as regulating AI at its source increases the visibility of its technical capabilities. Unlike AI models, which can be replicated exponentially and without control, compute must be purchased and is quantifiable. As computing chips are manufactured through highly concentrated supply chains and dominated by only a few companies, regulatory interventions can be more focused. Such regulation can purposefully occur with AI safety in mind to control the allocation of resources for AI projects by subsidizing or limiting access to compute or by building guardrails into hardware.

With compute governance gaining traction because of advanced AI systems, compute thresholds, i.e., numerical measures of computing power, are also being set legally, which helps distinguish AI systems with high capabilities from other AI systems.

For instance, U.S. Executive Order 14110 requires models using computing power greater than 10^{26} integer and models using biological sequence data and computing power greater than 10^{23} integer to provide the government with information and reports on the models testing and security on an ongoing basis.

Similarly, under the EU AI Act, GPAI is presumed to have high-impact capabilities when cumulative compute used for training is greater than 10^{25} floating-point operations. When a model meets this threshold, the provider must notify the Commission, as meeting the threshold leads to the presumption that this is a GPAI system with systemic risk. This means the model can have a significant impact on the internal market, and actual or reasonably foreseeable negative effects on health, safety, fundamental rights or society. Providers need to comply with requirements on model evaluation, adversarial testing, assessing and mitigating systemic risks, and reporting any serious incidents.

Implementing AI governance

The organizational practices for security and robustness discussed in this report, such as red teaming for adversarial testing, HITL and privacy-preserving technologies, can apply to AI safety. Similarly, organizational practices and laws requiring transparency and explainability, specifically watermarks, also apply to AI safety.

Prompt engineering

One of OpenAI's safety practices includes prompt engineering to help generative AI understand prompts in a given context. This practice is aimed at minimizing harmful and undesired outputs from generative AI, and it helps developers exercise more control over user interactions with AI to reduce misuse at the user level. Moreover, as part of product safety standards, OpenAI also has put in place usage policies.

Reports and complaints

Another safety practice of OpenAI is allowing users to report issues that can be monitored and responded to by human operators. This is not yet a popular practice. A 2023 study carried out by TrustibleAI found out of 100 random organizations, three provided an individual appeals process between the individual and the company. It is possible internal governance and complaint mechanisms may become more common post-EU AI Act, given that, under Article 27 (f), deployers of AI systems must carry out FRIAs of internal governance and complaint mechanisms where a risk has materialized into a harm.

Safety by design

To combat abusive AI-generated content, Microsoft is focused on building strong safety architecture through the safety by design approach, which can be applied at the AI platform, model and application levels. Some efforts include red teaming, preemptive classifiers, blocking abusive prompts, automated testing and rapid bans of users who abuse the system. With regard to balancing freedom of speech against abusive content, Microsoft is also committed to identifying and removing deceptive and abusive content on LinkedIn, Microsoft Gaming Network and other services.

> “Humans control AI, not the other way around. Generative AI models drift. The only way for companies to know when/how they are drifting is to continuously test, monitor and audit the AI applications for high risk use cases- every second of every minute of every day. This is the only way to ensure that the model output comports with the organization’s pre-installed guardrails for accuracy, health and safety, privacy, bias.
>
> **Dominique Shelton Leipzig**
> Mayer Brown Partner, Cybersecurity & Data Privacy and Leader, Global Data Innovation & AdTech

Safety policies

In preparation for the U.K. AI Safety Summit, Meta released an overview of its AI safety policies, specifically in relation to its generative AI Llama model. In addition to model evaluations and red-team analysis, the policy also detailed Meta's model reporting and sharing, reporting structure for vulnerabilities found after model release, post-deployment monitoring for patterns of misuse, identifiers of AI generated material, data input controls and audits, and priority research on societal, safety and security risks.

Industry best practices

Partnership on AI has invested extensively in AI safety research and resources. Some of its work includes Guidance for Safe Foundation Model Deployment. This framework is a living document targeted at model providers on ways to operationalize AI safety for responsible deployment. The framework provides custom guidance providers of foundation models can follow throughout the deployment process that is appropriate for their model's capabilities. Another resource is PAI's SafeLife, which is a benchmark focused on avoiding negative side effects in complex environments. SafeLife is a reinforcement learning environment that tests the "safety of reinforcement learning agents and the algorithms that train them." It allows agents to navigate a complex environment to accomplish a primary task. The aim is to create a "space for comparisons and improving techniques for training non-destructive agents."

Part VIII. The copyright challenge

Generative AI is raising new challenges for copyright law.

Copyright refers to the rights that creators have over the expression of their artistic or intellectual works. Although it is not possible to provide an exhaustive list of "works" covered by copyright legislation, globally copyright protection has been extended to include a wide range of works, such as literature, music, architecture and film. In the context of modern technology, computer software programs, e-books, online journal publications and the content of websites such as news reports and databases are also copyrightable.

> **The clear establishment of intellectual property rights around both inputs and outputs for generative AI models is of crucial importance to creative artists and the creative industries. In the face of dramatically growing machine capabilities, we need to make sure that incentives for human creation remain strong."**
>
> Lord Tim Clement-Jones
> U.K. House of Lords Liberal Democrat Peer and
> Spokesperson for Science, Innovation and Technology

Law and policy considerations

In most countries, and especially those party to the Berne Convention, copyright protection is obtained automatically upon creation of the work. In other words, copyright registration is not necessary for proprietarily safeguarding artistic and intellectual works. Regardless, while offering automatic copyright protection, many countries, including the U.S., also allow voluntary copyright registration.

Copyright provides owners two types of rights: economic rights, through which the owner can make financial gains by authorizing use of their work by others through a license, and moral rights, which include noneconomic interests such as the right to claim authorship of a work or to oppose changes to a work that could harm the owner's reputation.

Copyright protects artistic and intellectual works by preventing others from copying, adapting, distributing, performing or publicly displaying the work, or creating derivative works. When an individual does any of these without the authorization of the rights' owner, this may constitute copyright infringement.

The use of copyright protected content requires the authorization of the original author, unless a statutory copyright exception applies. A legitimate exception to copyright infringement in some jurisdictions is fair use or fair dealing. This is a limitation on the exclusive rights of a copyright holder, which sometimes allows the use of the work without the right holder's permission.

In the U.S., fair use is statutorily defined under 17 U.S. Code § 107, and four factors assist courts in making a fair-use determination. These include purpose and character of use, nature of copyrighted work, substantiality of use, and impact of use on the potential market of the copyrighted work. Similarly, Singapore's Copyright Act of 2021 also includes a fair-use exemption and takes into account the same four factors as the U.S. courts. Singapore's old copyright law also had a fifth factor, which considered the possibility of obtaining a work within a reasonable time at an ordinary commercial price. However, under the new law, the fifth factor may be considered by courts only when relevant.

The U.K. also has a permitted exemption to copyright infringement termed fair dealing. There is no statutory definition for fair dealing as, depending on the case, it will always be a matter of fact, degree and impression. Other factors the U.K. courts previously considered to determine fair dealing include the effect on the market for the original work and whether the amount of work copied was reasonable and appropriate.

Common remedies that can be granted by a court ruling on copyright infringement include injunctions, damages for the loss suffered, statutory damages, infringer's profits, surrender or destruction of infringing articles, and attorney fees and costs.

> “Though copyright has emerged as one of the first and foremost frontiers between AI and intellectual property, the full gamut of IP rights are engaged by AI, and specifically generative AI: design rights, performers’ rights, patents and trademarks. Anthropocentric approaches to IP will butt up against AI’s learning techniques, its scale and the nature of its outputs, leaving much uncertainty, complexity and variety in the implementation of AI and IP governance.
>
> **Joe Jones**
> IAPP Director of Research and Insights

→ SPOTLIGHT
Generative AI copyright litigation in the U.S.

Two main lines of argument are emerging in ongoing AI copyright litigation in the U.S.

Petitioners are arguing that:

- → Defendants made copies of copyrighted works when ingesting them for training foundation models.
- → As the generated outputs were trained on copyrighted material, the outputs themselves are also infringing derivative works.

More specifically, in a lawsuit against OpenAI, the New York Times argued that OpenAI and Microsoft's generative AI tools were built by copying years of journalistic work without permission or payment, and both companies are making high profits through their generative AI tools, which now compete with the news outlet as reliable sources of information.

OpenAI's motion to dismiss the lawsuit provides background on fair use law, and it argues courts have historically used fair use to protect useful innovations and copyright is not a veto right over transformative technologies that leverage existing work internally.

The assessment of fair use is likely to include an evaluation of exactly what was or is being copied, whether ingestion of copyrighted material amounts to transformative use, the substantiality of the copying and the economic harm caused by using copyrighted material in developing generative AI models on the potential market for the copyrighted work.

Similarly, in Tremblay v. OpenAI, various authors alleged copyright infringement based on the ingestion of training data that copied the works of the authors without consent, credit or compensation. A California court recently rejected claims on vicarious copyright infringements, Digital Millennium Copyright Act violations, negligence and unjust enrichment.

INDUSTRY EXAMPLE

Microsoft committed to absolve its users of liability by assuming vicarious responsibility for infringements when use of its Copilot service leads to legal issues for their commercial customers. Microsoft also requires customers to use content filters and other built-in product safety features, and asks customers to not generate infringing content, such as by providing Copilot with inputs they do not have the right to use.

Implementing AI governance

Numerous copyright-safety solutions and harm-mitigation strategies are emerging, notwithstanding the uncertainty present due to pending litigation.

→ **Opt outs.** As foundation models are trained on vast amounts of data online, organizations may not be aware that their copyrighted material is used for training. In those scenarios, when organizations are concerned about their webpages being scraped, an opt-out process, like that of OpenAI, may be a workable strategy to mitigate the risk of unwanted scraping.

→ **Liability considerations.** Given the fear of potentially becoming a copyright infringer as a user of generative AI, commercial users may avoid engaging with providers of generative AI services.

→ **Explore technical guardrails.** Organizations can also make use of technical guardrails that help them respect the copyrights of authors. Microsoft incorporated guardrails such as content filters, operational monitoring, classifiers, abuse detection and other technologies to reduce the likelihood of Copilot returning copyright-infringing content.

→ **Generative AI requirements.** To increase transparency around data used to train generative AI models, including copyrighted data, certain jurisdictions such as the EU require system providers to publish detailed summaries of the content used for training their models. Further, with respect to copyright compliance, the EU AI Act requires providers to implement a copyright policy mandating protocols to identify and observe applicable copyright laws.

Part IX. Third-party AI assurance

AI assurance methods are crucial for demonstrating accountability and establishing trust.

In a recent report released by the U.K. government, assurance is defined as "the process of measuring, evaluating and communicating something about a system or process, documentation, a product, or an organisation." Many of the AI governance implementation mechanisms discussed in this report are forms of assurance.

While establishing core competencies within an organization is beneficial to create strong AI-governance foundations across the different lines of defense, utilization of third-party AI assurance mechanisms may be an important or necessary consideration depending on the type of AI used and the organization's knowledge and capacity.

Integrating third-party assurance into an AI-governance strategy is a consideration at various stages of the life cycle.

Types of third-party assurance

Some of the most practical tools for the realization of safe and responsible AI are emerging from third-party AI assurance methods.

Assessment

Assessments are key mechanisms to evaluate various aspects of an AI system, including to determine the risk of a system or identify the source of bias or determine the reason a system is making inaccurate predictions. Various services and off-the-shelf products can be integrated into AI governance practices based on what an organization is trying to determine from its assessment.

Certain assessments must be conducted by the third party providing the system to their customers, such as conformity assessments and impact assessments focusing on the impacts of the datasets used and the model itself. From a deployer's perspective, third-party due diligence enquiries should be integrated into the organization's existing third-party risk management program and include screening at both the vendor enterprise and product levels.

Testing and validation

Testing techniques such as statistical tests to evaluate demographic fairness, assess system performance or detect generative AI that may lead to copyright breaches are becoming widely available through various third-party vendors. Before choosing a vendor, it is important to have a clear understanding of what the test is for and whether the context — which includes the type of AI used, applicable jurisdictions and the domain operating in — will impact the types of tests to run.

Conformity assessments

Conformity assessments are reviews completed by internal or external review functions to evaluate whether a product, system, process or individual adheres to an established set of requirements. This is typically performed in advance of a product or system being placed on the market. While most assessments focus on evaluating aspects of AI systems, conformity assessments have been designed to evaluate quality-management systems, a set of processes for those who build and deploy AI systems, and individuals who are involved in the development, management or auditing of AI systems.

From a deployer's perspective, the third-party due diligence process should include vendor inquiries into product documentation, such as technical specifications, user guides, conformity assessments and impact assessments.

> **Risk assessments should be done at several phases of development, starting with the proposal/idea phase.**
>
> **It's easier to incorporate some 'responsible by design' features early on, rather than tack them on at the end. For example, filtering for toxic content in your training data, before a model is trained, can be more effective than trying to catch toxic generated content afterwards.**
>
> **In contrast, a full impact assessment should be done once a model is fully developed and evaluated, because it's hard to assess the impact without a lot of information about the final system.**
>
> Andrew Gamino-Cheong
> Trustible AI Co-founder and Chief Technology Officer

> **Organizations need a clear understanding of how AI risk will affect their business through third-party relationships. They should proactively review their inventory of vendors and identify those that provide AI solutions or components. They also need to be aware of the development plans for all third-party products, including whether, how, and when AI will be integrated. With that understanding, partners, vendors and their products may need to be reassessed to account for AI risk with updated due diligence processes.**
>
> Amber Gosney
> FTI Technology Managing Director

Impact assessments

The risk profile of AI systems can vary widely based on the technical capabilities and intended purposes of the system, as well as the particular context of their implementation. Evaluating and mitigating the impacts of an AI system is therefore a shared responsibility that must be owned by providers and deployers alike in practice. The organization deploying a third-party AI system will have a closer understanding of the specific context and impacts of deploying the system. Similarly, the third-party vendor is best placed to evaluate the impacts of the training, testing and validation datasets, the model and infrastructure used to design and develop the system.

AI/algorithmic auditing

While there is not yet a formal audit practice as seen in financial services, there is a growing call for those who audit AI systems to demonstrate a common set of competencies, such as with a certification or formal designation. These audits may incorporate other third-party mechanisms discussed above to evaluate AI systems and ensure they are safe, secure, legally compliant and meet requisite standards, among other things. The National Telecommunications and Information Administration released recommendations for federal agencies to use audit and auditors for the use of high-risk AI systems.

Canada's proposed Bill C-27, the Digital Charter Implementation Act, identifies that the Minister of Innovation, Science and Industry can issue an independent audit if they have reasonable grounds to believe requirements outlined in the act have not been met. This may encourage organizations to ensure compliance via preventative third-party audits. Additionally, Canada identified the importance of international standards to help support the desired objectives of the act.

Certifications

Certifications are marks or declarations provided after evaluations or audits are performed against standards or conformity assessments. The mark indicates the AI system adheres to certain specified requirements. It is important to note certifications can also be provided to quality-management systems used throughout the life cycle of an AI system or to individuals, demonstrating that they met a set of competencies.

→ SPOTLIGHT
Algorithmic audits as airplane cockpits

At ORCAA, we use the analogy of an airplane cockpit to talk about algorithmic audits. In an airplane cockpit, the dials and gauges take measurements that relate to possible failure modes.

For instance, the fuel gauge says if the plane is about to run out of gas, and the attitude indicator says if it is going to dive or roll. These dials have 'redlines': threshold values that, if exceeded, mean the pilot needs to intervene. The auditor's job is to design a 'cockpit' for a given algorithmic system. This involves identifying failure modes — how the system could result in harm to various stakeholders — and building 'dials' that measure conditions that lead to failures. At ORCAA, we have developed frameworks for doing these critical tasks.

Some other aspects of this analogy are worth noting. A cockpit identifies problems but does not fix them. An indicator light will say an engine is out, but it won't say how to repair or restart the engine. Likewise, an algorithmic cockpit should indicate when a failure is imminent, but it is the job of the system deployer, the 'pilot,' to intervene. A cockpit is a critical piece of airplane safety, but it's not the whole picture. Planes are tested extensively before being put into service, both during the design phase and when they roll off the assembly line and are periodically taken out of service for regular inspections and maintenance.

Likewise, algorithmic cockpits, which are critical for safety while the system is deployed, should be complemented by predeployment testing and regular inspections and maintenance during deployment."

Cathy O'Neil
O'Neil Risk Consulting & Algorithmic Auditing CEO

Conclusion

Bringing it all together and putting it into action.

Organizations may seek to leverage existing organizational risk frameworks to tackle AI risk at enterprise, product and operational levels. Tailoring their approach to AI governance to their specific AI product risks, business needs and broader strategic objectives can help organizations establish the building blocks of trustworthy and responsible AI. A key goal of the AI governance program is to facilitate responsible innovation. Flexibly adapting existing governance processes can help businesses to move forward with exploring the disruptive competitive opportunities that AI technologies present, while minimizing associated financial, operational and reputational risks.

Contacts

Connect with the team

Uzma Chaudhry
IAPP AI Governance Center
Research Fellow
uchaudhry@iapp.org

Joe Jones
IAPP Director of Research
and Insights
jjones@iapp.org

Ashley Casovan
IAPP AI Governance Center
Managing Director
acasovan@iapp.org

Lynsey Burke
IAPP Research and Insights
Project Specialist
lburke@iapp.org

Nina Bryant
FTI Technology Senior
Managing Director
nina.bryant@fticonsulting.com

Luisa Resmerita
FTI Technology
Senior Director
luisa.resmerita@fticonsulting.com

Michael Spadea
FTI Technology Senior
Managing Director
micheal.spadea@fticonsulting.com

Follow the IAPP on social media

Published June 2024.