An official website of the United States Government Here's how you know

Newsroom   Business   Employees   Job Seekers   Students   Travelers   Visas

POLICY ISSUES   COUNTRIES & AREAS   BUREAUS & OFFICES   ABOUT

Risk Management Profile for Artificial Intelligence and Human Rights

BUREAU OF CYBERSPACE AND DIGITAL POLICY

JULY 25, 2024

Executive Summary:

The U.S. Department of State is releasing a "Risk Management Profile for Artificial Intelligence and Human Rights" (the "Profile") as a practical guide for organizations—including governments, the private sector, and civil society—to design, develop, deploy, use, and govern AI in a manner consistent with respect for international human rights.[1] When used in a rights-respecting manner, AI can propel technological advances that benefit societies and individuals, including by facilitating enjoyment of human rights. However, AI can be applied in ways that infringe on human rights unintentionally, such as through biased or inaccurate outputs from AI models. AI can also be intentionally misused to infringe on human rights, such as for mass surveillance and censorship. International human rights are uniquely valuable to AI risk management because they provide an internationally recognized, universally applicable normative basis for assessing the impacts of technology. However, human rights are not always familiar to those involved in AI design, development, deployment, and use, and there is a gap in translating human rights concepts for technologists.

The Profile aims to bridge the gap between human rights and risk management approaches, demonstrating how actions related to assessing,

addressing, and mitigating human rights risks fit naturally into other risk management practices. This Profile is anchored in international human rights and can serve as a common tool for stakeholders across sectors and around the world to increase their capacity to engage in AI risk management practices that promote enjoyment of human rights. It demonstrates how AI risk management processes from the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework ("AI RMF")—a voluntary framework intended to help AI actors throughout the AI lifecycle work together to make AI systems safe, secure, and trustworthy—provides a selection of actions that organizations can draw on as part of their comprehensive human rights due diligence processes.[2] Conversely, the Profile also shows how some human rights-related actions can be taken as part of implementing the AI RMF's four organizational functions. These functions are: 1) **Govern** (set up institutional structures and processes), 2) **Map** (understand context and identify risks), 3) **Measure** (assess and monitor risks and impacts), and 4) **Manage** (prioritize, prevent, and respond to incidents). The Profile can be applied across applications, stakeholders, and sectors, and throughout the AI lifecycle.

Section I outlines the rationale, scope, and function of this Profile. Section II provides analysis of some potential human rights impacts of AI design, development, deployment, and use. Section III provides a selection of recommended practices for incorporating human rights considerations into risk management practices from the four organizational functions of the AI RMF, and draws practices from international human rights documents.[3]

Section I: Why International Human Rights Matter for AI Governance

Rationale:

International human rights are uniquely well-placed to serve as a normative foundation for AI risk management, for three key reasons:

1) International human rights are universally applicable and already function as a shared international language to enable

effective due diligence and technology governance.

The Universal Declaration of Human Rights (UDHR) and international human rights law have a distinctive status internationally. International human rights law is referenced in multiple technology-related resolutions at the UN and can play an important role in AI governance around the world. In March 2024, all 193 United Nations member states affirmed this in adopting, by consensus, a resolution that emphasizes "that human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of artificial intelligence systems," and "calls upon all Member States and, where applicable, other stakeholders to refrain from or cease the use of artificial intelligence systems that are impossible to operate in compliance with international human rights law or that pose undue risks to the enjoyment of human rights."[4]

2) Human rights commitments are relevant to both governments and private sector actors, who play significant roles in AI design, development, deployment, use and governance.

Governments have obligations to protect human rights. The private sector has the responsibility to respect human rights and to engage in due diligence to ensure that their products and services do not infringe on human rights. The UN Guiding Principles on Business and Human Rights (UNGPs), which the UN Human Rights Council unanimously endorsed in 2011, provide both foundational and operational principles that can guide private sector companies in conducting human rights due diligence processes, which are also applicable to AI systems. To date, however, AI-specific guidance on applying human rights due diligence has been under-developed, and AI-specific normative frameworks have not relied heavily on the work from the human rights due diligence community.

3) Many risks posed by AI are related to human rights.

A broad spectrum of risks associated with AI have bearing on the enjoyment and exercise of human rights, including but not limited to privacy, equal protection under the law, freedom of opinion and expression, and freedom of peaceful assembly and association.[5] These risks will be outlined in greater detail in Section II.

Function:

The chief function of the Profile is to provide non-exhaustive, non-binding guidance on how organizations can utilize NIST's AI RMF to manage the risks of AI technologies related to human rights, throughout the AI lifecycle and in a context-specific manner.

The Profile has two key interrelated goals:

1) Show AI designers, developers, deployers, and users how to apply NIST's AI Risk Management Framework to contribute to human rights due diligence practices.

The NIST AI RMF provides a detailed framework of methods for managing risk throughout the AI lifecycle. This Profile complements the AI RMF by demonstrating how human rights can be considered when applying the AI RMF. By referencing universally applicable, internationally recognized human rights, the Profile provides a globally relevant normative basis for the AI RMF's recommended risk management actions. It also offers concrete suggestions based on the AI RMF on how to address human rights-related risks throughout the AI lifecycle.

2) Facilitate rights-respecting AI governance throughout AI design, development, deployment, and use by all stakeholders.

Part of the aim of this Profile is to develop a common language for AI developers, international policymakers, and civil society by linking the AI RMF with a universally applicable normative framework. This Profile can serve as a shared resource to bridge different stakeholder and disciplinary communities by illustrating for technologists how human rights considerations can be integrated into AI risk management and illustrating for policymakers and civil society actors how human rights-related risks can be identified, addressed, and mitigated when organizations are utilizing NIST's AI RMF in support of rights-respecting AI governance approaches.

Section II: Potential Human Rights Impacts of AI Systems

The design, development, deployment, and use of AI technologies can impact human rights in multiple ways – through both unintentional and intentional effects. Responsible use of AI can bring benefits including increased accessibility, enhanced detection of potential human rights harms, and support in achieving the UN's Sustainable Development Goals. On the other hand, misuse of AI can have detrimental impacts in many sectors including but not limited to criminal justice, immigration, finance, welfare, healthcare, education, and human resources. AI's misuse can infringe on human rights by facilitating arbitrary surveillance, enabling censorship and control of the information realm, or by entrenching bias and discrimination. Even if AI is designed and developed to be rights-respecting, misuse by the end user may result in the system being used to infringe on human rights.[6]

Table 1 contains a non-exhaustive list of examples identified in multistakeholder consultations as potential risks related to human rights, including privacy, equal protection under the law, freedom of opinion and expression, and freedom of peaceful assembly and association. Risks related to these human rights can arise throughout the AI lifecycle both as intended and unintended consequences of AI actors' actions. The table focuses on issues that are likely to apply across sectors and connects them to AI lifecycle stages identified in the AI RMF.[7] The table also identifies real-world impacts that can occur later in the AI lifecycle—generally during the deployment and use stages—if these issues are not addressed at earlier points.

Table 1: Potential Human Rights Risks from AI

AI Lifecycle Stages	Example Actions or Lapses that Can Pose Risks to Human Rights	Example Resulting Impacts
Plan and Design	System operators and AI designers plan or design systems without considering harms from potential failure modes and/or from foreseeable applications beyond systems' intended use(s).	AI models and applications may be used in ways that lead to discriminatory, unsafe, or other harmful outcomes, possibly via use cases beyond developers' intentions. For example, an image preprocessing tool that fails to properly handle darker skin tones could introduce harmful bias, or a system designed for monitoring vehicle or person movements could be abused for privacy intrusions or harassment.
	Operators and designers knowingly	Impacts could include unlawful

design a system, or select data and algorithms for a system, such that it will contravene human rights by design, e.g., by enabling arbitrary and unlawful surveillance.

and/or inaccurate surveillance and tracking—including inappropriate algorithmic management and AI-enabled workplace monitoring—wrongful arrest and detention, and harms from synthetic child sexual abuse material and non-consensual intimate imagery. These could additionally lead to chilling effects on freedom of expression and freedom of peaceful assembly and association.

Collect and Process Data

Data that was collected or scraped is used to train AI models, reused for other applications, or sold without users' knowledge. e.g., authorities or companies sell or otherwise transfer biometric data to other countries or private companies.

The accuracy of datasets used to train AI models is not adequately verified prior to use.

Datasets are not constructed to be adequately representative of race, gender, other personal characteristics, or cultural dimensions (including language).

Datasets draw upon biased content or real-world events, potentially reinforcing structural inequities or harmful stereotypes.

AI models may have higher error rates or fail to provide benefits for individuals who have characteristics not well-represented in training data, or from inaccurate data. If not adequately addressed, flawed model outputs could lead to unjustified arrests, denial of welfare benefits, denial of credit, or other harmful outcomes.

Specifically, discriminatory impacts on individuals who are either underrepresented or not accurately represented in training data can include:

In the criminal justice sector, for example, data may be used to train AI models whose use can enable predictive inferences that can perpetuate preexisting patterns of discrimination, including racial and ethnic profiling. Systems may produce false positives (false identifications) for individuals, which may lead to wrongful arrests.

Inclusion of harmful and biased stereotypes in datasets may result in the perpetuation of harmful stereotypes based on race, color, sex, gender, language, religion or belief, political or other opinion, national or social origin, property, or birth or other status, which can exacerbate

which can exacerbate discrimination and existing socio-economic inequalities.[8]

Impacts from inaccurate, non-representative, or otherwise harmful data may have chilling effects on individuals or groups who distrust AI systems' accuracy and efficacy, or who fear being tracked, targeted, or monitored for expressing their opinions in public spaces.

Build and Use Model; Verify and Validate	System designers implement insufficient technical safeguards to prevent data leakage, unauthorized disclosure, or de-anonymization of personally identifiable (PII), or other sensitive data such as biometric, health, or location data.	AI tools can infer PII that may violate privacy, including sensitive attributes such as location, gender, age, sexual orientation, and political beliefs. Users can be re-identified, possibly by combining anonymized data with other data points, and can be tracked across physical locations and online spaces. Users could extract training data from models, which may allow reconstructing individuals' PII without their consent.
	Developers or other actors fail to conduct testing and evaluation that detects inaccurate outputs, including biased responses or confabulations.	AI models may fail to perform their intended function, causing harm to those whose enjoyment of their human rights would rely on that functionality (e.g., to translate documents related to asylum, or determining consequential life-impacting decisions).
	The interaction between AI models and humans is configured such that the models' answers are acted upon or provided to users without sufficient further examination, including because explanations or justifications for "black-box" system outputs are not provided and	Services or resources may be denied or revoked based on unjustified decisions or inaccurate data. It can be difficult for those affected by AI outputs or decisions to identify when and how they have been harmed, decreasing accountability and the ability to mitigate or remediate harms.

evaluated.

Deploy and Use	Organizations release or deploy a new AI model for which one of the above actions or lapses remains unaddressed, or they release it without guidance and safeguards on acceptable and responsible uses. End users circumvent system guardrails to use the system to violate or abuse human rights.	The risks highlighted above may come to fruition. AI models may be misused to generate non-consensual intimate imagery (NCII) or child sexual abuse material (CSAM), which can be used to victimize individuals, or to enable unlawful surveillance or censorship. AI-enabled disinformation can be used to harass and intimidate actors such as journalists, political opponents, or human rights defenders into self-censorship.[9] This can undermine the ability to exercise the freedom to seek and receive information necessary to form opinions as well as freedom of peaceful assembly and association.[10]
Operate and Monitor	Organizations do not provide an avenue for people to report and access remedy for abuses, errors, or incidents with the system.	AI systems may continue amplifying inequities by continuing to produce inaccurate outputs that result in increased targeting of marginalized populations. AI systems may be used for technology-facilitated gender-based violence in ways that cause individuals to retreat from civic spaces. Discriminatory or otherwise unjust decisions may go unaddressed and be repeated in the future. Victims may not be able to access remedy after their human rights are violated or abused.

Section III: Recommended Actions to Help Address Human Rights Risks, Based on the AI RMF

To support efforts to protect and/or promote respect for human rights, this section provides examples of actions AI actors can take, using their leverage and resources, to incorporate human rights considerations throughout their risk management processes.[11] Organizations should use their resources and leverage to promote these actions, which help to identify, prevent, mitigate, and remedy human rights risks. It provides recommendations derived from the AI RMF, which outlines the four organizational functions "**Govern**," "**Map**," "**Measure**," and "**Manage**," to manage the risks identified across the lifecycle stages above.

This Profile provides references to specific AI RMF subcategories to help implement recommended actions. Actions are also suggested below based on business and human rights (BHR) practices from documents such as the UN Guiding Principles on Business and Human Rights (UNGPs) and the associated UN B-Tech Project, UNESCO's Recommendation on the Ethics of AI, and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (OECD MNE Guidelines).

NIST AI RMF Function: GOVERN: *Set up policies, procedures, and institutional structures to align operations with societal values, organizational values, and legal requirements and to foster a culture of risk management.*[12]

Examples of Recommended Human Rights-Related Actions Under **GOVERN**:

Issue publicly available policies regarding AI activities and human rights. Government agencies and departments can have publicly available policies regarding how they will protect human rights in the context of

their AI activities. Businesses can have a publicly available policy commitment to respect human rights in their AI activities, including to use their resources and leverage to identify, prevent, mitigate and remedy human rights risks in line with the expectations set out in the UNGPs and the OECD MNE Guidelines.[13] These policy commitments can exist alongside and/or be integrated into enterprise risk management systems, particularly organizational policies and practices that seek to minimize potential negative impacts of AI design, development, deployment, and use (**Govern 4.1**).

Establish or refine processes that make clear how they evaluate human rights risks that can emerge across the AI value chain (**Govern 1.1, Govern 1.2**), how they incorporate human rights considerations into risk mapping and stakeholder consultations (**Govern 3.1**), and how they document the results.[14] Table 1 can be used as a starting point.

Establish and incorporate algorithmic impact assessments, privacy impact assessments, and human rights due diligence processes as part of their organizational risk management processes (**Govern 1.4**). As reflected in the UNGPs, businesses should set up procedures for human rights due diligence, including assessing actual and potential human rights impacts, integrating and acting upon the findings, and tracking outcomes, where more significant risks are prioritized. This includes establishing access to remedy in the event of adverse impacts. [15] As reflected in the OECD MNE Guidelines, businesses' policies and procedures should include preventing or mitigating adverse human rights impacts that are directly linked to their business operations, products or services by a business relationship, even if they do not contribute to those impacts.[16]

In training policies for staff working on AI systems throughout the AI lifecycle (**Govern 2.2**), include considerations related to human rights from AI-related design, development, and deployment activities. For example, include guidance from human rights-focused organizations on how to identify risk factors to the enjoyment of human rights (such as the issues listed in Table 1). Organizations can require that people developing, deploying, and using AI systems have sufficient training in detecting and mitigating potential harmful bias.

When establishing policies and procedures to address AI risks associated with third-party entities (**Govern 6.1**), include risks of infringement on human rights that can arise in connection with the purchase and integration of data or AI systems from third-party

vendors, such as risks to privacy and gaps in transparency.[17]

Publicize information about how impacts are addressed, with an emphasis on how they will communicate about impacts in a transparent manner (**Govern 4.2**).

Processes and procedures for decommissioning and phasing out AI systems safely (**Govern 1.7**) should include procedures for cases where feedback channels have revealed that an AI system is impacting rights and safety in unacceptable ways.

NIST AI RMF Function: MAP: *Establish and understand the context in which risks might materialize. Characterize potential impacts, with input from many perspectives. Inform an initial go/no-go decision about whether to pursue a given AI solution.*[18]

Examples of Recommended Human Rights-Related Actions Under **MAP**:

Prior to designing, and during the development and deployment of an AI system, conduct consultations at regular intervals with diverse internal teams, as well as external collaborators, end users, and communities that could potentially be impacted (**Map 1.2**). Stakeholder consultations should involve civil society and especially affected stakeholders to learn about risks, impacts, challenges, and opportunities to advance meaningful AI risk assessment and mitigations.[19] Consultations can help with all of the MAP tasks below.

The intended purpose for the AI system should be well-specified and finite (**Map 1.1, Map 3.3**). Analyze whether the AI system can provide a net benefit for that purpose, accounting for both potential benefits (**Map 3.1**) and costs (**Map 3.2**) of the system. Where feasible, this analysis should be supported by specific metrics or quantitative analysis. Where quantification is not feasible, qualitative analysis should demonstrate an expected positive outcome.[20]

As part of this analysis, demonstrate that the AI system is on net better suited to accomplish the relevant task than alternative strategies that pose fewer risks related to human rights. Among those alternatives, consider solutions that do not involve AI technologies. Also consider and document what risk avoidance or reduction measures would be necessary to make the AI system a better choice.

When analyzing and documenting potential uses and impacts (**Map 1.1, Map 3.1, Map 3.2**), include unintended downstream harms that may arise, such as infringements on privacy from data collected without consent or data re-use, or chilling effects on freedom of expression or freedom of peaceful assembly and association upon individuals or members of groups. As reflected in the OECD MNE Guidelines, businesses should "conduct due diligence commensurate to the severity and likelihood of the adverse impact. When the likelihood and severity of an adverse impact is high, then due diligence should be more extensive. Due diligence should also be adapted to the nature of the adverse impact on responsible business conduct issues, such as human rights, the environment and corruption. This involves tailoring approaches for specific risks and taking into account how these risks affect different groups."[21]

Establish channels to integrate and document feedback about positive, negative, and unanticipated impacts related to human rights in consultation with civil society and impacted communities or users (**Map 5.2**).

As part of understanding the context of an AI system's deployment (**Map 1.1**), determine how the system could impact individuals, groups, and societies, and could run counter to international human rights and the UNGPs as applicable.

Assess the likelihood and magnitude of known and foreseeable negative impacts and limitations related to both intended and unintended uses of an AI system (**Map 5.1**), including potential infringements upon human rights. Consider context-specific deployment environments and how they could lead to different sets of risks (e.g., risks created in conflict settings).

When documenting potential beneficial uses and impacts (**Map 1.1**), include unintended downstream harms that may arise, such as privacy harms from data collected without consent, data re-use, or chilling effects on freedom of expression or freedom of peaceful assembly and association upon individuals or members of groups.

Developers can define system requirements that can promote respect for human rights (e.g., "the system shall respect the privacy of its users," "the system shall not be trained on non-representative datasets,"), drawing on input from those who might be affected by the AI systems' behavior. Developers can then make decisions in a way that accounts for these requirements and other human rights implications (**Map 1.6**),

and obtain agreement from users that they will follow developer-set requirements.

NIST AI RMF Function: MEASURE: *Employ quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts.*[22]

Examples of Recommended Human Rights-Related Actions Under **MEASURE**:

Using examples in Table 1 as a starting point, identify indicators of human rights related risks, including metrics that assess and account for errors and impacts for individuals, groups, and societies (**Measure 1.1**).

Update these metrics as needed, including to account for feedback about impacts (**Measure 1.2**).

Where quantitative metrics are difficult to establish, more qualitative forms of impact assessment can be used (**Measure 3.2**).

Impact assessments should identify impacts on human rights and fundamental freedoms, in particular but not limited to the rights of marginalized individuals or individuals in vulnerable situations, as well as impacts on gender equality, labor rights, the environment and ecosystems and ethical and social implications, and citizen participation.[23]

Regularly apply the metrics and impact assessment methods to AI systems to determine their risk level, including for human rights risks related to safety (**Measure 2.6**), security and resilience (**Measure 2.7**), transparency and accountability (**Measure 2.8**), privacy (**Measure 2.10**), and fairness and bias (**Measure 2.11**). This should be done in consultation with internal experts who did not serve as front-line developers for the system and/or independent assessors, external AI actors and affected parties (**Measure 1.3**).

NIST AI RMF Function: MANAGE: *Prioritize and address risks based on projected impact, with plans to prevent, respond to, recover from, and communicate about incidents or events.*[24]

Examples of Human Rights-Related Actions Under **MANAGE**:

Prioritize treatment of documented AI risks based on likelihood, available resources or methods, and its impact (**Manage 1.2**). As reflected in the OECD MNE Guidelines, "Where it is not feasible to address all identified impacts at once, an enterprise should prioritize the order in which it takes action based on the severity and likelihood of the adverse impact."[25]

Ensure that salient human rights-related risks are included among high-priority risks to respond to (**Manage 1.3**), and that any residual risks do not include undue risks related to human rights.

Publicly communicate incidents and errors that can impact human rights, including but not limited to privacy, and downstream effects on freedom of expression and freedom of peaceful assembly and association, including to affected communities. Follow and document processes for tracking, responding to, and recovering from incidents (**Manage 4.3**).

Establish redress mechanisms and offices that serve as a remedial point of contact for users whose rights have been negatively affected by AI systems (**Manage 4.1**). States should ensure access to judicial and non-judicial remedies where individuals may have been harmed by the development or deployment of AI technologies.[26]

When determining the resources required to manage AI risks and assessing viability of alternative systems (**Manage 2.1**), consider what resources would be required to reduce the magnitude or likelihood of potential impacts to privacy, discrimination, or any downstream chilling effects on rights such as freedom of expression or freedom of association and peaceful assembly. This includes considering the resources that would be needed to mitigate risks to privacy, such as through use of privacy-enhancing technologies (PETs).

When analyzing the potential human rights impacts of an AI system, AI actors should seek to maximize benefits to individuals and groups whenever possible and consider both the likelihood and magnitude of both human rights harms and human rights benefits. Well-tailored and well-governed AI systems can accelerate access to effective remedy (e.g., by helping to identify harms), reduce biases in human decision-making (e.g., by suggesting

considerations they may have overlooked), or even be used for applications that support respect for human rights by design (e.g., identifying indicators of forced labor).

Conclusion:

By following the recommended actions in this Risk Management Profile for AI and Human Rights Profile, AI actors around the world and across sectors can integrate human rights considerations into responsible and rights-respecting AI risk management and governance approaches. Integrating these considerations into their use of the NIST AI RMF can help such actors develop practices to detect and mitigate potential risks to human rights early on and throughout the AI lifecycle.

Annex: Additional Resources

AI RMF Documents:

1. **NIST AI RMF**

2. **NIST AI RMF Playbook**

3. **Crosswalks to the NIST AI RMF**

Select USG Documents Related to AI and Human Rights:

1. **Blueprint for an AI Bill of Rights**

2. **Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence**

3. **OMB M-Memo: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence**

4. **Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI**

5. **Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI**

6. **Democratic Roadmap: Building Civic Resilience to the Global Digital**

Information Manipulation Challenge

7. **Due Diligence Guidance on Implementing the "UN Guiding Principles" for Transactions Linked to Foreign Government End-Users for Products or Services with Surveillance Capabilities**

8. **The U.S. Government's National Action Plan on Responsible Business Conduct**

9. **Biden-Harris Administration Unveils Critical Steps to Protect Workers from Risks of Artificial Intelligence**

10. **U.S. Department of Labor: Artificial Intelligence and Worker Well-being: Principles for Developers and Employers**

International Bill of Human Rights:

1. **Universal Declaration of Human Rights**

2. **International Covenant on Civil and Political Rights**

3. **International Covenant on Economic, Social and Cultural Rights**

Multilateral Business and Human Rights Principles:

1. **UN Guiding Principles on Business and Human Rights**

2. **OECD Guidelines for MNEs on Responsible Business Conduct (RBC)**

3. **B-Tech Project | OHCHR**

Select Multilateral AI Documents:

1. **UNGA Resolution A/RES/78/265: Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development**

2. **OECD Recommendation of the Council on Artificial Intelligence**

3. **UNESCO Recommendation on the Ethics of Artificial Intelligence**

4. **UNESCO Ethical Impact Assessment**

Select Non-Governmental Resources:

1. **A Taxonomy of Trustworthiness for Artificial Intelligence – UC Berkeley**

2. **AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models – UC Berkeley**

1. This Profile is intended to be complementary to and build upon U.S. government initiatives to safeguard and uphold rights-respecting AI, which, in addition to the NIST AI Risk Management Framework, include the October 2023 Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; the March 2024 Office of Management and Budget's (OMB) Memorandum on Advancing Governance, Innovation, and Risk Management in Federal Agencies' Use of Artificial Intelligence; the July 2023 Voluntary AI Commitments developed in coordination with leading AI companies; and the October 2022 Blueprint for an AI Bill of Rights. The Profile also builds upon existing international AI governance efforts, including the U.S.-sponsored UN resolution, "Seizing the Opportunities of Safe, Secure, and Trustworthy Artificial Intelligence Systems for Sustainable Development," which the UN General Assembly adopted by consensus in March 2024; the Organization for Economic Co-operation and Development (OECD) Recommendation of the Council on Artificial Intelligence; the G7 voluntary code of conduct on AI; and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence. USAID will also release complementary stakeholder-informed implementation guidance specific to the international development context. ↑

2. This Profile is intended to be cross-sectoral. As defined **by NIST**, "AI RMF cross-sectoral profiles cover risks of models or applications that can be used across use cases or sectors. Cross-sectoral profiles can also cover how to govern, map, measure, and manage risks for activities or business processes common across sectors such as the use of large language models, cloud-based services or acquisition."Furthermore, while the scope of risk management is broad and covers areas including financial and reputational risk, this Profile makes the case that human rights risks should be addressed as an important component of organizations' overall risk management strategy. ↑

3. International human rights documents that the Profile draws upon includes the UN Guiding Principles on Business and Human Rights and the associated UN B-Tech Project, the UNESCO Recommendation on the Ethics of AI, and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct. ↑

4. UN, A/RES/78/265 *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*, April 1, 2024 ↑

5. Each of these areas are addressed in the UDHR. Privacy is addressed under Article 12; equal protection and non-discrimination are addressed under Articles 2 and 7; freedom of expression is addressed under Article 19; and freedom of peaceful assembly and association are addressed under Article 20. These five human rights are not intended to be a comprehensive list of the rights potentially impacted by AI. Rather, they are intended as illustrative and have been identified through multistakeholder consultations as important examples that broadly apply across sectors. ↑

6. Companies developing AI systems with surveillance capabilities are encouraged to consult the **Department of State Guidance on Implementing the *UN Guiding Principles* for Transactions Linked to Foreign Government End-Users for Products or Services with Surveillance Capabilities.** ↑

7. The AI lifecycle stages as defined in the NIST AI RMF are: Plan and Design Collect and Process Data Build and Use Model Verify and Validate Deploy and Use Operate and Monitor↑

8. **OHCHR B-Tech Project, Taxonomy of Human Rights Risks Connected to Generative AI, 2024** ↑

9. **OHCHR B-Tech Project, Taxonomy of Human Rights Risks Connected to Generative AI, 2024** ↑

10. **U.S. Department of State, Democratic Roadmap: Building Civic Resilience to the Global Digital Information Manipulation Challenge, March 2024** ↑

11. In line with the UNGPs and the OECD MNE Guidelines, organizations can use their resources and leverage to identify, prevent, mitigate and remedy human rights risks. For example, businesses could encourage subcontractors to incorporate human rights considerations throughout their risk management process. ↑

12. From the AI RMF: "The **GOVERN** function:
    • cultivates and implements a culture of risk management within organizations designing, developing, deploying, evaluating, or acquiring AI systems;
    • outlines processes, documents, and organizational schemes that

anticipate, identify, and manage the risks a system can pose, including to users and others across society – and procedures to achieve those outcomes;

• incorporates processes to assess potential impacts;

• provides a structure by which AI risk management functions can align with organizational principles, policies, and strategic priorities;

• connects technical aspects of AI system design and development to organizational values and principles, and enables organizational practices and competencies for the individuals involved in acquiring, training, deploying, and monitoring such systems; and

• addresses full product lifecycle and associated processes, including legal and other issues concerning use of third-party software or hardware systems and data." See **NIST, NIST AI Risk Management Framework, January 26, 2023** ↑

13. **OECD, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, June 8, 2023** ↑

14. Evaluating human rights risks across the AI value chain should include workers' rights for workers involved in data enrichment. ↑

15. For a more in-depth explanation of the U.S. Government's expectations for businesses on human rights due diligence, see U.S. Department of State, The U.S. Government's National Action Plan on Responsible Business Conduct, p. 7-10, 2024, **U.S. Government's National Action Plan on Responsible Business Conduct.** ↑

16. **OECD, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, June 8, 2023** ↑

17. Related business and human rights practices include:From the OECD MNE Guidelines: IX. Science, Technology and Innovation: "When collecting, sharing and using data, enhance transparency of data access and sharing arrangements, and encourage the adoption, throughout the data value cycle, of responsible data governance practices that meet standards and obligations that are applicable, widely recognized or accepted among Adherents to the Guidelines, including codes of conduct, ethical principles, rules regarding manipulation and coercion of consumers, and privacy and data protection norms. **OECD, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, June 8, 2023** , From B-Tech: "States should enforce laws that are aimed at, or have the effect of, requiring companies developing and deploying generative AI technology to respect human rights, periodically assess the adequacy of such laws and address any gaps. "**OHCHR B-Tech Project,**

Advancing Responsible Development and Deployment of Generative AI, 2024 From the UNGPs: "States should exercise adequate oversight in order to meet their international human rights obligations when they contract with, or legislate for, business enterprises to provide services that may impact upon the enjoyment of human rights." See OHCHR, Guiding Principles on Business and Human Rights: Implementing the United Nations "**Protect, Respect and Remedy" Framework, January 1, 2012** ↑

18. From the AI RMF: "The information gathered while carrying out the MAP function enables negative risk prevention and informs decisions for processes such as model management, as well as an initial decision about appropriateness or the need for an AI solution. [...] Implementation of this function is enhanced by incorporating perspectives from a diverse internal team and engagement with those external to the team that developed or deployed the AI system. Engagement with external collaborators, end users, potentially impacted communities, and others may vary based on the risk level of a particular AI system, the makeup of the internal team, and organizational policies. Gathering such broad perspectives can help organizations proactively prevent negative risks and develop more trustworthy AI systems by:
   • improving their capacity for understanding contexts;
   • checking their assumptions about context of use;
   • enabling recognition of when systems are not functional within or out of their intended context;
   • identifying positive and beneficial uses of their existing AI systems;
   • improving understanding of limitations in AI and ML processes;
   • identifying constraints in real-world applications that may lead to negative impacts;
   • identifying known and foreseeable negative impacts related to intended use of AI systems; and
   • anticipating risks of the use of AI systems beyond intended use." See **NIST, NIST AI Risk Management Framework, January 26, 2023** ↑

19. **OHCHR B-Tech Project, Advancing Responsible Development and Deployment of Generative AI, 2024** ↑

20. **OMB, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, Section 5(c)(iv)(A)(1), March 2024** ↑

21. **OECD, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, Annex Q2, June 8, 2023** ↑

22. From the AI RMF: "The **MEASURE** function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. It uses knowledge relevant to AI risks identified in the **MAP** function and informs the **MANAGE** function. AI systems should be tested before their deployment and regularly while in operation. AI risk measurements include documenting aspects of systems' functionality and trustworthiness." See **NIST, NIST AI Risk Management Framework, January 26, 2023** ↑

23. Additionally, "Member States and business enterprises should implement appropriate measures to monitor all phases of an AI system lifecycle, including the functioning of algorithms used for decision-making, the data, as well as AI actors involved in the process. Member States should establish ethical impact assessments, which can identify and assess benefits, concerns and risks of AI systems, as well as appropriate risk prevention, mitigation, and monitoring measures, among other assurance mechanisms." See **UNESCO, Recommendation on the Ethics of Artificial Intelligence, May 16, 2023** ↑

24. From the AI RMF: "The **MANAGE** function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the **GOVERN** function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events. ↑Contextual information gleaned from expert consultation and input from relevant AI actors – established in **GOVERN** and carried out in **MAP** – is utilized in this function to decrease the likelihood of system failures and negative impacts. Systematic documentation practices established in **GOVERN** and utilized in **MAP** and **MEASURE** bolster AI risk management efforts and increase transparency and accountability. Processes for assessing emergent risks are in place, along with mechanisms for continual improvement. After completing the **MANAGE** function, plans for prioritizing risk and regular monitoring and improvement will be in place. Framework users will have enhanced capacity to manage the risks of deployed AI systems and to allocate risk management resources based on assessed and prioritized risks. It is incumbent on Framework users to continue to apply the **MANAGE** function to deployed AI systems as methods, contexts, risks, and needs or expectations from relevant AI actors evolve over time." See **NIST, NIST AI Risk Management Framework, January 26, 2023** ↑

25. Furthermore, the OECD MNE Guidelines states, "Once the most significant

impacts are identified and dealt with, the enterprise should move on to address less significant impacts. The process of prioritization is also ongoing, and in some instances new or emerging adverse impacts may arise and be prioritized before moving on to less significant impacts." See **OECD, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, Annex Q3, June 8, 2023** ↑

26. Additionally related to **Manage 4.1** and as reflected in the UNESCO Recommendation on the Ethics of AI: "Member States should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and fundamental freedoms and the rule of law are respected in the digital world and in the physical world. Such mechanisms and actions should include remediation mechanisms provided by private and public sector companies. The auditability and traceability of AI systems should be promoted to this end. In addition, Member States should strengthen their institutional capacities to deliver on this commitment and should collaborate with researchers and other stakeholders to investigate, prevent and mitigate any potentially malicious uses of AI systems." See **UNESCO, Recommendation on the Ethics of Artificial Intelligence, May 16, 2023** ↑

TAGS   **Bureau of Cyberspace and Digital Policy**

White House

USA.gov Office of the Inspector General

Archives Contact Us

FOLLOW US   f   X   ⊙   ▶   ⋯   ✉   Privacy Policy

Accessibility Statement

Copyright Information

FOIA

No FEAR Act