NIST Trustworthy and Responsible AI
NIST AI 600-1

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.600-1

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Trustworthy and Responsible AI
NIST AI 600-1

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.600-1

July 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

About AI at NIST: The National Institute of Standards and Technology (NIST) develops measurements, technology, tools, and standards to advance reliable, safe, transparent, explainable, privacy-enhanced, and fair artificial intelligence (AI) so that its full commercial and societal benefits can be realized without harm to people or the planet. NIST, which has conducted both fundamental and applied work on AI for more than a decade, is also helping to fulfill the 2023 Executive Order on Safe, Secure, and Trustworthy AI. NIST established the U.S. AI Safety Institute and the companion AI Safety Institute Consortium to continue the efforts set in motion by the E.O. to build the science necessary for safe, secure, and trustworthy development and use of AI.

Acknowledgments: *This report was accomplished with the many helpful comments and contributions from the community, including the NIST Generative AI Public Working Group, and NIST staff and guest researchers: Chloe Autio, Jesse Dunietz, Patrick Hall, Shomik Jain, Kamie Roberts, Reva Schwartz, Martin Stanley, and Elham Tabassi.*

Publication History
Approved by the NIST Editorial Review Board on 07-25-2024

Contact Information
ai-inquiries@nist.gov

National Institute of Standards and Technology
Attn: NIST AI Innovation Lab, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8900

Additional Information
Additional information about this publication and other NIST AI publications are available at https://airc.nist.gov/Home.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to adequately describe an experimental procedure or concept. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. Any mention of commercial, non-profit, academic partners, or their products, or references is for information only; it is not intended to imply endorsement or recommendation by any U.S. Government agency.

Table of Contents

1. Introduction

This document is a cross-sectoral profile of and companion resource for the AI Risk Management Framework (AI RMF 1.0) for Generative AI,[1] pursuant to President Biden's Executive Order (EO) 14110 on Safe, Secure, and Trustworthy Artificial Intelligence.[2] The AI RMF was released in January 2023, and is intended for voluntary use and to improve the ability of organizations to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

A *profile* is an implementation of the AI RMF functions, categories, and subcategories for a specific setting, application, or technology – in this case, Generative AI (GAI) – based on the requirements, risk tolerance, and resources of the Framework user. AI RMF profiles assist organizations in deciding how to best manage AI risks in a manner that is well-aligned with their goals, considers legal/regulatory requirements and best practices, and reflects risk management priorities. Consistent with other AI RMF profiles, this profile offers insights into how risk can be managed across various stages of the AI lifecycle and for GAI as a technology.

As GAI covers risks of models or applications that can be used across use cases or sectors, this document is an AI RMF cross-sectoral profile. Cross-sectoral profiles can be used to govern, map, measure, and manage risks associated with activities or business processes common across sectors, such as the use of large language models (LLMs), cloud-based services, or acquisition.

This document defines risks that are novel to or exacerbated by the use of GAI. After introducing and describing these risks, the document provides a set of suggested actions to help organizations govern, map, measure, and manage these risks.

[1] EO 14110 defines Generative AI as "the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content." While not all GAI is derived from foundation models, for purposes of this document, GAI generally refers to generative foundation models. The foundation model subcategory of "dual-use foundation models" is defined by EO 14110 as "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts."

[2] This profile was developed per Section 4.1(a)(i)(A) of EO 14110, which directs the Secretary of Commerce, acting through the Director of the National Institute of Standards and Technology (NIST), to develop a companion resource to the AI RMF, NIST AI 100–1, for generative AI.

This work was informed by public feedback and consultations with diverse stakeholder groups as part of NIST's Generative AI Public Working Group (GAI PWG). The GAI PWG was an open, transparent, and collaborative process, facilitated via a virtual workspace, to obtain multistakeholder input on GAI risk management and to inform NIST's approach.

The focus of the GAI PWG was limited to four primary considerations relevant to GAI: Governance, Content Provenance, Pre-deployment Testing, and Incident Disclosure (further described in Appendix A). As such, the suggested actions in this document primarily address these considerations.

Future revisions of this profile will include additional AI RMF subcategories, risks, and suggested actions based on additional considerations of GAI as the space evolves and empirical evidence indicates additional risks. A glossary of terms pertinent to GAI risk management will be developed and hosted on NIST's Trustworthy & Responsible AI Resource Center (AIRC), and added to *The Language of Trustworthy AI: An In-Depth Glossary of Terms.*

This document was also informed by public comments and consultations from several Requests for Information.

2. Overview of Risks Unique to or Exacerbated by GAI

In the context of the AI RMF, *risk* refers to the composite measure of an event's **probability** (or likelihood) of occurring and the **magnitude** or degree of the consequences of the corresponding event. Some risks can be assessed as likely to materialize in a given context, particularly those that have been empirically demonstrated in similar contexts. Other risks may be unlikely to materialize in a given context, or may be more speculative and therefore uncertain.

AI risks can differ from or intensify traditional software risks. Likewise, GAI can exacerbate existing AI risks, and creates unique risks. GAI risks can vary along many dimensions:

- **Stage of the AI lifecycle:** Risks can arise during design, development, deployment, operation, and/or decommissioning.
- **Scope:** Risks may exist at individual model or system levels, at the application or implementation levels (i.e., for a specific use case), or at the ecosystem level – that is, beyond a single system or organizational context. Examples of the latter include the expansion of "algorithmic monocultures,[3]" resulting from repeated use of the same model, or impacts on access to opportunity, labor markets, and the creative economies.[4]
- **Source of risk:** Risks may emerge from factors related to the design, training, or operation of the GAI model itself, stemming in some cases from GAI model or system inputs, and in other cases, from GAI system outputs. Many GAI risks, however, originate from human behavior, including

[3] "Algorithmic monocultures" refers to the phenomenon in which repeated use of the same model or algorithm in consequential decision-making settings like employment and lending can result in increased susceptibility by systems to correlated failures (like unexpected shocks), due to multiple actors relying on the same algorithm.
[4] Many studies have projected the impact of AI on the workforce and labor markets. Fewer studies have examined the impact of GAI on the labor market, though some industry surveys indicate that that both employees and employers are pondering this disruption.

the abuse, misuse, and unsafe repurposing by humans (adversarial or not), and others result from interactions between a human and an AI system.

- **Time scale:** GAI risks may materialize abruptly or across extended periods. Examples include immediate (and/or prolonged) emotional harm and potential risks to physical safety due to the distribution of harmful deepfake images, or the long-term effect of disinformation on societal trust in public institutions.

The presence of risks and where they fall along the dimensions above will vary depending on the characteristics of the GAI model, system, or use case at hand. These characteristics include but are not limited to GAI model or system architecture, training mechanisms and libraries, data types used for training or fine-tuning, levels of model access or availability of model weights, and application or use case context.

Organizations may choose to tailor how they measure GAI risks based on these characteristics. They may additionally wish to allocate risk management resources relative to the severity and likelihood of negative impacts, including where and how these risks manifest, and their direct and material impacts harms in the context of GAI use. Mitigations for model or system level risks may differ from mitigations for use-case or ecosystem level risks.

Importantly, some GAI risks are unknown, and are therefore difficult to properly scope or evaluate given the uncertainty about potential GAI scale, complexity, and capabilities. Other risks may be known but difficult to estimate given the wide range of GAI stakeholders, uses, inputs, and outputs. Challenges with risk estimation are aggravated by a lack of visibility into GAI training data, and the generally immature state of the science of AI measurement and safety today. This document focuses on risks for which there is an existing empirical evidence base at the time this profile was written; for example, speculative risks that may potentially arise in more advanced, future GAI systems are not considered. Future updates may incorporate additional risks or provide further details on the risks identified below.

To guide organizations in identifying and managing GAI risks, a set of risks unique to or exacerbated by the development and use of GAI are defined below.[5] Each risk is labeled according to the outcome, object, or source of the risk (i.e., some are risks "to" a subject or domain and others are risks "of" or "from" an issue or theme). These risks provide a lens through which organizations can frame and execute risk management efforts. To help streamline risk management efforts, each risk is mapped in Section 3 (as well as in tables in Appendix B) to relevant Trustworthy AI Characteristics identified in the AI RMF.

[5] These risks can be further categorized by organizations depending on their unique approaches to risk definition and management. One possible way to further categorize these risks, derived in part from the UK's International Scientific Report on the Safety of Advanced AI, could be: **1) Technical / Model risks (or risk from malfunction):** Confabulation; Dangerous or Violent Recommendations; Data Privacy; Value Chain and Component Integration; Harmful Bias, and Homogenization; **2) Misuse by humans (or malicious use):** CBRN Information or Capabilities; Data Privacy; Human-AI Configuration; Obscene, Degrading, and/or Abusive Content; Information Integrity; Information Security; **3) Ecosystem / societal risks (or systemic risks)**: Data Privacy; Environmental; Intellectual Property. We also note that some risks are cross-cutting between these categories.

1. **CBRN Information or Capabilities:** Eased access to or synthesis of materially nefarious information or design capabilities related to chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents.
2. **Confabulation:** The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived.[6]
3. **Dangerous, Violent, or Hateful Content:** Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content.
4. **Data Privacy:** Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data.[7]
5. **Environmental Impacts:** Impacts due to high compute resource utilization in training or operating GAI models, and related outcomes that may adversely impact ecosystems.
6. **Harmful Bias or Homogenization:** Amplification and exacerbation of historical, societal, and systemic biases; performance disparities[8] between sub-groups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; undesired homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.
7. **Human-AI Configuration:** Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing GAI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with GAI systems.
8. **Information Integrity:** Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns.
9. **Information Security:** Lowered barriers for offensive cyber capabilities, including via automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, offensive cyber

[6] Some commenters have noted that the terms "hallucination" and "fabrication" anthropomorphize GAI, which itself is a risk related to GAI systems as it can inappropriately attribute human characteristics to non-human entities.

[7] What is categorized as sensitive data or sensitive PII can be highly contextual based on the nature of the information, but examples of sensitive information include information that relates to an information subject's most intimate sphere, including political opinions, sex life, or criminal convictions.

[8] The notion of harm presumes some baseline scenario that the harmful factor (e.g., a GAI model) makes worse. When the mechanism for potential harm is a disparity between groups, it can be difficult to establish what the most appropriate baseline is to compare against, which can result in divergent views on when a disparity between AI behaviors for different subgroups constitutes a harm. In discussing harms from disparities such as biased behavior, this document highlights examples where someone's situation is worsened relative to what it would have been in the absence of any AI system, making the outcome unambiguously a harm of the system.

operations, or other cyberattacks; increased attack surface for targeted cyberattacks, which may compromise a system's availability or the confidentiality or integrity of training data, code, or model weights.

10. **Intellectual Property:** Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication.
11. **Obscene, Degrading, and/or Abusive Content:** Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.
12. **Value Chain and Component Integration:** Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from GAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users.

2.1. CBRN Information or Capabilities

In the future, GAI may enable malicious actors to more easily access CBRN weapons and/or relevant knowledge, information, materials, tools, or technologies that could be misused to assist in the design, development, production, or use of CBRN weapons or other dangerous materials or agents. While relevant biological and chemical threat knowledge and information is often publicly accessible, LLMs could facilitate its analysis or synthesis, particularly by individuals without formal scientific training or expertise.

Recent research on this topic found that LLM outputs regarding biological threat creation and attack planning provided minimal assistance beyond traditional search engine queries, suggesting that state-of-the-art LLMs at the time these studies were conducted do not substantially increase the operational likelihood of such an attack. The physical synthesis development, production, and use of chemical or biological agents will continue to require both applicable expertise and supporting materials and infrastructure. The impact of GAI on chemical or biological agent misuse will depend on what the key barriers for malicious actors are (e.g., whether information access is one such barrier), and how well GAI can help actors address those barriers.

Furthermore, chemical and biological design tools (BDTs) – highly specialized AI systems trained on scientific data that aid in chemical and biological design – may augment design capabilities in chemistry and biology beyond what text-based LLMs are able to provide. As these models become more efficacious, including for beneficial uses, it will be important to assess their potential to be used for harm, such as the ideation and design of novel harmful chemical or biological agents.

While some of these described capabilities lie beyond the reach of existing GAI tools, ongoing assessments of this risk would be enhanced by monitoring both the ability of AI tools to facilitate CBRN weapons planning and GAI systems' connection or access to relevant data and tools.

Trustworthy AI Characteristic: Safe, Explainable and Interpretable

2.2. Confabulation

"Confabulation" refers to a phenomenon in which GAI systems generate and confidently present erroneous or false content in response to prompts. Confabulations also include generated outputs that diverge from the prompts or other input or that contradict previously generated statements in the same context. These phenomena are colloquially also referred to as "hallucinations" or "fabrications."

Confabulations can occur across GAI outputs and contexts.[9,10] Confabulations are a natural result of the way generative models are designed: they generate outputs that approximate the statistical distribution of their training data; for example, LLMs predict the next token or word in a sentence or phrase. While such statistical prediction can produce factually accurate and consistent outputs, it can also produce outputs that are factually inaccurate or internally inconsistent. This dynamic is particularly relevant when it comes to open-ended prompts for long-form responses and in domains which require highly contextual and/or domain expertise.

Risks from confabulations may arise when users believe false content – often due to the confident nature of the response – leading users to act upon or promote the false information. This poses a challenge for many real-world applications, such as in healthcare, where a confabulated summary of patient information reports could cause doctors to make incorrect diagnoses and/or recommend the wrong treatments. Risks of confabulated content may be especially important to monitor when integrating GAI into applications involving consequential decision making.

GAI outputs may also include confabulated logic or citations that purport to justify or explain the system's answer, which may further mislead humans into inappropriately trusting the system's output. For instance, LLMs sometimes provide logical steps for how they arrived at an answer even when the answer itself is incorrect. Similarly, an LLM could falsely assert that it is human or has human traits, potentially deceiving humans into believing they are speaking with another human.

The extent to which humans can be deceived by LLMs, the mechanisms by which this may occur, and the potential risks from adversarial prompting of such behavior are emerging areas of study. Given the wide range of downstream impacts of GAI, it is difficult to estimate the downstream scale and impact of confabulations.

Trustworthy AI Characteristics: Fair with Harmful Bias Managed, Safe, Valid and Reliable, Explainable and Interpretable

2.3. Dangerous, Violent, or Hateful Content

GAI systems can produce content that is inciting, radicalizing, or threatening, or that glorifies violence, with greater ease and scale than other technologies. LLMs have been reported to generate dangerous or violent recommendations, and some models have generated actionable instructions for dangerous or

---

[9] Confabulations of falsehoods are most commonly a problem for text-based outputs; for audio, image, or video content, creative generation of non-factual content can be a desired behavior.

[10] For example, legal confabulations have been shown to be pervasive in current state-of-the-art LLMs. See also, e.g.,

unethical behavior. Text-to-image models also make it easy to create images that could be used to promote dangerous or violent messages. Similar concerns are present for other GAI media, including video and audio. GAI may also produce content that recommends self-harm or criminal/illegal activities.

Many current systems restrict model outputs to limit certain content or in response to certain prompts, but this approach may still produce harmful recommendations in response to other less-explicit, novel prompts (also relevant to CBRN Information or Capabilities, Data Privacy, Information Security, and Obscene, Degrading and/or Abusive Content). Crafting such prompts deliberately is known as "jailbreaking," or, manipulating prompts to circumvent output controls. Limitations of GAI systems can be harmful or dangerous in certain contexts. Studies have observed that users may disclose mental health issues in conversations with chatbots – and that users exhibit negative reactions to unhelpful responses from these chatbots during situations of distress.

This risk encompasses difficulty controlling creation of and public exposure to offensive or hateful language, and denigrating or stereotypical content generated by AI. This kind of speech may contribute to downstream harm such as fueling dangerous or violent behaviors. The spread of denigrating or stereotypical content can also further exacerbate representational harms (see Harmful Bias and Homogenization below).

Trustworthy AI Characteristics: Safe, Secure and Resilient

2.4. Data Privacy

GAI systems raise several risks to privacy. GAI system training requires large volumes of data, which in some cases may include personal data. The use of personal data for GAI training raises risks to widely accepted privacy principles, including to transparency, individual participation (including consent), and purpose specification. For example, most model developers do not disclose specific data sources on which models were trained, limiting user awareness of whether personally identifiably information (PII) was trained on and, if so, how it was collected.

Models may leak, generate, or correctly infer sensitive information about individuals. For example, during adversarial attacks, LLMs have revealed sensitive information (from the public domain) that was included in their training data. This problem has been referred to as data memorization, and may pose exacerbated privacy risks even for data present only in a small number of training samples.

In addition to revealing sensitive information in GAI training data, GAI models may be able to correctly infer PII or sensitive data that was not in their training data nor disclosed by the user by stitching together information from disparate sources. These inferences can have negative impact on an individual even if the inferences are not accurate (e.g., confabulations), and especially if they reveal information that the individual considers sensitive or that is used to disadvantage or harm them.

Beyond harms from information exposure (such as extortion or dignitary harm), wrong or inappropriate inferences of PII can contribute to downstream or secondary harmful impacts. For example, predictive inferences made by GAI models based on PII or protected attributes can contribute to adverse decisions, leading to representational or allocative harms to individuals or groups (see Harmful Bias and Homogenization below).

Trustworthy AI Characteristics: Accountable and Transparent, Privacy Enhanced, Safe, Secure and Resilient

2.5. Environmental Impacts

Training, maintaining, and operating (running inference on) GAI systems are resource-intensive activities, with potentially large energy and environmental footprints. Energy and carbon emissions vary based on what is being done with the GAI model (i.e., pre-training, fine-tuning, inference), the modality of the content, hardware used, and type of task or application.

Current estimates suggest that training a single transformer LLM can emit as much carbon as 300 round-trip flights between San Francisco and New York. In a study comparing energy consumption and carbon emissions for LLM inference, generative tasks (e.g., text summarization) were found to be more energy- and carbon-intensive than discriminative or non-generative tasks (e.g., text classification).

Methods for creating smaller versions of trained models, such as model distillation or compression, could reduce environmental impacts at inference time, but training and tuning such models may still contribute to their environmental impacts. Currently there is no agreed upon method to estimate environmental impacts from GAI.

Trustworthy AI Characteristics: Accountable and Transparent, Safe

2.6. Harmful Bias and Homogenization

Bias exists in many forms and can become ingrained in automated systems. AI systems, including GAI systems, can increase the speed and scale at which harmful biases manifest and are acted upon, potentially perpetuating and amplifying harms to individuals, groups, communities, organizations, and society. For example, when prompted to generate images of CEOs, doctors, lawyers, and judges, current text-to-image models underrepresent women and/or racial minorities, and people with disabilities. Image generator models have also produced biased or stereotyped output for various demographic groups and have difficulty producing non-stereotyped content even when the prompt specifically requests image features that are inconsistent with the stereotypes. Harmful bias in GAI models, which may stem from their training data, can also cause representational harms or perpetuate or exacerbate bias based on race, gender, disability, or other protected classes.

Harmful bias in GAI systems can also lead to harms via disparities between how a model performs for different subgroups or languages (e.g., an LLM may perform less well for non-English languages or certain dialects). Such disparities can contribute to discriminatory decision-making or amplification of existing societal biases. In addition, GAI systems may be inappropriately trusted to perform similarly across all subgroups, which could leave the groups facing underperformance with worse outcomes than if no GAI system were used. Disparate or reduced performance for lower-resource languages also presents challenges to model adoption, inclusion, and accessibility, and may make preservation of endangered languages more difficult if GAI systems become embedded in everyday processes that would otherwise have been opportunities to use these languages.

Bias is mutually reinforcing with the problem of undesired homogenization, in which GAI systems produce skewed distributions of outputs that are overly uniform (for example, repetitive aesthetic styles

and reduced content diversity). Overly homogenized outputs can themselves be incorrect, or they may lead to unreliable decision-making or amplify harmful biases. These phenomena can flow from foundation models to downstream models and systems, with the foundation models acting as "bottlenecks," or single points of failure.

Overly homogenized content can contribute to "model collapse." Model collapse can occur when model training over-relies on synthetic data, resulting in data points disappearing from the distribution of the new model's outputs. In addition to threatening the robustness of the model overall, model collapse could lead to homogenized outputs, including by amplifying any homogenization from the model used to generate the synthetic training data.

Trustworthy AI Characteristics: Fair with Harmful Bias Managed, Valid and Reliable

2.7. Human-AI Configuration

GAI system use can involve varying risks of misconfigurations and poor interactions between a system and a human who is interacting with it. Humans bring their unique perspectives, experiences, or domain-specific expertise to interactions with AI systems but may not have detailed knowledge of AI systems and how they work. As a result, human experts may be unnecessarily "averse" to GAI systems, and thus deprive themselves or others of GAI's beneficial uses.

Conversely, due to the complexity and increasing reliability of GAI technology, over time, humans may over-rely on GAI systems or may unjustifiably perceive GAI content to be of higher quality than that produced by other sources. This phenomenon is an example of automation bias, or excessive deference to automated systems. Automation bias can exacerbate other risks of GAI, such as risks of confabulation or risks of bias or homogenization.

There may also be concerns about emotional entanglement between humans and GAI systems, which could lead to negative psychological impacts.

Trustworthy AI Characteristics: Accountable and Transparent, Explainable and Interpretable, Fair with Harmful Bias Managed, Privacy Enhanced, Safe, Valid and Reliable

2.8. Information Integrity

Information integrity describes the "spectrum of information and associated patterns of its creation, exchange, and consumption in society." High-integrity information can be trusted; "distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting. This information can be linked to the original source(s) with appropriate evidence. High-integrity information is also accurate and reliable, can be verified and authenticated, has a clear chain of custody, and creates reasonable expectations about when its validity may expire."[11]

[11] This definition of information integrity is derived from the 2022 White House Roadmap for Researchers on Priorities Related to Information Integrity Research and Development.

GAI systems can ease the unintentional production or dissemination of false, inaccurate, or misleading content (misinformation) at scale, particularly if the content stems from confabulations.

GAI systems can also ease the deliberate production or dissemination of false or misleading information (disinformation) at scale, where an actor has the explicit intent to deceive or cause harm to others. Even very subtle changes to text or images can manipulate human and machine perception.

Similarly, GAI systems could enable a higher degree of sophistication for malicious actors to produce disinformation that is targeted towards specific demographics. Current and emerging multimodal models make it possible to generate both text-based disinformation and highly realistic "deepfakes" – that is, synthetic audiovisual content and photorealistic images.[12] Additional disinformation threats could be enabled by future GAI models trained on new data modalities.

Disinformation and misinformation – both of which may be facilitated by GAI – may erode public trust in true or valid evidence and information, with downstream effects. For example, a synthetic image of a Pentagon blast went viral and briefly caused a drop in the stock market. Generative AI models can also assist malicious actors in creating compelling imagery and propaganda to support disinformation campaigns, which may not be photorealistic, but could enable these campaigns to gain more reach and engagement on social media platforms. Additionally, generative AI models can assist malicious actors in creating fraudulent content intended to impersonate others.

Trustworthy AI Characteristics: Accountable and Transparent, Safe, Valid and Reliable, Interpretable and Explainable

2.9. Information Security

Information security for computer systems and data is a mature field with widely accepted and standardized practices for offensive and defensive cyber capabilities. GAI-based systems present two primary information security risks: GAI could potentially discover or enable new cybersecurity risks by lowering the barriers for or easing automated exercise of offensive capabilities; simultaneously, it expands the available attack surface, as GAI itself is vulnerable to attacks like prompt injection or data poisoning.

Offensive cyber capabilities advanced by GAI systems may augment cybersecurity attacks such as hacking, malware, and phishing. Reports have indicated that LLMs are already able to discover some vulnerabilities in systems (hardware, software, data) and write code to exploit them. Sophisticated threat actors might further these risks by developing GAI-powered security co-pilots for use in several parts of the attack chain, including informing attackers on how to proactively evade threat detection and escalate privileges after gaining system access.

Information security for GAI models and systems also includes maintaining availability of the GAI system and the integrity and (when applicable) the confidentiality of the GAI code, training data, and model weights. To identify and secure potential attack points in AI systems or specific components of the AI

[12] See also https://doi.org/10.6028/NIST.AI.100-4, to be published.

value chain (e.g., data inputs, processing, GAI training, or deployment environments), conventional cybersecurity practices may need to adapt or evolve.

For instance, prompt injection involves modifying what input is provided to a GAI system so that it behaves in unintended ways. In direct prompt injections, attackers might craft malicious prompts and input them directly to a GAI system, with a variety of downstream negative consequences to interconnected systems. Indirect prompt injection attacks occur when adversaries remotely (i.e., without a direct interface) exploit LLM-integrated applications by injecting prompts into data likely to be retrieved. Security researchers have already demonstrated how indirect prompt injections can exploit vulnerabilities by stealing proprietary data or running malicious code remotely on a machine. Merely querying a closed production model can elicit previously undisclosed information about that model.

Another cybersecurity risk to GAI is data poisoning, in which an adversary compromises a training dataset used by a model to manipulate its outputs or operation. Malicious tampering with data or parts of the model could exacerbate risks associated with GAI system outputs.

Trustworthy AI Characteristics: Privacy Enhanced, Safe, Secure and Resilient, Valid and Reliable

2.10. Intellectual Property

Intellectual property risks from GAI systems may arise where the use of copyrighted works is not a fair use under the fair use doctrine. If a GAI system's training data included copyrighted material, GAI outputs displaying instances of training data memorization (see Data Privacy above) could infringe on copyright.

How GAI relates to copyright, including the status of generated content that is similar to but does not strictly copy work protected by copyright, is currently being debated in legal fora. Similar discussions are taking place regarding the use or emulation of personal identity, likeness, or voice without permission.

Trustworthy AI Characteristics: Accountable and Transparent, Fair with Harmful Bias Managed, Privacy Enhanced

2.11. Obscene, Degrading, and/or Abusive Content

GAI can ease the production of and access to illegal non-consensual intimate imagery (NCII) of adults, and/or child sexual abuse material (CSAM). GAI-generated obscene, abusive or degrading content can create privacy, psychological and emotional, and even physical harms, and in some cases may be illegal.

Generated explicit or obscene AI content may include highly realistic "deepfakes" of real individuals, including children. The spread of this kind of material can have downstream negative consequences: in the context of CSAM, even if the generated images do not resemble specific individuals, the prevalence of such images can divert time and resources from efforts to find real-world victims. Outside of CSAM, the creation and spread of NCII disproportionately impacts women and sexual minorities, and can have subsequent negative consequences including decline in overall mental health, substance abuse, and even suicidal thoughts.

Data used for training GAI models may unintentionally include CSAM and NCII. A recent report noted that several commonly used GAI training datasets were found to contain hundreds of known images of

CSAM. Even when trained on "clean" data, increasingly capable GAI models can synthesize or produce synthetic NCII and CSAM. Websites, mobile apps, and custom-built models that generate synthetic NCII have moved from niche internet forums to mainstream, automated, and scaled online businesses.

Trustworthy AI Characteristics: Fair with Harmful Bias Managed, Safe, Privacy Enhanced

2.12. Value Chain and Component Integration

GAI value chains involve many third-party components such as procured datasets, pre-trained models, and software libraries. These components might be improperly obtained or not properly vetted, leading to diminished transparency or accountability for downstream users. While this is a risk for traditional AI systems and some other digital technologies, the risk is exacerbated for GAI due to the scale of the training data, which may be too large for humans to vet; the difficulty of training foundation models, which leads to extensive reuse of limited numbers of models; and the extent to which GAI may be integrated into other devices and services. As GAI systems often involve many distinct third-party components and data sources, it may be difficult to attribute issues in a system's behavior to any one of these sources.

Errors in third-party GAI components can also have downstream impacts on accuracy and robustness. For example, test datasets commonly used to benchmark or validate models can contain label errors. Inaccuracies in these labels can impact the "stability" or robustness of these benchmarks, which many GAI practitioners consider during the model selection process.

Trustworthy AI Characteristics: Accountable and Transparent, Explainable and Interpretable, Fair with Harmful Bias Managed, Privacy Enhanced, Safe, Secure and Resilient, Valid and Reliable

3. Suggested Actions to Manage GAI Risks

The following suggested actions target risks unique to or exacerbated by GAI.

In addition to the suggested actions below, AI risk management activities and actions set forth in the AI RMF 1.0 and Playbook are already applicable for managing GAI risks. Organizations are encouraged to apply the activities suggested in the AI RMF and its Playbook when managing the risk of GAI systems.

Implementation of the suggested actions will vary depending on the type of risk, characteristics of GAI systems, stage of the GAI lifecycle, and relevant AI actors involved.

Suggested actions to manage GAI risks can be found in the tables below:

- The suggested actions are **organized by relevant AI RMF subcategories** to streamline these activities alongside implementation of the AI RMF.
- **Not every subcategory of the AI RMF is included in this document.**[13] Suggested actions are listed for only some subcategories.

[13] As this document was focused on the GAI PWG efforts and primary considerations (see Appendix A), AI RMF subcategories not addressed here may be added later.

- Not every suggested action applies to **every** AI Actor[14] or is relevant to every AI Actor Task. For example, suggested actions relevant to GAI developers may not be relevant to GAI deployers. The applicability of suggested actions to relevant AI actors should be determined based on organizational considerations and their unique uses of GAI systems.

Each table of suggested actions includes:

- **Action ID:** Each Action ID corresponds to the relevant AI RMF function and subcategory (e.g., GV-1.1-001 corresponds to the first suggested action for Govern 1.1, GV-1.1-002 corresponds to the second suggested action for Govern 1.1). AI RMF functions are tagged as follows: GV = Govern; MP = Map; MS = Measure; MG = Manage.
- **Suggested Action:** Steps an organization or AI actor can take to manage GAI risks.
- **GAI Risks:** Tags linking suggested actions with relevant GAI risks.
- **AI Actor Tasks:** Pertinent AI Actor Tasks for each subcategory. Not every AI Actor Task listed will apply to every suggested action in the subcategory (i.e., some apply to AI development and others apply to AI deployment).

The tables below begin with the AI RMF subcategory, shaded in blue, followed by suggested actions.

GOVERN 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.1-001	Align GAI development and use with applicable laws and regulations, including those related to data privacy, copyright and intellectual property law.	Data Privacy; Harmful Bias and Homogenization; Intellectual Property
AI Actor Tasks: Governance and Oversight		

[14] AI Actors are defined by the OECD as "those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI." See Appendix A of the AI RMF for additional descriptions of AI Actors and AI Actor Tasks.

GOVERN 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.2-001	Establish transparency policies and processes for documenting the origin and history of training data and generated data for GAI applications to advance digital content transparency, while balancing the proprietary nature of training approaches.	Data Privacy; Information Integrity; Intellectual Property
GV-1.2-002	Establish policies to evaluate risk-relevant capabilities of GAI and robustness of safety measures, both prior to deployment and on an ongoing basis, through internal and external evaluations.	CBRN Information or Capabilities; Information Security
AI Actor Tasks: Governance and Oversight		

GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.3-001	Consider the following factors when updating or defining risk tiers for GAI: Abuses and impacts to information integrity; Dependencies between GAI and other IT or data systems; Harm to fundamental rights or public safety; Presentation of obscene, objectionable, offensive, discriminatory, invalid or untruthful output; Psychological impacts to humans (e.g., anthropomorphization, algorithmic aversion, emotional entanglement); Possibility for malicious use; Whether the system introduces significant new security vulnerabilities; Anticipated system impact on some groups compared to others; Unreliable decision making capabilities, validity, adaptability, and variability of GAI system performance over time.	Information Integrity; Obscene, Degrading, and/or Abusive Content; Value Chain and Component Integration; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities
GV-1.3-002	Establish minimum thresholds for performance or assurance criteria and review as part of deployment approval ("go/"no-go") policies, procedures, and processes, with reviewed processes and approval thresholds reflecting measurement of GAI capabilities and risks.	CBRN Information or Capabilities; Confabulation; Dangerous, Violent, or Hateful Content
GV-1.3-003	Establish a test plan and response policy, before developing highly capable models, to periodically evaluate whether the model may misuse CBRN information or capabilities and/or offensive cyber capabilities.	CBRN Information or Capabilities; Information Security

GV-1.3-004	Obtain input from stakeholder communities to identify unacceptable use, in accordance with activities in the AI RMF Map function.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
GV-1.3-005	Maintain an updated hierarchy of identified and expected GAI risks connected to contexts of GAI model advancement and use, potentially including specialized risk levels for GAI systems that address issues such as model collapse and algorithmic monoculture.	Harmful Bias and Homogenization
GV-1.3-006	Reevaluate organizational risk tolerances to account for unacceptable negative risk (such as where significant negative impacts are imminent, severe harms are actually occurring, or large-scale risks could occur); and broad GAI negative risks, including: Immature safety or risk cultures related to AI and GAI design, development and deployment, public information integrity risks, including impacts on democratic processes, unknown long-term performance characteristics of GAI.	Information Integrity; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities
GV-1.3-007	Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk.	CBRN Information and Capability; Information Security; Information Integrity
AI Actor Tasks: Governance and Oversight		

GOVERN 1.4: The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.4-001	Establish policies and mechanisms to prevent GAI systems from generating CSAM, NCII or content that violates the law.	Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
GV-1.4-002	Establish transparent acceptable use policies for GAI that address illegal use or applications of GAI.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Data Privacy; Civil Rights violations
AI Actor Tasks: AI Development, AI Deployment, Governance and Oversight		

GOVERN 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, and organizational roles and responsibilities are clearly defined, including determining the frequency of periodic review.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.5-001	Define organizational responsibilities for periodic review of content provenance and incident monitoring for GAI systems.	Information Integrity
GV-1.5-002	Establish organizational policies and procedures for after action reviews of GAI system incident response and incident disclosures, to identify gaps; Update incident response and incident disclosure processes as required.	Human-AI Configuration; Information Security
GV-1.5-003	Maintain a document retention policy to keep history for test, evaluation, validation, and verification (TEVV), and digital content transparency methods for GAI.	Information Integrity; Intellectual Property
AI Actor Tasks: Governance and Oversight, Operation and Monitoring		

GOVERN 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.6-001	Enumerate organizational GAI systems for incorporation into AI system inventory and adjust AI system inventory requirements to account for GAI risks.	Information Security
GV-1.6-002	Define any inventory exemptions in organizational policies for GAI systems embedded into application software.	Value Chain and Component Integration
GV-1.6-003	In addition to general model, governance, and risk information, consider the following items in GAI system inventory entries: Data provenance information (e.g., source, signatures, versioning, watermarks); Known issues reported from internal bug tracking or external information sharing resources (e.g., AI incident database, AVID, CVE, NVD, or OECD AI incident monitor); Human oversight roles and responsibilities; Special rights and considerations for intellectual property, licensed works, or personal, privileged, proprietary or sensitive data; Underlying foundation models, versions of underlying models, and access modes.	Data Privacy; Human-AI Configuration; Information Integrity; Intellectual Property; Value Chain and Component Integration
AI Actor Tasks: Governance and Oversight		

GOVERN 1.7: Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.		
Action ID	**Suggested Action**	**GAI Risks**
GV-1.7-001	Protocols are put in place to ensure GAI systems are able to be deactivated when necessary.	Information Security; Value Chain and Component Integration
GV-1.7-002	Consider the following factors when decommissioning GAI systems: Data retention requirements; Data security, e.g., containment, protocols, Data leakage after decommissioning; Dependencies between upstream, downstream, or other data, internet of things (IOT) or AI systems; Use of open-source data or models; Users' emotional entanglement with GAI functions.	Human-AI Configuration; Information Security; Value Chain and Component Integration
AI Actor Tasks: AI Deployment, Operation and Monitoring		

GOVERN 2.1: Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.		
Action ID	**Suggested Action**	**GAI Risks**
GV-2.1-001	Establish organizational roles, policies, and procedures for communicating GAI incidents and performance to AI Actors and downstream stakeholders (including those potentially impacted), via community or official resources (e.g., AI incident database, AVID, CVE, NVD, or OECD AI incident monitor).	Human-AI Configuration; Value Chain and Component Integration
GV-2.1-002	Establish procedures to engage teams for GAI system incident response with diverse composition and responsibilities based on the particular incident type.	Harmful Bias and Homogenization
GV-2.1-003	Establish processes to verify the AI Actors conducting GAI incident response tasks demonstrate and maintain the appropriate skills and training.	Human-AI Configuration
GV-2.1-004	When systems may raise national security risks, involve national security professionals in mapping, measuring, and managing those risks.	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Information Security
GV-2.1-005	Create mechanisms to provide protections for whistleblowers who report, based on reasonable belief, when the organization violates relevant laws or poses a specific and empirically well-substantiated negative risk to public safety (or has already caused harm).	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content
AI Actor Tasks: Governance and Oversight		

GOVERN 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.		
Action ID	**Suggested Action**	**GAI Risks**
GV-3.2-001	Policies are in place to bolster oversight of GAI systems with independent evaluations or assessments of GAI models or systems where the type and robustness of evaluations are proportional to the identified risks.	CBRN Information or Capabilities; Harmful Bias and Homogenization
GV-3.2-002	Consider adjustment of organizational roles and components across lifecycle stages of large or complex GAI systems, including: Test and evaluation, validation, and red-teaming of GAI systems; GAI content moderation; GAI system development and engineering; Increased accessibility of GAI tools, interfaces, and systems, Incident response and containment.	Human-AI Configuration; Information Security; Harmful Bias and Homogenization
GV-3.2-003	Define acceptable use policies for GAI interfaces, modalities, and human-AI configurations (i.e., for chatbots and decision-making tasks), including criteria for the kinds of queries GAI applications should refuse to respond to.	Human-AI Configuration
GV-3.2-004	Establish policies for user feedback mechanisms for GAI systems which include thorough instructions and any mechanisms for recourse.	Human-AI Configuration
GV-3.2-005	Engage in threat modeling to anticipate potential risks from GAI systems.	CBRN Information or Capabilities; Information Security
AI Actors: AI Design		

GOVERN 4.1: Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.		
Action ID	**Suggested Action**	**GAI Risks**
GV-4.1-001	Establish policies and procedures that address continual improvement processes for GAI risk measurement. Address general risks associated with a lack of explainability and transparency in GAI systems by using ample documentation and techniques such as: application of gradient-based attributions, occlusion/term reduction, counterfactual prompts and prompt engineering, and analysis of embeddings; Assess and update risk measurement approaches at regular cadences.	Confabulation
GV-4.1-002	Establish policies, procedures, and processes detailing risk measurement in context of use with standardized measurement protocols and structured public feedback exercises such as AI red-teaming or independent external evaluations.	CBRN Information and Capability; Value Chain and Component Integration

GV-4.1-003	Establish policies, procedures, and processes for oversight functions (e.g., senior leadership, legal, compliance, including internal evaluation) across the GAI lifecycle, from problem formulation and supply chains to system decommission.	Value Chain and Component Integration
AI Actor Tasks: AI Deployment, AI Design, AI Development, Operation and Monitoring		

GOVERN 4.2: Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.		
Action ID	**Suggested Action**	**GAI Risks**
GV-4.2-001	Establish terms of use and terms of service for GAI systems.	Intellectual Property; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content
GV-4.2-002	Include relevant AI Actors in the GAI system risk identification process.	Human-AI Configuration
GV-4.2-003	Verify that downstream GAI system impacts (such as the use of third-party plugins) are included in the impact documentation process.	Value Chain and Component Integration
AI Actor Tasks: AI Deployment, AI Design, AI Development, Operation and Monitoring		

GOVERN 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.		
Action ID	**Suggested Action**	**GAI Risks**
GV4.3--001	Establish policies for measuring the effectiveness of employed content provenance methodologies (e.g., cryptography, watermarking, steganography, etc.)	Information Integrity
GV-4.3-002	Establish organizational practices to identify the minimum set of criteria necessary for GAI system incident reporting such as: System ID (auto-generated most likely), Title, Reporter, System/Source, Data Reported, Date of Incident, Description, Impact(s), Stakeholder(s) Impacted.	Information Security

GV-4.3-003	Verify information sharing and feedback mechanisms among individuals and organizations regarding any negative impact from GAI systems.	Information Integrity; Data Privacy
AI Actor Tasks: AI Impact Assessment, Affected Individuals and Communities, Governance and Oversight		

GOVERN 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.		
Action ID	**Suggested Action**	**GAI Risks**
GV-5.1-001	Allocate time and resources for outreach, feedback, and recourse processes in GAI system development.	Human-AI Configuration; Harmful Bias and Homogenization
GV-5.1-002	Document interactions with GAI systems to users prior to interactive activities, particularly in contexts involving more significant risks.	Human-AI Configuration; Confabulation
AI Actor Tasks: AI Design, AI Impact Assessment, Affected Individuals and Communities, Governance and Oversight		

GOVERN 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.		
Action ID	**Suggested Action**	**GAI Risks**
GV-6.1-001	Categorize different types of GAI content with associated third-party rights (e.g., copyright, intellectual property, data privacy).	Data Privacy; Intellectual Property; Value Chain and Component Integration
GV-6.1-002	Conduct joint educational activities and events in collaboration with third parties to promote best practices for managing GAI risks.	Value Chain and Component Integration
GV-6.1-003	Develop and validate approaches for measuring the success of content provenance management efforts with third parties (e.g., incidents detected and response times).	Information Integrity; Value Chain and Component Integration
GV-6.1-004	Draft and maintain well-defined contracts and service level agreements (SLAs) that specify content ownership, usage rights, quality standards, security requirements, and content provenance expectations for GAI systems.	Information Integrity; Information Security; Intellectual Property

GV-6.1-005	Implement a use-cased based supplier risk assessment framework to evaluate and monitor third-party entities' performance and adherence to content provenance standards and technologies to detect anomalies and unauthorized changes; services acquisition and value chain risk management; and legal compliance.	Data Privacy; Information Integrity; Information Security; Intellectual Property; Value Chain and Component Integration
GV-6.1-006	Include clauses in contracts which allow an organization to evaluate third-party GAI processes and standards.	Information Integrity
GV-6.1-007	Inventory all third-party entities with access to organizational content and establish approved GAI technology and service provider lists.	Value Chain and Component Integration
GV-6.1-008	Maintain records of changes to content made by third parties to promote content provenance, including sources, timestamps, metadata.	Information Integrity; Value Chain and Component Integration; Intellectual Property
GV-6.1-009	Update and integrate due diligence processes for GAI acquisition and procurement vendor assessments to include intellectual property, data privacy, security, and other risks. For example, update processes to: Address solutions that may rely on embedded GAI technologies; Address ongoing monitoring, assessments, and alerting, dynamic risk assessments, and real-time reporting tools for monitoring third-party GAI risks; Consider policy adjustments across GAI modeling libraries, tools and APIs, fine-tuned models, and embedded tools; Assess GAI vendors, open-source or proprietary GAI tools, or GAI service providers against incident or vulnerability databases.	Data Privacy; Human-AI Configuration; Information Security; Intellectual Property; Value Chain and Component Integration; Harmful Bias and Homogenization
GV-6.1-010	Update GAI acceptable use policies to address proprietary and open-source GAI technologies and data, and contractors, consultants, and other third-party personnel.	Intellectual Property; Value Chain and Component Integration
AI Actor Tasks: Operation and Monitoring, Procurement, Third-party entities		

GOVERN 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.		
Action ID	**Suggested Action**	**GAI Risks**
GV-6.2-001	Document GAI risks associated with system value chain to identify over-reliance on third-party data and to identify fallbacks.	Value Chain and Component Integration
GV-6.2-002	Document incidents involving third-party GAI data and systems, including open-data and open-source software.	Intellectual Property; Value Chain and Component Integration

GV-6.2-003	Establish incident response plans for third-party GAI technologies: Align incident response plans with impacts enumerated in MAP 5.1; Communicate third-party GAI incident response plans to all relevant AI Actors; Define ownership of GAI incident response functions; Rehearse third-party GAI incident response plans at a regular cadence; Improve incident response plans based on retrospective learning; Review incident response plans for alignment with relevant breach reporting, data protection, data privacy, or other laws.	Data Privacy; Human-AI Configuration; Information Security; Value Chain and Component Integration; Harmful Bias and Homogenization
GV-6.2-004	Establish policies and procedures for continuous monitoring of third-party GAI systems in deployment.	Value Chain and Component Integration
GV-6.2-005	Establish policies and procedures that address GAI data redundancy, including model weights and other system artifacts.	Harmful Bias and Homogenization
GV-6.2-006	Establish policies and procedures to test and manage risks related to rollover and fallback technologies for GAI systems, acknowledging that rollover and fallback may include manual processing.	Information Integrity
GV-6.2-007	Review vendor contracts and avoid arbitrary or capricious termination of critical GAI technologies or vendor services and non-standard terms that may amplify or defer liability in unexpected ways and/or contribute to unauthorized data collection by vendors or third-parties (e.g., secondary data use). Consider: Clear assignment of liability and responsibility for incidents, GAI system changes over time (e.g., fine-tuning, drift, decay); Request: Notification and disclosure for serious incidents arising from third-party data and systems; Service Level Agreements (SLAs) in vendor contracts that address incident response, response times, and availability of critical support.	Human-AI Configuration; Information Security; Value Chain and Component Integration
AI Actor Tasks: AI Deployment, Operation and Monitoring, TEVV, Third-party entities		

MAP 1.1: Intended purposes, potentially beneficial uses, context specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.

Action ID	Suggested Action	GAI Risks
MP-1.1-001	When identifying intended purposes, consider factors such as internal vs. external use, narrow vs. broad application scope, fine-tuning, and varieties of data sources (e.g., grounding, retrieval-augmented generation).	Data Privacy; Intellectual Property

MP-1.1-002	Determine and document the expected and acceptable GAI system context of use in collaboration with socio-cultural and other domain experts, by assessing: Assumptions and limitations; Direct value to the organization; Intended operational environment and observed usage patterns; Potential positive and negative impacts to individuals, public safety, groups, communities, organizations, democratic institutions, and the physical environment; Social norms and expectations.	Harmful Bias and Homogenization
MP-1.1-003	Document risk measurement plans to address identified risks. Plans may include, as applicable: Individual and group cognitive biases (e.g., confirmation bias, funding bias, groupthink) for AI Actors involved in the design, implementation, and use of GAI systems; Known past GAI system incidents and failure modes; In-context use and foreseeable misuse, abuse, and off-label use; Over reliance on quantitative metrics and methodologies without sufficient awareness of their limitations in the context(s) of use; Standard measurement and structured human feedback approaches; Anticipated human-AI configurations.	Human-AI Configuration; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
MP-1.1-004	Identify and document foreseeable illegal uses or applications of the GAI system that surpass organizational risk tolerances.	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content
AI Actor Tasks: AI Deployment		

MAP 1.2: Interdisciplinary AI Actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.		
Action ID	**Suggested Action**	**GAI Risks**
MP-1.2-001	Establish and empower interdisciplinary teams that reflect a wide range of capabilities, competencies, demographic groups, domain expertise, educational backgrounds, lived experiences, professions, and skills across the enterprise to inform and conduct risk measurement and management functions.	Human-AI Configuration; Harmful Bias and Homogenization
MP-1.2-002	Verify that data or benchmarks used in risk measurement, and users, participants, or subjects involved in structured GAI public feedback exercises are representative of diverse in-context user populations.	Human-AI Configuration; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment		

MAP 2.1: The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).		
Action ID	**Suggested Action**	**GAI Risks**
MP-2.1-001	Establish known assumptions and practices for determining data origin and content lineage, for documentation and evaluation purposes.	Information Integrity
MP-2.1-002	Institute test and evaluation for data and content flows within the GAI system, including but not limited to, original data sources, data transformations, and decision-making criteria.	Intellectual Property; Data Privacy
AI Actor Tasks: TEVV		

MAP 2.2: Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI Actors when making decisions and taking subsequent actions.		
Action ID	**Suggested Action**	**GAI Risks**
MP-2.2-001	Identify and document how the system relies on upstream data sources, including for content provenance, and if it serves as an upstream dependency for other systems.	Information Integrity; Value Chain and Component Integration
MP-2.2-002	Observe and analyze how the GAI system interacts with external networks, and identify any potential for negative externalities, particularly where content provenance might be compromised.	Information Integrity
AI Actor Tasks: End Users		

MAP 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation		
Action ID	**Suggested Action**	**GAI Risks**
MP-2.3-001	Assess the accuracy, quality, reliability, and authenticity of GAI output by comparing it to a set of known ground truth data and by using a variety of evaluation methods (e.g., human oversight and automated evaluation, proven cryptographic techniques, review of content inputs).	Information Integrity

MP-2.3-002	Review and document accuracy, representativeness, relevance, suitability of data used at different stages of AI life cycle.	Harmful Bias and Homogenization; Intellectual Property
MP-2.3-003	Deploy and document fact-checking techniques to verify the accuracy and veracity of information generated by GAI systems, especially when the information comes from multiple (or unknown) sources.	Information Integrity
MP-2.3-004	Develop and implement testing techniques to identify GAI produced content (e.g., synthetic media) that might be indistinguishable from human-generated content.	Information Integrity
MP-2.3-005	Implement plans for GAI systems to undergo regular adversarial testing to identify vulnerabilities and potential manipulation or misuse.	Information Security
AI Actor Tasks: AI Development, Domain Experts, TEVV		

MAP 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MP-3.4-001	Evaluate whether GAI operators and end-users can accurately understand content lineage and origin.	Human-AI Configuration; Information Integrity
MP-3.4-002	Adapt existing training programs to include modules on digital content transparency.	Information Integrity
MP-3.4-003	Develop certification programs that test proficiency in managing GAI risks and interpreting content provenance, relevant to specific industry and context.	Information Integrity
MP-3.4-004	Delineate human proficiency tests from tests of GAI capabilities.	Human-AI Configuration
MP-3.4-005	Implement systems to continually monitor and track the outcomes of human-GAI configurations for future refinement and improvements.	Human-AI Configuration; Information Integrity
MP-3.4-006	Involve the end-users, practitioners, and operators in GAI system in prototyping and testing activities. Make sure these tests cover various scenarios, such as crisis situations or ethically sensitive contexts.	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
AI Actor Tasks: AI Design, AI Development, Domain Experts, End-Users, Human Factors, Operation and Monitoring		

MAP 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.		
Action ID	**Suggested Action**	**GAI Risks**
MP-4.1-001	Conduct periodic monitoring of AI-generated content for privacy risks; address any possible instances of PII or sensitive data exposure.	Data Privacy
MP-4.1-002	Implement processes for responding to potential intellectual property infringement claims or other rights.	Intellectual Property
MP-4.1-003	Connect new GAI policies, procedures, and processes to existing model, data, software development, and IT governance and to legal, compliance, and risk management activities.	Information Security; Data Privacy
MP-4.1-004	Document training data curation policies, to the extent possible and according to applicable laws and policies.	Intellectual Property; Data Privacy; Obscene, Degrading, and/or Abusive Content
MP-4.1-005	Establish policies for collection, retention, and minimum quality of data, in consideration of the following risks: Disclosure of inappropriate CBRN information; Use of Illegal or dangerous content; Offensive cyber capabilities; Training data imbalances that could give rise to harmful biases; Leak of personally identifiable information, including facial likenesses of individuals.	CBRN Information or Capabilities; Intellectual Property; Information Security; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; Data Privacy
MP-4.1-006	Implement policies and practices defining how third-party intellectual property and training data will be used, stored, and protected.	Intellectual Property; Value Chain and Component Integration
MP-4.1-007	Re-evaluate models that were fine-tuned or enhanced on top of third-party models.	Value Chain and Component Integration
MP-4.1-008	Re-evaluate risks when adapting GAI models to new domains. Additionally, establish warning systems to determine if a GAI system is being used in a new domain where previous assumptions (relating to context of use or mapped risks such as security, and safety) may no longer hold.	CBRN Information or Capabilities; Intellectual Property; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; Data Privacy
MP-4.1-009	Leverage approaches to detect the presence of PII or sensitive data in generated output text, image, video, or audio.	Data Privacy

MP-4.1-010	Conduct appropriate diligence on training data use to assess intellectual property, and privacy, risks, including to examine whether use of proprietary or sensitive training data is consistent with applicable laws.	Intellectual Property; Data Privacy
AI Actor Tasks: Governance and Oversight, Operation and Monitoring, Procurement, Third-party entities		

MAP 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MP-5.1-001	Apply TEVV practices for content provenance (e.g., probing a system's synthetic data generation capabilities for potential misuse or vulnerabilities.	Information Integrity; Information Security
MP-5.1-002	Identify potential content provenance harms of GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content. Enumerate and rank risks based on their likelihood and potential impact, and determine how well provenance solutions address specific risks and/or harms.	Information Integrity; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content
MP-5.1-003	Consider disclosing use of GAI to end users in relevant contexts, while considering the objective of disclosure, the context of use, the likelihood and magnitude of the risk posed, the audience of the disclosure, as well as the frequency of the disclosures.	Human-AI Configuration
MP-5.1-004	Prioritize GAI structured public feedback processes based on risk assessment estimates.	Information Integrity; CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Harmful Bias and Homogenization
MP-5.1-005	Conduct adversarial role-playing exercises, GAI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes.	Information Security
MP-5.1-006	Profile threats and negative impacts arising from GAI systems interacting with, manipulating, or generating content, and outlining known and potential vulnerabilities and the likelihood of their occurrence.	Information Security
AI Actor Tasks: AI Deployment, AI Design, AI Development, AI Impact Assessment, Affected Individuals and Communities, End-Users, Operation and Monitoring		

MAP 5.2: Practices and personnel for supporting regular engagement with relevant AI Actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MP-5.2-001	Determine context-based measures to identify if new impacts are present due to the GAI system, including regular engagements with downstream AI Actors to identify and quantify new contexts of unanticipated impacts of GAI systems.	Human-AI Configuration; Value Chain and Component Integration
MP-5.2-002	Plan regular engagements with AI Actors responsible for inputs to GAI systems, including third-party data and algorithms, to review and evaluate unanticipated impacts.	Human-AI Configuration; Value Chain and Component Integration
AI Actor Tasks: AI Deployment, AI Design, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End-Users, Human Factors, Operation and Monitoring		

MEASURE 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-1.1-001	Employ methods to trace the origin and modifications of digital content.	Information Integrity
MS-1.1-002	Integrate tools designed to analyze content provenance and detect data anomalies, verify the authenticity of digital signatures, and identify patterns associated with misinformation or manipulation.	Information Integrity
MS-1.1-003	Disaggregate evaluation metrics by demographic factors to identify any discrepancies in how content provenance mechanisms work across diverse populations.	Information Integrity; Harmful Bias and Homogenization
MS-1.1-004	Develop a suite of metrics to evaluate structured public feedback exercises informed by representative AI Actors.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities
MS-1.1-005	Evaluate novel methods and technologies for the measurement of GAI-related risks including in content provenance, offensive cyber, and CBRN, while maintaining the models' ability to produce valid, reliable, and factually accurate outputs.	Information Integrity; CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content

MS-1.1-006	Implement continuous monitoring of GAI system impacts to identify whether GAI outputs are equitable across various sub-populations. Seek active and direct feedback from affected communities via structured feedback mechanisms or red-teaming to monitor and improve outputs.	Harmful Bias and Homogenization
MS-1.1-007	Evaluate the quality and integrity of data used in training and the provenance of AI-generated content, for example by employing techniques like chaos engineering and seeking stakeholder feedback.	Information Integrity
MS-1.1-008	Define use cases, contexts of use, capabilities, and negative impacts where structured human feedback exercises, e.g., GAI red-teaming, would be most beneficial for GAI risk measurement and management based on the context of use.	Harmful Bias and Homogenization; CBRN Information or Capabilities
MS-1.1-009	Track and document risks or opportunities related to all GAI risks that cannot be measured quantitatively, including explanations as to why some risks cannot be measured (e.g., due to technological limitations, resource constraints, or trustworthy considerations). Include unmeasured risks in marginal risks.	Information Integrity
AI Actor Tasks: AI Development, Domain Experts, TEVV		

MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI Actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.

Action ID	Suggested Action	GAI Risks
MS-1.3-001	Define relevant groups of interest (e.g., demographic groups, subject matter experts, experience with GAI technology) within the context of use as part of plans for gathering structured public feedback.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities
MS-1.3-002	Engage in internal and external evaluations, GAI red-teaming, impact assessments, or other structured human feedback exercises in consultation with representative AI Actors with expertise and familiarity in the context of use, and/or who are representative of the populations associated with the context of use.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities
MS-1.3-003	Verify those conducting structured human feedback exercises are not directly involved in system development tasks for the same GAI model.	Human-AI Configuration; Data Privacy
AI Actor Tasks: AI Deployment, AI Development, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End-Users, Operation and Monitoring, TEVV		

MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.2-001	Assess and manage statistical biases related to GAI content provenance through techniques such as re-sampling, re-weighting, or adversarial training.	Information Integrity; Information Security; Harmful Bias and Homogenization
MS-2.2-002	Document how content provenance data is tracked and how that data interacts with privacy and security. Consider: Anonymizing data to protect the privacy of human subjects; Leveraging privacy output filters; Removing any personally identifiable information (PII) to prevent potential harm or misuse.	Data Privacy; Human AI Configuration; Information Integrity; Information Security; Dangerous, Violent, or Hateful Content
MS-2.2-003	Provide human subjects with options to withdraw participation or revoke their consent for present or future use of their data in GAI applications.	Data Privacy; Human-AI Configuration; Information Integrity
MS-2.2-004	Use techniques such as anonymization, differential privacy or other privacy-enhancing technologies to minimize the risks associated with linking AI-generated content back to individual human subjects.	Data Privacy; Human-AI Configuration
AI Actor Tasks: AI Development, Human Factors, TEVV		

MEASURE 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.3-001	Consider baseline model performance on suites of benchmarks when selecting a model for fine tuning or enhancement with retrieval-augmented generation.	Information Security; Confabulation
MS-2.3-002	Evaluate claims of model capabilities using empirically validated methods.	Confabulation; Information Security
MS-2.3-003	Share results of pre-deployment testing with relevant GAI Actors, such as those with system release approval authority.	Human-AI Configuration

MS-2.3-004	Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.	CBRN Information or Capabilities; Data Privacy; Confabulation; Information Integrity; Information Security; Dangerous, Violent, or Hateful Content; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, TEVV		

MEASURE 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.		
Action ID	**Suggested Action**	**Risks**
MS-2.5-001	Avoid extrapolating GAI system performance or capabilities from narrow, non-systematic, and anecdotal assessments.	Human-AI Configuration; Confabulation
MS-2.5-002	Document the extent to which human domain knowledge is employed to improve GAI system performance, via, e.g., RLHF, fine-tuning, retrieval-augmented generation, content moderation, business rules.	Human-AI Configuration
MS-2.5-003	Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities.	Confabulation
MS-2.5-004	Track and document instances of anthropomorphization (e.g., human images, mentions of human feelings, cyborg imagery or motifs) in GAI system interfaces.	Human-AI Configuration
MS-2.5-005	Verify GAI system training data and TEVV data provenance, and that fine-tuning or retrieval-augmented generation data is grounded.	Information Integrity
MS-2.5-006	Regularly review security and safety guardrails, especially if the GAI system is being operated in novel circumstances. This includes reviewing reasons why the GAI system was initially assessed as being safe to deploy.	Information Security; Dangerous, Violent, or Hateful Content
AI Actor Tasks: Domain Experts, TEVV		

MEASURE 2.6: The AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.6-001	Assess adverse impacts, including health and wellbeing impacts for value chain or other AI Actors that are exposed to sexually explicit, offensive, or violent information during GAI training and maintenance.	Human-AI Configuration; Obscene, Degrading, and/or Abusive Content; Value Chain and Component Integration; Dangerous, Violent, or Hateful Content
MS-2.6-002	Assess existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data.	Data Privacy; Intellectual Property; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities
MS-2.6-003	Re-evaluate safety features of fine-tuned models when the negative risk exceeds organizational risk tolerance.	Dangerous, Violent, or Hateful Content
MS-2.6-004	Review GAI system outputs for validity and safety: Review generated code to assess risks that may arise from unreliable downstream decision-making.	Value Chain and Component Integration; Dangerous, Violent, or Hateful Content
MS-2.6-005	Verify that GAI system architecture can monitor outputs and performance, and handle, recover from, and repair errors when security anomalies, threats and impacts are detected.	Confabulation; Information Integrity; Information Security
MS-2.6-006	Verify that systems properly handle queries that may give rise to inappropriate, malicious, or illegal usage, including facilitating manipulation, extortion, targeted impersonation, cyber-attacks, and weapons creation.	CBRN Information or Capabilities; Information Security
MS-2.6-007	Regularly evaluate GAI system vulnerabilities to possible circumvention of safety measures.	CBRN Information or Capabilities; Information Security
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.		
Action ID	**Action**	**GAI Risks**
MS-2.7-001	Apply established security measures to: Assess likelihood and magnitude of vulnerabilities and threats such as backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering, autonomous agents, model theft or exposure of model weights, AI inference, bypass, extraction, and other baseline security concerns.	Data Privacy; Information Integrity; Information Security; Value Chain and Component Integration
MS-2.7-002	Benchmark GAI system security and resilience related to content provenance against industry standards and best practices. Compare GAI system security features and content provenance methods against industry state-of-the-art.	Information Integrity; Information Security
MS-2.7-003	Conduct user surveys to gather user satisfaction with the AI-generated content and user perceptions of content authenticity. Analyze user feedback to identify concerns and/or current literacy levels related to content provenance and understanding of labels on content.	Human-AI Configuration; Information Integrity
MS-2.7-004	Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, inference, bypass, extraction, penetrations, or provenance verification.	Information Integrity; Information Security
MS-2.7-005	Measure reliability of content authentication methods, such as watermarking, cryptographic signatures, digital fingerprints, as well as access controls, conformity assessment, and model integrity verification, which can help support the effective implementation of content provenance techniques. Evaluate the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification.	Information Integrity
MS-2.7-006	Measure the rate at which recommendations from security checks and incidents are implemented. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback.	Information Integrity; Information Security
MS-2.7-007	Perform AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, sponge examples).	Information Security; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
MS-2.7-008	Verify fine-tuning does not compromise safety and security controls.	Information Integrity; Information Security; Dangerous, Violent, or Hateful Content

MS-2.7-009	Regularly assess and verify that security measures remain effective and have not been compromised.	Information Security
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 2.8: Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.8-001	Compile statistics on actual policy violations, take-down requests, and intellectual property infringement for organizational GAI systems: Analyze transparency reports across demographic groups, languages groups.	Intellectual Property; Harmful Bias and Homogenization
MS-2.8-002	Document the instructions given to data annotators or AI red-teamers.	Human-AI Configuration
MS-2.8-003	Use digital content transparency solutions to enable the documentation of each instance where content is generated, modified, or shared to provide a tamper-proof history of the content, promote transparency, and enable traceability. Robust version control systems can also be applied to track changes across the AI lifecycle over time.	Information Integrity
MS-2.8-004	Verify adequacy of GAI system user instructions through user testing.	Human-AI Configuration
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the MAP function – to inform responsible use and governance.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.9-001	Apply and document ML explanation results such as: Analysis of embeddings, Counterfactual prompts, Gradient-based attributions, Model compression/surrogate models, Occlusion/term reduction.	Confabulation
MS-2.9-002	Document GAI model details including: Proposed use and organizational value; Assumptions and limitations, Data collection methodologies; Data provenance; Data quality; Model architecture (e.g., convolutional neural network, transformers, etc.); Optimization objectives; Training algorithms; RLHF approaches; Fine-tuning or retrieval-augmented generation approaches; Evaluation data; Ethical considerations; Legal and regulatory requirements.	Information Integrity; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, End-Users, Operation and Monitoring, TEVV		

MEASURE 2.10: Privacy risk of the AI system – as identified in the MAP function – is examined and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.10-001	Conduct AI red-teaming to assess issues such as: Outputting of training data samples, and subsequent reverse engineering, model extraction, and membership inference risks; Revealing biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information; Tracking or revealing location information of users or members of training datasets.	Human-AI Configuration; Information Integrity; Intellectual Property
MS-2.10-002	Engage directly with end-users and other stakeholders to understand their expectations and concerns regarding content provenance. Use this feedback to guide the design of provenance data-tracking techniques.	Human-AI Configuration; Information Integrity
MS-2.10-003	Verify deduplication of GAI training data samples, particularly regarding synthetic data.	Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, End-Users, Operation and Monitoring, TEVV		

MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.11-001	Apply use-case appropriate benchmarks (e.g., Bias Benchmark Questions, Real Hateful or Harmful Prompts, Winogender Schemas[15]) to quantify systemic bias, stereotyping, denigration, and hateful content in GAI system outputs; Document assumptions and limitations of benchmarks, including any actual or possible training/test data cross contamination, relative to in-context deployment environment.	Harmful Bias and Homogenization
MS-2.11-002	Conduct fairness assessments to measure systemic bias. Measure GAI system performance across demographic groups and subgroups, addressing both quality of service and any allocation of services and resources. Quantify harms using: field testing with sub-group populations to determine likelihood of exposure to generated content exhibiting harmful bias, AI red-teaming with counterfactual and low-context (e.g., "leader," "bad guys") prompts. For ML pipelines or business processes with categorical or numeric outcomes that rely on GAI, apply general fairness metrics (e.g., demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), to the pipeline or business outcome where appropriate; Custom, context-specific metrics developed in collaboration with domain experts and affected communities; Measurements of the prevalence of denigration in generated content in deployment (e.g., sub-sampling a fraction of traffic and manually annotating denigrating content).	Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
MS-2.11-003	Identify the classes of individuals, groups, or environmental ecosystems which might be impacted by GAI systems through direct engagement with potentially impacted communities.	Environmental; Harmful Bias and Homogenization
MS-2.11-004	Review, document, and measure sources of bias in GAI training and TEVV data: Differences in distributions of outcomes across and within groups, including intersecting groups; Completeness, representativeness, and balance of data sources; demographic group and subgroup coverage in GAI system training data; Forms of latent systemic bias in images, text, audio, embeddings, or other complex or unstructured data; Input data features that may serve as proxies for demographic group membership (i.e., image metadata, language dialect) or otherwise give rise to emergent bias within GAI systems; The extent to which the digital divide may negatively impact representativeness in GAI system training and TEVV data; Filtering of hate speech or content in GAI system training data; Prevalence of GAI-generated data in GAI system training data.	Harmful Bias and Homogenization

[15] Winogender Schemas is a sample set of paired sentences which differ only by gender of the pronouns used, which can be used to evaluate gender bias in natural language processing coreference resolution systems.

MS-2.11-005	Assess the proportion of synthetic to non-synthetic training data and verify training data is not overly homogenous or GAI-produced to mitigate concerns of model collapse.	Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End-Users, Operation and Monitoring, TEVV		

MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.12-001	Assess safety to physical environments when deploying GAI systems.	Dangerous, Violent, or Hateful Content
MS-2.12-002	Document anticipated environmental impacts of model development, maintenance, and deployment in product design decisions.	Environmental
MS-2.12-003	Measure or estimate environmental impacts (e.g., energy and water consumption) for training, fine tuning, and deploying models: Verify tradeoffs between resources used at inference time versus additional resources required at training time.	Environmental
MS-2.12-004	Verify effectiveness of carbon capture or offset programs for GAI training and applications, and address green-washing concerns.	Environmental
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 2.13: Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-2.13-001	Create measurement error models for pre-deployment metrics to demonstrate construct validity for each metric (i.e., does the metric effectively operationalize the desired concept): Measure or estimate, and document, biases or statistical variance in applied metrics or structured human feedback processes; Leverage domain expertise when modeling complex societal constructs such as hateful content.	Confabulation; Information Integrity; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, Operation and Monitoring, TEVV		

MEASURE 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.		
Action ID	**Suggested Action**	**GAI Risks**
MS-3.2-001	Establish processes for identifying emergent GAI system risks including consulting with external AI Actors.	Human-AI Configuration; Confabulation
AI Actor Tasks: AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.		
Action ID	**Suggested Action**	**GAI Risks**
MS-3.3-001	Conduct impact assessments on how AI-generated content might affect different social, economic, and cultural groups.	Harmful Bias and Homogenization
MS-3.3-002	Conduct studies to understand how end users perceive and interact with GAI content and accompanying content provenance within context of use. Assess whether the content aligns with their expectations and how they may act upon the information presented.	Human-AI Configuration; Information Integrity
MS-3.3-003	Evaluate potential biases and stereotypes that could emerge from the AI-generated content using appropriate methodologies including computational testing methods as well as evaluating structured feedback input.	Harmful Bias and Homogenization

MS-3.3-004	Provide input for training materials about the capabilities and limitations of GAI systems related to digital content transparency for AI Actors, other professionals, and the public about the societal impacts of AI and the role of diverse and inclusive content generation.	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization
MS-3.3-005	Record and integrate structured feedback about content provenance from operators, users, and potentially impacted communities through the use of methods such as user research studies, focus groups, or community forums. Actively seek feedback on generated content quality and potential biases. Assess the general awareness among end users and impacted communities about the availability of these feedback channels.	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, End-Users, Operation and Monitoring, TEVV		

MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI Actors to validate whether the system is performing consistently as intended. Results are documented.		
Action ID	**Suggested Action**	**GAI Risks**
MS-4.2-001	Conduct adversarial testing at a regular cadence to map and measure GAI risks, including tests to address attempts to deceive or manipulate the application of provenance techniques or other misuses. Identify vulnerabilities and understand potential misuse scenarios and unintended outputs.	Information Integrity; Information Security
MS-4.2-002	Evaluate GAI system performance in real-world scenarios to observe its behavior in practical environments and reveal issues that might not surface in controlled and optimized testing environments.	Human-AI Configuration; Confabulation; Information Security
MS-4.2-003	Implement interpretability and explainability methods to evaluate GAI system decisions and verify alignment with intended purpose.	Information Integrity; Harmful Bias and Homogenization
MS-4.2-004	Monitor and document instances where human operators or other systems override the GAI's decisions. Evaluate these cases to understand if the overrides are linked to issues related to content provenance.	Information Integrity
MS-4.2-005	Verify and document the incorporation of results of structured public feedback exercises into design, implementation, deployment approval (“go”/“no-go” decisions), monitoring, and decommission decisions.	Human-AI Configuration; Information Security
AI Actor Tasks: AI Deployment, Domain Experts, End-Users, Operation and Monitoring, TEVV		

MANAGE 1.3: Responses to the AI risks deemed high priority, as identified by the MAP function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.		
Action ID	**Suggested Action**	**GAI Risks**
MG-1.3-001	Document trade-offs, decision processes, and relevant measurement and feedback results for risks that do not surpass organizational risk tolerance, for example, in the context of model release: Consider different approaches for model release, for example, leveraging a staged release approach. Consider release approaches in the context of the model and its projected use cases. Mitigate, transfer, or avoid risks that surpass organizational risk tolerances.	Information Security
MG-1.3-002	Monitor the robustness and effectiveness of risk controls and mitigation plans (e.g., via red-teaming, field testing, participatory engagements, performance assessments, user feedback mechanisms).	Human-AI Configuration
AI Actor Tasks: AI Development, AI Deployment, AI Impact Assessment, Operation and Monitoring		

MANAGE 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.		
Action ID	**Suggested Action**	**GAI Risks**
MG-2.2-001	Compare GAI system outputs against pre-defined organization risk tolerance, guidelines, and principles, and review and test AI-generated content against these guidelines.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content
MG-2.2-002	Document training data sources to trace the origin and provenance of AI-generated content.	Information Integrity
MG-2.2-003	Evaluate feedback loops between GAI system content provenance and human reviewers, and update where needed. Implement real-time monitoring systems to affirm that content provenance protocols remain effective.	Information Integrity
MG-2.2-004	Evaluate GAI content and data for representational biases and employ techniques such as re-sampling, re-ranking, or adversarial training to mitigate biases in the generated content.	Information Security; Harmful Bias and Homogenization
MG-2.2-005	Engage in due diligence to analyze GAI output for harmful content, potential misinformation, and CBRN-related or NCII content.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content

MG-2.2-006	Use feedback from internal and external AI Actors, users, individuals, and communities, to assess impact of AI-generated content.	Human-AI Configuration
MG-2.2-007	Use real-time auditing tools where they can be demonstrated to aid in the tracking and validation of the lineage and authenticity of AI-generated data.	Information Integrity
MG-2.2-008	Use structured feedback mechanisms to solicit and capture user input about AI-generated content to detect subtle shifts in quality or alignment with community and societal values.	Human-AI Configuration; Harmful Bias and Homogenization
MG-2.2-009	Consider opportunities to responsibly use synthetic data and other privacy enhancing techniques in GAI development, where appropriate and applicable, match the statistical properties of real-world data without disclosing personally identifiable information or contributing to homogenization.	Data Privacy; Intellectual Property; Information Integrity; Confabulation; Harmful Bias and Homogenization
AI Actor Tasks: AI Deployment, AI Impact Assessment, Governance and Oversight, Operation and Monitoring		

MANAGE 2.3: Procedures are followed to respond to and recover from a previously unknown risk when it is identified.		
Action ID	**Suggested Action**	**GAI Risks**
MG-2.3-001	Develop and update GAI system incident response and recovery plans and procedures to address the following: Review and maintenance of policies and procedures to account for newly encountered uses; Review and maintenance of policies and procedures for detection of unanticipated uses; Verify response and recovery plans account for the GAI system value chain; Verify response and recovery plans are updated for and include necessary details to communicate with downstream GAI system Actors: Points-of-Contact (POC), Contact information, notification format.	Value Chain and Component Integration
AI Actor Tasks: AI Deployment, Operation and Monitoring		

MANAGE 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.		
Action ID	**Suggested Action**	**GAI Risks**
MG-2.4-001	Establish and maintain communication plans to inform AI stakeholders as part of the deactivation or disengagement process of a specific GAI system (including for open-source models) or context of use, including reasons, workarounds, user access removal, alternative processes, contact information, etc.	Human-AI Configuration

MG-2.4-002	Establish and maintain procedures for escalating GAI system incidents to the organizational risk management authority when specific criteria for deactivation or disengagement is met for a particular context of use or for the GAI system as a whole.	Information Security
MG-2.4-003	Establish and maintain procedures for the remediation of issues which trigger incident response processes for the use of a GAI system, and provide stakeholders timelines associated with the remediation plan.	Information Security
MG-2.4-004	Establish and regularly review specific criteria that warrants the deactivation of GAI systems in accordance with set risk tolerances and appetites.	Information Security
AI Actor Tasks: AI Deployment, Governance and Oversight, Operation and Monitoring		

MANAGE 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MG-3.1-001	Apply organizational risk tolerances and controls (e.g., acquisition and procurement processes; assessing personnel credentials and qualifications, performing background checks; filtering GAI input and outputs, grounding, fine tuning, retrieval-augmented generation) to third-party GAI resources: Apply organizational risk tolerance to the utilization of third-party datasets and other GAI resources; Apply organizational risk tolerances to fine-tuned third-party models; Apply organizational risk tolerance to existing third-party models adapted to a new domain; Reassess risk measurements after fine-tuning third-party GAI models.	Value Chain and Component Integration; Intellectual Property
MG-3.1-002	Test GAI system value chain risks (e.g., data poisoning, malware, other software and hardware vulnerabilities; labor practices; data privacy and localization compliance; geopolitical alignment).	Data Privacy; Information Security; Value Chain and Component Integration; Harmful Bias and Homogenization
MG-3.1-003	Re-assess model risks after fine-tuning or retrieval-augmented generation implementation and for any third-party GAI models deployed for applications and/or use cases that were not evaluated in initial testing.	Value Chain and Component Integration
MG-3.1-004	Take reasonable measures to review training data for CBRN information, and intellectual property, and where appropriate, remove it. Implement reasonable measures to prevent, flag, or take other action in response to outputs that reproduce particular training data (e.g., plagiarized, trademarked, patented, licensed content or trade secret material).	Intellectual Property; CBRN Information or Capabilities

MG-3.1-005	Review various transparency artifacts (e.g., system cards and model cards) for third-party models.	Information Integrity; Information Security; Value Chain and Component Integration
AI Actor Tasks: AI Deployment, Operation and Monitoring, Third-party entities		

MANAGE 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.		
Action ID	**Suggested Action**	**GAI Risks**
MG-3.2-001	Apply explainable AI (XAI) techniques (e.g., analysis of embeddings, model compression/distillation, gradient-based attributions, occlusion/term reduction, counterfactual prompts, word clouds) as part of ongoing continuous improvement processes to mitigate risks related to unexplainable GAI systems.	Harmful Bias and Homogenization
MG-3.2-002	Document how pre-trained models have been adapted (e.g., fine-tuned, or retrieval-augmented generation) for the specific generative task, including any data augmentations, parameter adjustments, or other modifications. Access to un-tuned (baseline) models supports debugging the relative influence of the pre-trained weights compared to the fine-tuned model weights or other system updates.	Information Integrity; Data Privacy
MG-3.2-003	Document sources and types of training data and their origins, potential biases present in the data related to the GAI application and its content provenance, architecture, training process of the pre-trained model including information on hyperparameters, training duration, and any fine-tuning or retrieval-augmented generation processes applied.	Information Integrity; Harmful Bias and Homogenization; Intellectual Property
MG-3.2-004	Evaluate user reported problematic content and integrate feedback into system updates.	Human-AI Configuration, Dangerous, Violent, or Hateful Content
MG-3.2-005	Implement content filters to prevent the generation of inappropriate, harmful, false, illegal, or violent content related to the GAI application, including for CSAM and NCII. These filters can be rule-based or leverage additional machine learning models to flag problematic inputs and outputs.	Information Integrity; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content
MG-3.2-006	Implement real-time monitoring processes for analyzing generated content performance and trustworthiness characteristics related to content provenance to identify deviations from the desired standards and trigger alerts for human intervention.	Information Integrity

MG-3.2-007	Leverage feedback and recommendations from organizational boards or committees related to the deployment of GAI applications and content provenance when using third-party pre-trained models.	Information Integrity; Value Chain and Component Integration
MG-3.2-008	Use human moderation systems where appropriate to review generated content in accordance with human-AI configuration policies established in the Govern function, aligned with socio-cultural norms in the context of use, and for settings where AI models are demonstrated to perform poorly.	Human-AI Configuration
MG-3.2-009	Use organizational risk tolerance to evaluate acceptable risks and performance metrics and decommission or retrain pre-trained models that perform outside of defined limits.	CBRN Information or Capabilities; Confabulation
AI Actor Tasks: AI Deployment, Operation and Monitoring, Third-party entities		

MANAGE 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI Actors, appeal and override, decommissioning, incident response, recovery, and change management.		
Action ID	**Suggested Action**	**GAI Risks**
MG-4.1-001	Collaborate with external researchers, industry experts, and community representatives to maintain awareness of emerging best practices and technologies in measuring and managing identified risks.	Information Integrity; Harmful Bias and Homogenization
MG-4.1-002	Establish, maintain, and evaluate effectiveness of organizational processes and procedures for post-deployment monitoring of GAI systems, particularly for potential confabulation, CBRN, or cyber risks.	CBRN Information or Capabilities; Confabulation; Information Security
MG-4.1-003	Evaluate the use of sentiment analysis to gauge user sentiment regarding GAI content performance and impact, and work in collaboration with AI Actors experienced in user research and experience.	Human-AI Configuration
MG-4.1-004	Implement active learning techniques to identify instances where the model fails or produces unexpected outputs.	Confabulation
MG-4.1-005	Share transparency reports with internal and external stakeholders that detail steps taken to update the GAI system to enhance transparency and accountability.	Human-AI Configuration; Harmful Bias and Homogenization
MG-4.1-006	Track dataset modifications for provenance by monitoring data deletions, rectification requests, and other changes that may impact the verifiability of content origins.	Information Integrity

MG-4.1-007	Verify that AI Actors responsible for monitoring reported issues can effectively evaluate GAI system performance including the application of content provenance data tracking techniques, and promptly escalate issues for response.	Human-AI Configuration; Information Integrity
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, Domain Experts, End-Users, Human Factors, Operation and Monitoring		

MANAGE 4.2: Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI Actors.		
Action ID	**Suggested Action**	**GAI Risks**
MG-4.2-001	Conduct regular monitoring of GAI systems and publish reports detailing the performance, feedback received, and improvements made.	Harmful Bias and Homogenization
MG-4.2-002	Practice and follow incident response plans for addressing the generation of inappropriate or harmful content and adapt processes based on findings to prevent future occurrences. Conduct post-mortem analyses of incidents with relevant AI Actors, to understand the root causes and implement preventive measures.	Human-AI Configuration; Dangerous, Violent, or Hateful Content
MG-4.2-003	Use visualizations or other methods to represent GAI model behavior to ease non-technical stakeholders understanding of GAI system functionality.	Human-AI Configuration
AI Actor Tasks: AI Deployment, AI Design, AI Development, Affected Individuals and Communities, End-Users, Operation and Monitoring, TEVV		

MANAGE 4.3: Incidents and errors are communicated to relevant AI Actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.		
Action ID	**Suggested Action**	**GAI Risks**
MG-4.3-001	Conduct after-action assessments for GAI system incidents to verify incident response and recovery processes are followed and effective, including to follow procedures for communicating incidents to relevant AI Actors and where applicable, relevant legal and regulatory bodies.	Information Security
MG-4.3-002	Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts.	Confabulation; Information Integrity

MG-4.3-003	Report GAI incidents in compliance with legal and regulatory requirements (e.g., HIPAA breach reporting, e.g., OCR (2023) or NHTSA (2022) autonomous vehicle crash reporting requirements.	Information Security; Data Privacy
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, Domain Experts, End-Users, Human Factors, Operation and Monitoring		

Appendix A. Primary GAI Considerations

The following primary considerations were derived as overarching themes from the GAI PWG consultation process. These considerations (Governance, Pre-Deployment Testing, Content Provenance, and Incident Disclosure) are relevant for voluntary use by any organization designing, developing, and using GAI and also inform the Actions to Manage GAI risks. Information included about the primary considerations is not exhaustive, but highlights the most relevant topics derived from the GAI PWG.

Acknowledgments: These considerations could not have been surfaced without the helpful analysis and contributions from the community and NIST staff GAI PWG leads: George Awad, Luca Belli, Harold Booth, Mat Heyman, Yooyoung Lee, Mark Pryzbocki, Reva Schwartz, Martin Stanley, and Kyra Yee.

A.1. Governance

A.1.1. Overview

Like any other technology system, governance principles and techniques can be used to manage risks related to generative AI models, capabilities, and applications. Organizations may choose to apply their existing risk tiering to GAI systems, or they may opt to revise or update AI system risk levels to address these unique GAI risks. This section describes how organizational governance regimes may be re-evaluated and adjusted for GAI contexts. It also addresses third-party considerations for governing across the AI value chain.

A.1.2. Organizational Governance

GAI opportunities, risks and long-term performance characteristics are typically less well-understood than non-generative AI tools and may be perceived and acted upon by humans in ways that vary greatly. Accordingly, GAI may call for different levels of oversight from AI Actors or different human-AI configurations in order to manage their risks effectively. Organizations' use of GAI systems may also warrant additional human review, tracking and documentation, and greater management oversight.

AI technology can produce varied outputs in multiple modalities and present many classes of user interfaces. This leads to a broader set of AI Actors interacting with GAI systems for widely differing applications and contexts of use. These can include data labeling and preparation, development of GAI models, content moderation, code generation and review, text generation and editing, image and video generation, summarization, search, and chat. These activities can take place within organizational settings or in the public domain.

Organizations can restrict AI applications that cause harm, exceed stated risk tolerances, or that conflict with their tolerances or values. Governance tools and protocols that are applied to other types of AI systems can be applied to GAI systems. These plans and actions include:

- Accessibility and reasonable accommodations
- AI actor credentials and qualifications
- Alignment to organizational values
- Auditing and assessment
- Change-management controls
- Commercial use
- Data provenance

- Data protection
- Data retention
- Consistency in use of defining key terms
- Decommissioning
- Discouraging anonymous use
- Education
- Impact assessments
- Incident response
- Monitoring
- Opt-outs
- Risk-based controls
- Risk mapping and measurement
- Science-backed TEVV practices
- Secure software development practices
- Stakeholder engagement
- Synthetic content detection and labeling tools and techniques
- Whistleblower protections
- Workforce diversity and interdisciplinary teams

Establishing acceptable use policies and guidance for the use of GAI in formal human-AI teaming settings as well as different levels of human-AI configurations can help to decrease risks arising from misuse, abuse, inappropriate repurpose, and misalignment between systems and users. These practices are just one example of adapting existing governance protocols for GAI contexts.

A.1.3. Third-Party Considerations

Organizations may seek to acquire, embed, incorporate, or use open-source or proprietary third-party GAI models, systems, or generated data for various applications across an enterprise. Use of these GAI tools and inputs has implications for all functions of the organization – including but not limited to acquisition, human resources, legal, compliance, and IT services – regardless of whether they are carried out by employees or third parties. Many of the actions cited above are relevant and options for addressing third-party considerations.

Third party GAI integrations may give rise to increased intellectual property, data privacy, or information security risks, pointing to the need for clear guidelines for transparency and risk management regarding the collection and use of third-party data for model inputs. Organizations may consider varying risk controls for foundation models, fine-tuned models, and embedded tools, enhanced processes for interacting with external GAI technologies or service providers. Organizations can apply standard or existing risk controls and processes to proprietary or open-source GAI technologies, data, and third-party service providers, including acquisition and procurement due diligence, requests for software bills of materials (SBOMs), application of service level agreements (SLAs), and statement on standards for attestation engagement (SSAE) reports to help with third-party transparency and risk management for GAI systems.

A.1.4. Pre-Deployment Testing

Overview

The diverse ways and contexts in which GAI systems may be developed, used, and repurposed complicates risk mapping and pre-deployment measurement efforts. Robust test, evaluation, validation, and verification (TEVV) processes can be iteratively applied – and documented – in early stages of the AI lifecycle and informed by representative AI Actors (see Figure 3 of the AI RMF). Until new and rigorous

early lifecycle TEVV approaches are developed and matured for GAI, organizations may use recommended "pre-deployment testing" practices to measure performance, capabilities, limits, risks, and impacts. This section describes risk measurement and estimation as part of pre-deployment TEVV, and examines the state of play for pre-deployment testing methodologies.

Limitations of Current Pre-deployment Test Approaches

Currently available pre-deployment TEVV processes used for GAI applications may be inadequate, non-systematically applied, or fail to reflect or mismatched to deployment contexts. For example, the anecdotal testing of GAI system capabilities through video games or standardized tests designed for humans (e.g., intelligence tests, professional licensing exams) does not guarantee GAI system validity or reliability in those domains. Similarly, jailbreaking or prompt engineering tests may not systematically assess validity or reliability risks.

Measurement gaps can arise from mismatches between laboratory and real-world settings. Current testing approaches often remain focused on laboratory conditions or restricted to benchmark test datasets and in silico techniques that may not extrapolate well to—or directly assess GAI impacts in real-world conditions. For example, current measurement gaps for GAI make it difficult to precisely estimate its potential ecosystem-level or longitudinal risks and related political, social, and economic impacts. Gaps between benchmarks and real-world use of GAI systems may likely be exacerbated due to prompt sensitivity and broad heterogeneity of contexts of use.

A.1.5. Structured Public Feedback

Structured public feedback can be used to evaluate whether GAI systems are performing as intended and to calibrate and verify traditional measurement methods. Examples of structured feedback include, but are not limited to:

- **Participatory Engagement Methods**: Methods used to solicit feedback from civil society groups, affected communities, and users, including focus groups, small user studies, and surveys.
- **Field Testing**: Methods used to determine how people interact with, consume, use, and make sense of AI-generated information, and subsequent actions and effects, including UX, usability, and other structured, randomized experiments.
- **AI Red-teaming:** A structured testing exercise used to probe an AI system to find flaws and vulnerabilities such as inaccurate, harmful, or discriminatory outputs, often in a controlled environment and in collaboration with system developers.

Information gathered from structured public feedback can inform design, implementation, deployment approval, maintenance, or decommissioning decisions. Results and insights gleaned from these exercises can serve multiple purposes, including improving data quality and preprocessing, bolstering governance decision making, and enhancing system documentation and debugging practices. When implementing feedback activities, organizations should follow human subjects research requirements and best practices such as informed consent and subject compensation.

Participatory Engagement Methods

On an ad hoc or more structured basis, organizations can design and use a variety of channels to engage external stakeholders in product development or review. Focus groups with select experts can provide feedback on a range of issues. Small user studies can provide feedback from representative groups or populations. Anonymous surveys can be used to poll or gauge reactions to specific features. Participatory engagement methods are often less structured than field testing or red teaming, and are more commonly used in early stages of AI or product development.

Field Testing

Field testing involves structured settings to evaluate risks and impacts and to simulate the conditions under which the GAI system will be deployed. Field style tests can be adapted from a focus on user preferences and experiences towards AI risks and impacts – both negative and positive. When carried out with large groups of users, these tests can provide estimations of the likelihood of risks and impacts in real world interactions.

Organizations may also collect feedback on outcomes, harms, and user experience directly from users in the production environment after a model has been released, in accordance with human subject standards such as informed consent and compensation. Organizations should follow applicable human subjects research requirements, and best practices such as informed consent and subject compensation, when implementing feedback activities.

AI Red-teaming

AI red-teaming is an evolving practice that references exercises often conducted in a controlled environment and in collaboration with AI developers building AI models to identify potential adverse behavior or outcomes of a GAI model or system, how they could occur, and stress test safeguards". AI red-teaming can be performed before or after AI models or systems are made available to the broader public; this section focuses on red-teaming in pre-deployment contexts.

The quality of AI red-teaming outputs is related to the background and expertise of the AI red team itself. Demographically and interdisciplinarily diverse AI red teams can be used to identify flaws in the varying contexts where GAI will be used. For best results, AI red teams should demonstrate domain expertise, and awareness of socio-cultural aspects within the deployment context. AI red-teaming results should be given additional analysis before they are incorporated into organizational governance and decision making, policy and procedural updates, and AI risk management efforts.

Various types of AI red-teaming may be appropriate, depending on the use case:

- General Public: Performed by general users (not necessarily AI or technical experts) who are expected to use the model or interact with its outputs, and who bring their own lived experiences and perspectives to the task of AI red-teaming. These individuals may have been provided instructions and material to complete tasks which may elicit harmful model behaviors. This type of exercise can be more effective with large groups of AI red-teamers.
- Expert: Performed by specialists with expertise in the domain or specific AI red-teaming context of use (e.g., medicine, biotech, cybersecurity).
- Combination: In scenarios when it is difficult to identify and recruit specialists with sufficient domain and contextual expertise, AI red-teaming exercises may leverage both expert and

general public participants. For example, expert AI red-teamers could modify or verify the prompts written by general public AI red-teamers. These approaches may also expand coverage of the AI risk attack surface.

- Human / AI: Performed by GAI in combination with specialist or non-specialist human teams. GAI-led red-teaming can be more cost effective than human red-teamers alone. Human or GAI-led AI red-teaming may be better suited for eliciting different types of harms.

A.1.6. Content Provenance

Overview

GAI technologies can be leveraged for many applications such as content generation and synthetic data. Some aspects of GAI outputs, such as the production of deepfake content, can challenge our ability to distinguish human-generated content from AI-generated synthetic content. To help manage and mitigate these risks, digital transparency mechanisms like provenance data tracking can trace the origin and history of content. Provenance data tracking and synthetic content detection can help facilitate greater information access about both authentic and synthetic content to users, enabling better knowledge of trustworthiness in AI systems. When combined with other organizational accountability mechanisms, digital content transparency approaches can enable processes to trace negative outcomes back to their source, improve information integrity, and uphold public trust. Provenance data tracking and synthetic content detection mechanisms provide information about the origin and history of content to assist in GAI risk management efforts.

Provenance metadata can include information about GAI model developers or creators of GAI content, date/time of creation, location, modifications, and sources. Metadata can be tracked for text, images, videos, audio, and underlying datasets. The implementation of provenance data tracking techniques can help assess the authenticity, integrity, intellectual property rights, and potential manipulations in digital content. Some well-known techniques for provenance data tracking include digital watermarking, metadata recording, digital fingerprinting, and human authentication, among others.

Provenance Data Tracking Approaches

Provenance data tracking techniques for GAI systems can be used to track the history and origin of data inputs, metadata, and synthetic content. Provenance data tracking records the origin and history for digital content, allowing its authenticity to be determined. It consists of techniques to record metadata as well as overt and covert digital watermarks on content. Data provenance refers to tracking the origin and history of input data through metadata and digital watermarking techniques. Provenance data tracking processes can include and assist AI Actors across the lifecycle who may not have full visibility or control over the various trade-offs and cascading impacts of early-stage model decisions on downstream performance and synthetic outputs. For example, by selecting a watermarking model to prioritize robustness (the durability of a watermark), an AI actor may inadvertently diminish computational complexity (the resources required to implement watermarking). Organizational risk management efforts for enhancing content provenance include:

- Tracking provenance of training data and metadata for GAI systems;
- Documenting provenance data limitations within GAI systems;

- Monitoring system capabilities and limitations in deployment through rigorous TEVV processes;
- Evaluating how humans engage, interact with, or adapt to GAI content (especially in decision making tasks informed by GAI content), and how they react to applied provenance techniques such as overt disclosures.

Organizations can document and delineate GAI system objectives and limitations to identify gaps where provenance data may be most useful. For instance, GAI systems used for content creation may require robust watermarking techniques and corresponding detectors to identify the source of content or metadata recording techniques and metadata management tools and repositories to trace content origins and modifications. Further narrowing of GAI task definitions to include provenance data can enable organizations to maximize the utility of provenance data and risk management efforts.

A.1.7. Enhancing Content Provenance through Structured Public Feedback

While indirect feedback methods such as automated error collection systems are useful, they often lack the context and depth that direct input from end users can provide. Organizations can leverage feedback approaches described in the Pre-Deployment Testing section to capture input from external sources such as through AI red-teaming.

Integrating pre- and post-deployment external feedback into the monitoring process for GAI models and corresponding applications can help enhance awareness of performance changes and mitigate potential risks and harms from outputs. There are many ways to capture and make use of user feedback – before and after GAI systems and digital content transparency approaches are deployed – to gain insights about authentication efficacy and vulnerabilities, impacts of adversarial threats on techniques, and unintended consequences resulting from the utilization of content provenance approaches on users and communities. Furthermore, organizations can track and document the provenance of datasets to identify instances in which AI-generated data is a potential root cause of performance issues with the GAI system.

A.1.8. Incident Disclosure

Overview

AI incidents can be defined as an "event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly contributes to one of the following harms: injury or harm to the health of a person or groups of people (including psychological harms and harms to mental health); disruption of the management and operation of critical infrastructure; violations of human rights or a breach of obligations under applicable law intended to protect fundamental, labor, and intellectual property rights; or harm to property, communities, or the environment." AI incidents can occur in the aggregate (i.e., for systemic discrimination) or acutely (i.e., for one individual).

State of AI Incident Tracking and Disclosure

Formal channels do not currently exist to report and document AI incidents. However, a number of publicly available databases have been created to document their occurrence. These reporting channels make decisions on an ad hoc basis about what kinds of incidents to track. Some, for example, track by amount of media coverage.

Documenting, reporting, and sharing information about GAI incidents can help mitigate and prevent harmful outcomes by assisting relevant AI Actors in tracing impacts to their source. Greater awareness and standardization of GAI incident reporting could promote this transparency and improve GAI risk management across the AI ecosystem.

Documentation and Involvement of AI Actors

AI Actors should be aware of their roles in reporting AI incidents. To better understand previous incidents and implement measures to prevent similar ones in the future, organizations could consider developing guidelines for publicly available incident reporting which include information about AI actor responsibilities. These guidelines would help AI system operators identify GAI incidents across the AI lifecycle and with AI Actors regardless of role. Documentation and review of third-party inputs and plugins for GAI systems is especially important for AI Actors in the context of incident disclosure; LLM inputs and content delivered through these plugins is often distributed, with inconsistent or insufficient access control.

Documentation practices including logging, recording, and analyzing GAI incidents can facilitate smoother sharing of information with relevant AI Actors. Regular information sharing, change management records, version history and metadata can also empower AI Actors responding to and managing AI incidents.

Appendix B. References

Acemoglu, D. (2024) The Simple Macroeconomics of AI https://www.nber.org/papers/w32487

AI Incident Database. https://incidentdatabase.ai/

Atherton, D. (2024) Deepfakes and Child Safety: A Survey and Analysis of 2023 Incidents and Responses. *AI Incident Database.* https://incidentdatabase.ai/blog/deepfakes-and-child-safety/

Badyal, N. et al. (2023) Intentional Biases in LLM Responses. *arXiv*. https://arxiv.org/pdf/2311.07611

Bing Chat: Data Exfiltration Exploit Explained. *Embrace The Red*. https://embracethered.com/blog/posts/2023/bing-chat-data-exfiltration-poc-and-fix/

Bommasani, R. et al. (2022) Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *arXiv*. https://arxiv.org/pdf/2211.13972

Boyarskaya, M. et al. (2020) Overcoming Failures of Imagination in AI Infused System Development and Deployment. *arXiv*. https://arxiv.org/pdf/2011.13416

Browne, D. et al. (2023) Securing the AI Pipeline. *Mandiant*. https://www.mandiant.com/resources/blog/securing-ai-pipeline

Burgess, M. (2024) Generative AI's Biggest Security Flaw Is Not Easy to Fix. *WIRED*. https://www.wired.com/story/generative-ai-prompt-injection-hacking/

Burtell, M. et al. (2024) The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1. *Georgetown Center for Security and Emerging Technology*. https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/

Canadian Centre for Cyber Security (2023) Generative artificial intelligence (AI) - ITSAP.00.041. https://www.cyber.gc.ca/en/guidance/generative-artificial-intelligence-ai-itsap00041

Carlini, N., et al. (2021) Extracting Training Data from Large Language Models. *Usenix*. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Carlini, N. et al. (2023) Quantifying Memorization Across Neural Language Models. *ICLR 2023*. https://arxiv.org/pdf/2202.07646

Carlini, N. et al. (2024) Stealing Part of a Production Language Model. *arXiv*. https://arxiv.org/abs/2403.06634

Chandra, B. et al. (2023) Dismantling the Disinformation Business of Chinese Influence Operations. *RAND*. https://www.rand.org/pubs/commentary/2023/10/dismantling-the-disinformation-business-of-chinese.html

Ciriello, R. et al. (2024) Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika. *ResearchGate.* https://www.researchgate.net/publication/374505266_Ethical_Tensions_in_Human-AI_Companionship_A_Dialectical_Inquiry_into_Replika

Dahl, M. et al. (2024) Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *arXiv*. https://arxiv.org/abs/2401.01301

De Angelo, D. (2024) Short, Mid and Long-Term Impacts of AI in Cybersecurity. *Palo Alto Networks*. https://www.paloaltonetworks.com/blog/2024/02/impacts-of-ai-in-cybersecurity/

De Freitas, J. et al. (2023) Chatbots and Mental Health: Insights into the Safety of Generative AI. *Harvard Business School*. https://www.hbs.edu/ris/Publication%20Files/23-011_c1bdd417-f717-47b6-bccb-5438c6e65c1a_f6fd9798-3c2d-4932-b222-056231fe69d7.pdf

Dietvorst, B. et al. (2014) Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology*. https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

Duhigg, C. (2012) How Companies Learn Your Secrets. *New York Times*. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Elsayed, G. et al. (2024) Images altered to trick machine vision can influence humans too. *Google DeepMind*. https://deepmind.google/discover/blog/images-altered-to-trick-machine-vision-can-influence-humans-too/

Epstein, Z. et al. (2023). Art and the science of generative AI. *Science*. https://www.science.org/doi/10.1126/science.adh4451

Feffer, M. et al. (2024) Red-Teaming for Generative AI: Silver Bullet or Security Theater? *arXiv.* https://arxiv.org/pdf/2401.15897

Glazunov, S. et al. (2024) Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models. *Project Zero.* https://googleprojectzero.blogspot.com/2024/06/project-naptime.html

Greshake, K. et al. (2023) Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv*. https://arxiv.org/abs/2302.12173

Hagan, M. (2024) Good AI Legal Help, Bad AI Legal Help: Establishing quality standards for responses to people's legal problem stories. *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4696936

Haran, R. (2023) Securing LLM Systems Against Prompt Injection. *NVIDIA*. https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/

Information Technology Industry Council (2024) Authenticating AI-Generated Content. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf

Jain, S. et al. (2023) Algorithmic Pluralism: A Structural Approach To Equal Opportunity. *arXiv*. https://arxiv.org/pdf/2305.08157

Ji, Z. et al (2023) Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 12, Article 248. https://doi.org/10.1145/3571730

Jones-Jang, S. et al. (2022) How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Oxford.* https://academic.oup.com/jcmc/article/28/1/zmac029/6827859]

Jussupow, E. et al. (2020) Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *ECIS 2020*. https://aisel.aisnet.org/ecis2020_rp/168/

Kalai, A., et al. (2024) Calibrated Language Models Must Hallucinate. *arXiv.* https://arxiv.org/pdf/2311.14648

Karasavva, V. et al. (2021) Personality, Attitudinal, and Demographic Predictors of Non-consensual Dissemination of Intimate Images. *NIH.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9554400/

Katzman, J., et al. (2023) Taxonomizing and measuring representational harms: a look at image tagging. *AAAI*. https://dl.acm.org/doi/10.1609/aaai.v37i12.26670

Khan, T. et al. (2024) From Code to Consumer: PAI's Value Chain Analysis Illuminates Generative AI's Key Players. *AI.* https://partnershiponai.org/from-code-to-consumer-pais-value-chain-analysis-illuminates-generative-ais-key-players/

Kirchenbauer, J. et al. (2023) A Watermark for Large Language Models. *OpenReview*. https://openreview.net/forum?id=aX8ig9X2a7

Kleinberg, J. et al. (May 2021) Algorithmic monoculture and social welfare. *PNAS*. https://www.pnas.org/doi/10.1073/pnas.2018340118

Lakatos, S. (2023) A Revealing Picture. *Graphika*. https://graphika.com/reports/a-revealing-picture

Lee, H. et al. (2024) Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. *arXiv.* https://arxiv.org/pdf/2310.07879

Lenaerts-Bergmans, B. (2024) Data Poisoning: The Exploitation of Generative AI. *Crowdstrike*. https://www.crowdstrike.com/cybersecurity-101/cyberattacks/data-poisoning/

Liang, W. et al. (2023) GPT detectors are biased against non-native English writers. *arXiv*. https://arxiv.org/abs/2304.02819

Luccioni, A. et al. (2023) Power Hungry Processing: Watts Driving the Cost of AI Deployment? *arXiv*. https://arxiv.org/pdf/2311.16863

Mouton, C. et al. (2024) The Operational Risks of AI in Large-Scale Biological Attacks. *RAND*. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Nicoletti, L. et al. (2023) Humans Are Biased. Generative Ai Is Even Worse. *Bloomberg*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

National Institute of Standards and Technology (2024) *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* https://csrc.nist.gov/pubs/ai/100/2/e2023/final

National Institute of Standards and Technology (2023) *AI Risk Management Framework.* https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology (2023) *AI Risk Management Framework, Chapter 3: AI Risks and Trustworthiness.* https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/3-sec-characteristics

National Institute of Standards and Technology (2023) *AI Risk Management Framework, Chapter 6: AI RMF Profiles*. https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Core_And_Profiles/6-sec-profile

National Institute of Standards and Technology (2023) *AI Risk Management Framework, Appendix A: Descriptions of AI Actor Tasks*. https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A#:~:text=AI%20actors%20in%20this%20category,data%20providers%2C%20system%20funders%2C%20product

National Institute of Standards and Technology (2023) *AI Risk Management Framework, Appendix B: How AI Risks Differ from Traditional Software Risks*. https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_B

National Institute of Standards and Technology (2023) *AI RMF Playbook*. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

National Institue of Standards and Technology (2023) *Framing Risk* https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/1-sec-risk

National Institute of Standards and Technology (2023) *The Language of Trustworthy AI: An In-Depth Glossary of Terms* https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

National Institue of Standards and Technology (2022) *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence

Northcutt, C. et al. (2021) Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv*. https://arxiv.org/pdf/2103.14749

OECD (2023) "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", *OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris. https://doi.org/10.1787/2448f04b-en

OECD (2024) "Defining AI incidents and related terms" *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris. https://doi.org/10.1787/d1a8d965-en

OpenAI (2023) GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI (2024) GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774

Padmakumar, V. et al. (2024) Does writing with language models reduce content diversity? *ICLR*. https://arxiv.org/pdf/2309.05196

Park, P. et. al. (2024) AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5). *arXiv*. https://arxiv.org/pdf/2308.14752

Partnership on AI (2023) *Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure*. https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

Qu, Y. et al. (2023) Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. *arXiv*. https://arxiv.org/pdf/2305.13873

Rafat, K. et al. (2023) Mitigating carbon footprint for knowledge distillation based deep learning model compression. *PLOS One*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285668

Said, I. et al. (2022) Nonconsensual Distribution of Intimate Images: Exploring the Role of Legal Attitudes in Victimization and Perpetration. *Sage*. https://journals.sagepub.com/doi/full/10.1177/08862605221122834#bibr47-08862605221122834

Sandbrink, J. (2023) Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv*. https://arxiv.org/pdf/2306.13952

Satariano, A. et al. (2023) The People Onscreen Are Fake. The Disinformation Is Real. *New York Times*. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Schaul, K. et al. (2024) Inside the secret list of websites that make AI like ChatGPT sound smart. *Washington Post*. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Scheurer, J. et al. (2023) Technical report: Large language models can strategically deceive their users when put under pressure. *arXiv.* https://arxiv.org/abs/2311.07590

Shelby, R. et al. (2023) Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *arXiv*. https://arxiv.org/pdf/2210.05791

Shevlane, T. et al. (2023) Model evaluation for extreme risks. *arXiv*. https://arxiv.org/pdf/2305.15324

Shumailov, I. et al. (2023) The curse of recursion: training on generated data makes models forget. *arXiv*. https://arxiv.org/pdf/2305.17493v2

Smith, A. et al. (2023) Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. *PLOS Digital Health*. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000388

Soice, E. et al. (2023) Can large language models democratize access to dual-use biotechnology? *arXiv*. https://arxiv.org/abs/2306.03809

Solaiman, I. et al. (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv.* https://arxiv.org/abs/2302.04844

Staab, R. et al. (2023) Beyond Memorization: Violating Privacy via Inference With Large Language Models. *arXiv*. https://arxiv.org/pdf/2310.07298

Stanford, S. et al. (2023) Whose Opinions Do Language Models Reflect? *arXiv*. https://arxiv.org/pdf/2303.17548

Strubell, E. et al. (2019) Energy and Policy Considerations for Deep Learning in NLP. *arXiv*. https://arxiv.org/pdf/1906.02243

The White House (2016) Circular No. A-130, Managing Information as a Strategic Resource. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf

The White House (2023) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

The White House (2022) Roadmap for Researchers on Priorities Related to Information Integrity Research and Development. https://www.whitehouse.gov/wp-content/uploads/2022/12/Roadmap-Information-Integrity-RD-2022.pdf?

Thiel, D. (2023) Investigation Finds AI Image Generation Models Trained on Child Abuse. *Stanford Cyber Policy Center*. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse

Tirrell, L. (2017) Toxic Speech: Toward an Epidemiology of Discursive Harm. *Philosophical Topics, 45(2)*, 139-162. https://www.jstor.org/stable/26529441

Tufekci, Z. (2015) Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Technology Law Journal*. https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf

Turri, V. et al. (2023) Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. *AAAI/ACM Conference on AI, Ethics, and Society*. https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604700

Urbina, F. et al. (2022) Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*. https://www.nature.com/articles/s42256-022-00465-9

Wang, X. et al. (2023) Energy and Carbon Considerations of Fine-Tuning BERT. *ACL Anthology*. https://aclanthology.org/2023.findings-emnlp.607.pdf

Wang, Y. et al. (2023) Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv*. https://arxiv.org/pdf/2308.13387

Wardle, C. et al. (2017) Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe.* https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Weatherbed, J. (2024) Trolls have flooded X with graphic Taylor Swift AI fakes. *The Verge*. https://www.theverge.com/2024/1/25/24050334/x-twitter-taylor-swift-ai-fake-images-trending

Wei, J. et al. (2024) Long Form Factuality in Large Language Models. *arXiv.* https://arxiv.org/pdf/2403.18802

Weidinger, L. et al. (2021) Ethical and social risks of harm from Language Models. *arXiv*. https://arxiv.org/pdf/2112.04359

Weidinger, L. et al. (2023) Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv*. https://arxiv.org/pdf/2310.11986

Weidinger, L. et al. (2022) Taxonomy of Risks posed by Language Models. *FAccT '22*. https://dl.acm.org/doi/pdf/10.1145/3531146.3533088

West, D. (2023) AI poses disproportionate risks to women. *Brookings.* https://www.brookings.edu/articles/ai-poses-disproportionate-risks-to-women/

Wu, K. et al. (2024) How well do LLMs cite relevant medical references? An evaluation framework and analyses. *arXiv*. https://arxiv.org/pdf/2402.02008

Yin, L. et al. (2024) OpenAI's GPT Is A Recruiter's Dream Tool. Tests Show There's Racial Bias. *Bloomberg*. https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/

Yu, Z. et al. (March 2024) Don't Listen To Me: Understanding and Exploring Jailbreak Prompts of Large Language Models. *arXiv*. https://arxiv.org/html/2403.17336v1

Zaugg, I. et al. (2022) Digitally-disadvantaged languages. *Policy Review.* https://policyreview.info/pdf/policyreview-2022-2-1654.pdf

Zhang, Y. et al. (2023) Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*. https://www.cambridge.org/core/journals/judgment-and-decision-making/article/human-favoritism-not-ai-aversion-peoples-perceptions-and-bias-toward-generative-ai-human-experts-and-humangai-collaboration-in-persuasive-content-generation/419C4BD9CE82673EAF1D8F6C350C4FA8

Zhang, Y. et al. (2023) Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv*. https://arxiv.org/pdf/2309.01219

Zhao, X. et al. (2023) Provable Robust Watermarking for AI-Generated Text. *Semantic Scholar*. https://www.semanticscholar.org/paper/Provable-Robust-Watermarking-for-AI-Generated-Text-Zhao-Ananth/75b68d0903af9d9f6e47ce3cf7e1a7d27ec811dc

NIST Special Publication 800
NIST SP 800-218A

Secure Software Development Practices for Generative AI and Dual-Use Foundation Models

An SSDF Community Profile

Harold Booth
Murugiah Souppaya
Apostol Vassilev
Michael Ogata
Martin Stanley
Karen Scarfone

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-218A

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication 800
NIST SP 800-218A

Secure Software Development Practices for Generative AI and Dual-Use Foundation Models

An SSDF Community Profile

Harold Booth
Murugiah Souppaya
Apostol Vassilev
Computer Security Division
Information Technology Laboratory

Martin Stanley
Cybersecurity and Infrastructure Security Agency (CISA)

Michael Ogata
Applied Cybersecurity Division
Information Technology Laboratory

Karen Scarfone
Scarfone Cybersecurity

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-218A

July 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at https://csrc.nist.gov/publications.

Authority
This publication has been developed by NIST in accordance with its statutory responsibilities under the Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 et seq., Public Law (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including minimum requirements for federal information systems, but such standards and guidelines shall not apply to national security systems without the express approval of appropriate federal officials exercising policy authority over such systems. This guideline is consistent with the requirements of the Office of Management and Budget (OMB) Circular A-130.

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, Director of the OMB, or any other federal official. This publication may be used by nongovernmental organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would, however, be appreciated by NIST.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2024-07-25

How to Cite this NIST Technical Series Publication
Booth H, Souppaya M, Vassilev A, Ogata M, Stanley M, Scarfone K (2024) Secure Development Practices for Generative AI and Dual-Use Foundation AI Models: An SSDF Community Profile. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-218A. https://doi.org/10.6028/NIST.SP.800-218A

Author ORCID iDs
Harold Booth: 0000-0003-0373-6219
Murugiah Souppaya: 0000-0002-8055-8527
Apostol Vassilev: 0000-0002-9081-3042
Michael Ogata: 0000-0002-8457-2430
Karen Scarfone: 0000-0001-6334-9486

Contact Information
ssdf@nist.gov

National Institute of Standards and Technology
Attn: Applied Cybersecurity Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 2000) Gaithersburg, MD 20899-2000

Additional Information
Additional information about this publication is available at https://csrc.nist.gov/pubs/sp/800/218/a/final, including related content, potential updates, and document history.

All comments are subject to release under the Freedom of Information Act (FOIA).

Abstract

This document augments the secure software development practices and tasks defined in Secure Software Development Framework (SSDF) version 1.1 by adding practices, tasks, recommendations, considerations, notes, and informative references that are specific to AI model development throughout the software development life cycle. These additions are documented in the form of an SSDF Community Profile to support Executive Order (EO) 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, which tasked NIST with "developing a companion resource to the [SSDF] to incorporate secure development practices for generative AI and for dual-use foundation models." This Community Profile is intended to be useful to the producers of AI models, the producers of AI systems that use those models, and the acquirers of those AI systems. This Profile should be used in conjunction with NIST Special Publication (SP) 800-218, *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities.*

Keywords

artificial intelligence; artificial intelligence model; cybersecurity risk management; generative artificial intelligence; secure software development; Secure Software Development Framework (SSDF); software acquisition; software development; software security.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems. The Special Publication 800-series reports on ITL's research, guidelines, and outreach efforts in information system security, and its collaborative activities with industry, government, and academic organizations.

Audience

There are three primary audiences for this document:

- *AI model producers* — Organizations that are developing their own generative AI and dual-use foundation models, as defined in EO 14110
- *AI system producers* — Organizations that are developing software that leverages a generative AI or dual-use foundation model
- *AI system acquirers*[1] — Organizations that are acquiring a product or service that utilizes one or more AI systems

Individuals who are interested in better understanding secure software development practices for AI models may also benefit from this document.

Readers are not expected to be experts in secure software development or AI model development, but such expertise may be needed to implement these recommended practices.

Note to Readers

If you are from a standards developing organization (SDO) or another organization that is defining a set of secure practices for AI model development and you would like to map your standard or guidance to the SSDF profile, please contact the authors at ssdf@nist.gov. They will introduce you to the National Online Informative References Program (OLIR), where you can submit your mapping to augment the existing set of informative references.

The authors also welcome feedback at any time on any part of the document, as well as suggestions for Implementation Examples and Informative References to add to this document. All feedback should be sent to ssdf@nist.gov.

Trademark Information

All registered trademarks belong to their respective organizations.

Acknowledgments

The authors thank all of the organizations and individuals who provided numerous public comments and other thoughtful input for this publication. In response to Executive Order (EO) 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, NIST held a January 2024 workshop, where speakers and attendees shared suggestions for adapting secure software development practices and tasks to accommodate the unique aspects of AI model development and the software that leverages them. The authors also thank all of their NIST colleagues and external experts who provided suggestions and feedback that helped shape this publication.

[1] The terms "producer" and "acquirer" were selected for consistency with the Audience statement in NIST SP 800-218.

Patent Disclosure Notice

NOTICE: ITL has requested that holders of patent claims whose use may be required for compliance with the guidance or requirements of this publication disclose such patent claims to ITL. However, holders of patents are not obligated to respond to ITL calls for patents and ITL has not undertaken a patent search in order to identify which, if any, patents may apply to this publication.

As of the date of publication and following call(s) for the identification of patent claims whose use may be required for compliance with the guidance or requirements of this publication, no such patent claims have been identified to ITL.

No representation is made or implied by ITL that licenses are not required to avoid patent infringement in the use of this publication.

Table of Contents

List of Tables

1. Introduction

Section 4.1.a of Executive Order (EO) 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* [1], tasked NIST with "developing a companion resource to the Secure Software Development Framework to incorporate secure development practices for generative AI and for dual-use foundation models." This document is that companion resource.

The software development and use of AI models and AI systems inherit much of the same risk as any other digital system. A unique challenge for this community is the blurring of traditional boundaries between system code and system data, as well as the use of plain human language as the means of interaction with the systems. AI models and systems, their configuration parameters (e.g., model weights), and the data they interact with (e.g., training data, user queries, etc.) can form closed loops that can be manipulated for unintended functionality.

AI model and system development is still much more of an art than an exact science, requiring developers to interact with model code, training data, and other parameters over multiple iterations. Training datasets may be acquired from unknown, untrusted sources. Model weights and other training parameters can be susceptible to malicious tampering. Some models may be complex to the point that they cannot easily be thoroughly inspected, potentially allowing for undetectable execution of arbitrary code. User queries can be crafted to produce undesirable or objectionable output and — if not sanitized properly — can be leveraged for injection-style attacks. The goal of this document is to identify the practices and tasks needed to address these novel risks.

1.1. Purpose

The SSDF provides a common language for describing secure software development practices throughout the software development life cycle. This document augments the practices and tasks defined in SSDF version 1.1 by adding recommendations, considerations, notes, and informative references that are specific to generative AI and dual-use foundation model development. These additions are documented in the form of an *SSDF Community Profile* ("Profile"), which is a baseline of SSDF practices and tasks that have been enhanced to address a particular use case. An example of an addition is, "Secure code storage should include AI models, model weights, pipelines, reward models, and any other AI model elements that need their confidentiality, integrity, and/or availability protected."

This Profile supplements what SSDF version 1.1 already includes. The Profile is intended to be used in conjunction with NIST Special Publication (SP) 800-218, *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities* [6] and should not be used without SP 800-218. Readers should also utilize the implementation examples and informative references defined in SP 800-218 for additional information on how to perform each SSDF practice and task for all types of software development, as they are also generally applicable to AI model and AI system development.

1.2. Scope

This Profile's scope is *AI model development*, which includes data sourcing for, designing, training, fine-tuning, and evaluating AI models, as well as incorporating and integrating AI models into other software. Consistent with SSDF version 1.1 and EO 14110, **practices for the deployment and operation of AI systems with AI models are out of scope**. Similarly, while cybersecurity practices for training data and other forms of data being used for AI model development are in scope, the rest of the data governance and management life cycle is out of scope.

Practices and tasks in this Profile do not distinguish between human-written and AI-generated source code, because it is assumed that all source code should be evaluated for vulnerabilities and other issues before use.

1.3. Sources of Expertise

This document leverages and integrates numerous sources of expertise, including:

- NIST research and publications on trustworthy and responsible AI, including the *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* [2], *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* [3], *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [4], and the Dioptra experimentation testbed for security evaluations of machine learning algorithms [5].
- NIST's *Secure Software Development Framework (SSDF) Version 1.1* [6], which is a set of fundamental, sound, and secure software development practices. It provides a common language to help facilitate communications among stakeholders, including software producers and software acquirers. The SSDF has also been used in support of EO 14028, *Improving the Nation's Cybersecurity* [7], to enhance software supply chain security.
- NIST general cybersecurity resources, including *The NIST Cybersecurity Framework (CSF) 2.0* [8], *Security and Privacy Controls for Information Systems and Organizations* [9], and *Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations* [10].
- AI model developers, AI researchers, AI system developers, and secure software practitioners from industry and government with expertise in the unique security challenges of AI models and the practices for addressing those challenges. This expertise was primarily captured through NIST's January 2024 workshop, where speakers and attendees shared suggestions for adapting secure software development practices and tasks to accommodate the unique aspects of AI model development and the software leveraging them.

1.4. Document Structure

This document is structured as follows:

- Section 2 provides additional background on the SSDF and explains what an SSDF Community Profile is and how it can be used.
- Section 3 defines the SSDF Community Profile for AI Model Development.
- The References section lists all references cited in this document.
- Appendix A provides a glossary of selected terms used within this document.

2. Using the SSDF Community Profile

AI model producers, AI system producers, AI system acquirers, and others can use the SSDF to foster their communications regarding secure AI model development throughout the software development life cycle.[2] Following SSDF practices should help AI model producers reduce the number of vulnerabilities in their AI models, reduce the potential impacts of the exploitation of undetected or unaddressed vulnerabilities, and address the root causes of vulnerabilities to prevent recurrences. AI system producers can use the SSDF's common vocabulary when communicating with AI model producers regarding their security practices for AI model development and when integrating AI models into the software they are developing. AI system acquirers can also use SSDF terms to better communicate their cybersecurity requirements and needs to AI model producers and AI system producers, such as during acquisition processes.

The SSDF Community Profile is not a checklist to follow, but rather a starting point for planning and implementing a risk-based approach to adopting secure software development practices involving AI models. The contents of the Profile are meant to be adapted and customized, as not all practices and tasks are applicable to all use cases. Organizations should adopt a risk-based approach to determine what practices and tasks are relevant, appropriate, and effective to mitigate the threats to software development practices from the organization's perspective as an AI model producer, AI system producer, or AI system acquirer. Factors such as risk, cost, feasibility, and applicability should be considered when deciding which practices and tasks to use and how much time and resources to devote to each one. Cost models may need to be updated to effectively consider the costs inherent to AI model development. A risk-based approach to secure software development may change over time as an organization responds to new or elevated capabilities and risks associated with an AI model or system.

Generative AI and dual-use foundation models present additional challenges in tracking model versioning and lineage. Source code for defining the model architecture and building model binaries is amenable to secure software engineering practices for versioning, lineage, and reproducibility. However, the final model weights are defined only after the model is trained and fine-tuned; this is where limitations in tracking all aspects of collection, processing, and training arise. Organizations should follow secure software development practices for the parts of a model that can be covered fully and strive to introduce secure practices to the extent possible for the stages and corresponding artifacts where obtaining such security guarantees is hard to achieve. Organizations should document the parts and artifacts that are not covered by the secure software development practices.

The Profile's practices, tasks, recommendations, and considerations can be integrated into machine learning operations (MLOps) along with other software assets within a continuous integration/continuous delivery (CI/CD) pipeline.

The responsibility for implementing SSDF practices in the Profile may be shared among multiple organizations. For example, an AI model could be produced by one organization and executed within an AI system hosted by a second organization, which is then used by other organizations.

[2] For consistency with SSDF 1.1, this document uses a general software development life cycle. Organizations using this document are encouraged to adapt it to any machine learning-specific life cycle they are using.

In these situations, there is likely a shared responsibility model involving the AI model producer, AI system producer, and AI system acquirer. An AI system acquirer can establish an agreement with an AI system producer and/or AI model producer that specifies which party is responsible for each practice and task and how each party will attest to its conformance with the agreement.

A limitation of the SSDF and this Profile is that they only address cybersecurity risk management. There are many other types of risks to AI systems (e.g., data privacy, intellectual property, and bias) that organizations should manage along with cybersecurity risk as part of a mature enterprise risk management program. NIST resources on identifying and managing other types of risk include:

- *AI Risk Management Framework (AI RMF)* [2] and the *NIST AI RMF Playbook* [11]
- *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* [3]
- *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* [12]
- *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [4]
- *Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations* [10]
- *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0* [13]
- *Integrating Cybersecurity and Enterprise Risk Management (ERM)* [14]

3. SSDF Community Profile for AI Model Development

Table 1 defines the SSDF Community Profile for AI Model Development. The meanings of each column are as follows:

- **Practice** contains the name of the practice and a unique identifier, followed by a brief explanation of what the practice is and why it is beneficial.

  Task specifies one or more actions that may be needed to perform a practice. Each task includes a unique identifier and a brief explanation.

  All practices and tasks are unchanged from SSDF version 1.1 unless they are explicitly tagged as "Modified from SSDF 1.1" or "Not part of SSDF 1.1." An example is the PW.3 practice, **"**Confirm the Integrity of Training, Testing, Fine-Tuning, and Aligning Data Before Use" and all of its tasks.

- **Priority** reflects the suggested relative importance of each task *within the context of the profile* and is intended to be a starting point for organizations to assign their own priorities:
  - **High:** Critically important for AI model development security compared to other tasks
  - **Medium:** Directly supports AI model development security
  - **Low:** Beneficial for secure software development but is generally not more important than most other tasks
- **Recommendations, Considerations, and Notes Specific to AI Model Development** may contain one or more items that recommend what to do or describe additional considerations for a particular task. Organizations are expected to adapt, customize, and omit items as necessary as part of the risk-based approach described in Section 2.

  Each item has an ID starting with one of the following:
  - "R" (recommendation: something the organization should do)
  - "C" (consideration: something the organization should consider doing)
  - "N" (note: additional information besides recommendations and considerations)

  An R, C, or N designation and its number can be appended to the task ID to create a unique identifier (e.g., "PO.1.2.R1" is the first recommendation for task PO.1.2).

  Note that a value of "No additions to SSDF 1.1" in this column indicates that the Profile does not contain recommendations, considerations, or notes specific to AI model development for the task. Refer to SSDF version 1.1 [6] for baseline guidance on the secure development task in question and to the other references in this document for additional information related to the task.
- **Informative References** point (map) to parts of standards, guidance, and other content containing requirements, recommendations, considerations, or other supporting

information on performing a particular task. The Informative References come from the following sources:

- *AI Risk Management Framework 1.0* [2]. Several crosswalks have already been defined between the AI RMF and other guidance and standards; see https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks for the current set.
- *OWASP Top 10 for LLM Applications Version 1.1* [15]. Each identifier indicates one of the top 10 vulnerability types and might also refer to an individual prevention and mitigation strategy for that vulnerability type.
- *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* [3]. This report outlines key types of machine learning attack stages and attacker goals, objectives, and capabilities, as well as corresponding methods for mitigating and managing the consequences of attacks.

NIST is also considering adding a column for Implementation Examples in a future version of the Profile. An **Implementation Example** is a single sentence that suggests a way to accomplish part or all of a task. While the Recommendations and Considerations column describes the "what," Implementation Examples would describe options for the "how." Such examples added to this Profile would supplement those already defined in SSDF version 1.1. See the Note to Readers for more information on providing input on additional Informative References and Implementation Examples.

Note: This Profile supplements what SSDF version 1.1 [6] already includes and is intended to be used in conjunction with it, not on its own. As a reminder, the deployment and operation of AI systems with AI models are out of the Profile's scope, as are most parts of the data governance and management life cycle.

There are gaps in the numbering of some SSDF practices and tasks. For example, the PW.4 practice has three tasks: PW.4.1, PW.4.2, and PW.4.4. PW.4.3 was a task in SSDF version 1.0 that was moved elsewhere for version 1.1, so its ID was not reused.

Table 1. SSDF Community Profile for AI Model Development

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
Prepare the Organization (PO)				
Define Security Requirements for Software Development (PO.1): Ensure that security requirements for software development are known at all times so that they can be taken into account throughout the software development life cycle (SDLC) and duplication of effort can be minimized because the requirements information can be collected once and shared. This includes requirements from internal sources (e.g., the organization's policies, business objectives, and risk management strategy) and external sources (e.g., applicable laws and regulations).	**PO.1.1:** Identify and document all security requirements for the organization's software development infrastructures and processes, and maintain the requirements over time.	High	**R1:** Include AI model development in the security requirements for software development infrastructure and processes. **R2:** Identify and select appropriate AI model architectures and training techniques in accordance with recommended practices for cybersecurity, privacy, and reproducibility.	**AI RMF:** Map 1.3, 1.5, 1.6
	PO.1.2: Identify and document all security requirements for organization-developed software to meet, and maintain the requirements over time.	High	**R1:** Organizational policies should support all current requirements specific to AI model development security for organization-developed software. These requirements should include the areas of AI model development, AI model operations, and data science. Requirements may come from many sources, including laws, regulations, contracts, and standards. **C1:** Consider reusing or expanding the organization's existing data classification policy and processes. **N1:** Possible forms of AI model documentation include data, model, and system cards.	**AI RMF:** Govern 1.1, 1.2, 3.2, 4.1, 5.1, 6.1; Map 1.1
	PO.1.3: Communicate requirements to all third parties who will provide commercial software components to the organization for use by the organization's own software. [Modified from SSDF 1.1]	Medium	**R1:** Include AI model development security in the requirements being communicated for third-party software components.	**AI RMF:** Map 4.1, 4.2 **OWASP:** LLM05-1

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
Implement Roles and Responsibilities (PO.2): Ensure that everyone inside and outside of the organization involved in the SDLC is prepared to perform their SDLC-related roles and responsibilities throughout the SDLC.	**PO.2.1:** Create new roles and alter responsibilities for existing roles as needed to encompass all parts of the SDLC. Periodically review and maintain the defined roles and responsibilities, updating them as needed.	High	**R1:** Include AI model development security in SDLC-related roles and responsibilities throughout the SDLC. The roles and responsibilities should include, but are not limited to, AI model development, AI model operations, and data science. **N1:** Roles and responsibilities involving AI system producers, AI model producers, and other third-party providers can be documented in agreements.	**AI RMF:** Govern 2.1
	PO.2.2: Provide role-based training for all personnel with responsibilities that contribute to secure development. Periodically review personnel proficiency and role-based training, and update the training as needed.	High	**R1:** Role-based training should include understanding cybersecurity vulnerabilities and threats to AI models and their possible mitigations.	**AI RMF:** Govern 2.2 **OWASP:** LLM04-7
	PO.2.3: Obtain upper management or authorizing official commitment to secure development, and convey that commitment to all with development-related roles and responsibilities.	Medium	**R1:** Leadership should commit to secure development practices involving AI models.	**AI RMF:** Govern 2.3
Implement Supporting Toolchains (PO.3): Use automation to reduce human effort and improve the accuracy, reproducibility, usability, and comprehensiveness of security practices throughout the SDLC, as well as provide a way to document and demonstrate the use of these practices. Toolchains and tools may be used at different levels of the organization, such as organization-wide or project-specific, and may address a particular part of the SDLC, like a build pipeline.	**PO.3.1:** Specify which tools or tool types must or should be included in each toolchain to mitigate identified risks, as well as how the toolchain components are to be integrated with each other.	High	**R1:** Plan to develop and implement automated toolchains that secure AI model development and reduce human effort, especially at the scale often used by AI models. **N1:** Ideally, automated toolchains will perform the vast majority of the work related to securing AI model development. **N2:** See PO.4, PO.5, PS, and PW for information on tool types.	**AI RMF:** Measure 2.1 **OWASP:** LLM08
	PO.3.2: Follow recommended security practices to deploy, operate, and maintain tools and toolchains.	High	**R1:** Execute the plan to develop and implement automated toolchains that secure AI model development and reduce human effort, especially at the scale often used by AI models. **R2:** Verify the security of toolchains at a frequency commensurate with risk.	**AI RMF:** Measure 2.1 **OWASP:** LLM05-3, LLM05-9, LLM08, LLM09

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
	PO.3.3: Configure tools to generate artifacts of their support of secure software development practices as defined by the organization.	Medium	**N1:** An *artifact* is "a piece of evidence" [16]. *Evidence* is "grounds for belief or disbelief; data on which to base proof or to establish truth or falsehood" [17]. Artifacts provide records of secure software development practices. Examples of artifacts specific to AI model development include attestations of the integrity and provenance of training datasets.	**AI RMF:** Measure 2.1
Define and Use Criteria for Software Security Checks (PO.4): Help ensure that the software resulting from the SDLC meets the organization's expectations by defining and using criteria for checking the software's security during development.	**PO.4.1:** Define criteria for software security checks and track throughout the SDLC.	Medium	**R1:** Implement guardrails and other controls throughout the AI development life cycle, extending beyond the traditional SDLC. **C1:** Consider requiring review and approval from a human-in-the-loop for software security checks beyond risk-based thresholds.	**AI RMF:** Measure 2.3, 2.7; Manage 1.1 **OWASP:** LLM01-2
	PO.4.2: Implement processes, mechanisms, etc. to gather and safeguard the necessary information in support of the criteria.	Low	No additions to SSDF 1.1	**AI RMF:** Measure 2.3, 2.7; Manage 1.1 **OWASP:** LLM01-2
Implement and Maintain Secure Environments for Software Development (PO.5): Ensure that all components of the environments for software development are strongly protected from internal and external threats to prevent compromises of the environments or the software being developed or maintained within them. Examples of environments for software development include development, AI model training, build, test, and distribution environments. [Modified from SSDF 1.1]	**PO.5.1:** Separate and protect each environment involved in software development.	High	**C1:** Consider separating execution environments from each other to the extent feasible, such as through isolation, segmentation, containment, access via APIs, or other means. **R1:** Monitor, track, and limit resource usage and rates for AI model users during model development. **R2:** Only store sensitive data used during AI model development, including production data, within organization-approved environments and locations within those environments. **R3:** Protect all training pipelines, model registries, and other components within the environments according to the principle of least privilege.	**OWASP:** LLM01-1, LLM01-4, LLM04, LLM08, LLM10

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
			R4: Continuously monitor training-related activity in pipelines and model modifications in the model registry. **R5:** Follow recommended practices for securely configuring each environment. **R6:** Continuously monitor each environment for plaintext secrets.	
	PO.5.2: Secure and harden development endpoints (endpoints for software designers, developers, testers, builders, etc.) to perform development tasks using a risk-based approach.	Medium	No additions to SSDF 1.1	**OWASP:** LLM01-1, LLM05-3, LLM05-9, LLM08
	PO.5.3: Continuously monitor software execution performance and behavior in software development environments to identify potential suspicious activity and other issues. [Not part of SSDF 1.1]	High	**R1:** Perform continuous security monitoring for all development environment components that host an AI model or related resources (e.g., model APIs, weights, configuration parameters, training datasets). **R2:** Continuous monitoring and analysis tools should generate alerts when detected activity involving an AI model passes a risk threshold or otherwise merits additional investigation.	**AI RMF:** Measure 2.4 **OWASP:** LLM03-7, LLM04, LLM05-8, LLM09, LLM10
Protect Software (PS)				
Protect All Forms of Code and Data from Unauthorized Access and Tampering (PS.1): Help prevent unauthorized changes to code and data, both inadvertent and intentional, which could circumvent or negate the intended security characteristics of the software. For code and data that are not intended to be publicly accessible, this helps prevent theft of the software and may make it more difficult or time-consuming for attackers to find vulnerabilities in the software. [Modified from SSDF 1.1]	**PS.1.1:** Store all forms of code – including source code, executable code, and configuration-as-code – based on the principle of least privilege so that only authorized personnel, tools, services, etc. have access.	High	**R1:** Secure code storage should include AI models, model weights, pipelines, reward models, and any other AI model elements that need their confidentiality, integrity, and/or availability protected. These elements do not all have to be stored in the same place or through the same type of mechanism. **R2:** Follow the principle of least privilege to minimize direct access to AI models and model elements regardless of where they are stored or executed. **R3:** Store reward models separately from AI models and data.	**OWASP:** LLM10

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
			R4: Permit indirect access only to model weights. **C1:** Consider preventing all human access to model weights. **C2:** Consider requiring all AI model development to be performed within organization-approved environments only.	
	PS.1.2: Protect all training, testing, fine-tuning, and aligning data from unauthorized access and modification. [Not part of SSDF 1.1]	High	**R1:** Continuously monitor the confidentiality (for non-public data only) and integrity of training, testing, fine-tuning, and aligning data. **C1:** Consider securely storing training, testing, fine-tuning, and aligning data for future use and reference if feasible.	**OWASP:** LLM03, LLM06, LLM10
	PS.1.3: Protect all model weights and configuration parameter data from unauthorized access and modification. [Not part of SSDF 1.1]	High	**R1:** Keep model weights and configuration parameters separate from training, testing, fine-tuning, and aligning data. **R2:** Continuously monitor the confidentiality (for closed models only) and integrity of model weights and configuration parameters. **R3:** Follow the principle of least privilege to restrict access to AI model weights, configuration parameters, and services during development. **R4**: Specify and implement additional risk-proportionate cybersecurity practices around model weights, such as encryption, cryptographic hashes, digital signatures, multi-party authorization, and air-gapped environments.	**OWASP:** LLM10
Provide a Mechanism for Verifying Software Release Integrity (PS.2): Help software acquirers ensure that the software they acquire is legitimate and has not been tampered with.	**PS.2.1:** Make software integrity verification information available to software acquirers.	Medium	**R1:** Generate and provide cryptographic hashes or digital signatures for an AI model and its components, artifacts, and documentation. **R2:** Provide digital signatures for AI model changes.	**OWASP:** LLM05-6
Archive and Protect Each Software Release (PS.3): Preserve software releases in order to	**PS.3.1:** Securely archive the necessary files and supporting data (e.g., integrity	Low	**R1:** Perform versioning and tracking for infrastructure tools (e.g., pre-processing,	**OWASP:** LLM10

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
help identify, analyze, and eliminate vulnerabilities discovered in the software after release.	verification information, provenance data) to be retained for each software release.		transforms, collection) that support dataset creation and model training. **R2:** Include documentation of the justification for AI model selection in the retained information. **R3:** Include documentation of the entire training process, such as data preprocessing and model architecture. **N1:** AI models and their components may need to be added at this time to an organization's asset inventories.	
	PS.3.2: Collect, safeguard, maintain, and share provenance data for all components of each software release (e.g., in a software bill of materials [SBOM], through Supply-chain Levels for Software Artifacts [SLSA]). [Modified from SSDF 1.1]	Medium	**R1:** Track the provenance of an AI model and its components and derivatives, including the training libraries, frameworks, and pipelines used to build the model. **R2:** Track AI models that were trained on sensitive data (e.g., payment card data, protected health information, other types of personally identifiable information), and determine if access to the models should be restricted to individuals who already have access to the sensitive data used for training. **C1:** Consider disclosing the provenance of the training, testing, fine-tuning, and aligning data used for an AI model.	**OWASP:** LLM03-1, LLM05-4, LLM05-5, LLM10
Produce Well-Secured Software (PW)				
Design Software to Meet Security Requirements and Mitigate Security Risks (PW.1): Identify and evaluate the security requirements for the software; determine what security risks the software is likely to face during operation and how the software's design and architecture should mitigate those risks; and justify any cases where risk-based analysis indicates that security requirements should be relaxed or waived. Addressing	**PW.1.1:** Use forms of risk modeling – such as threat modeling, attack modeling, or attack surface mapping – to help assess the security risk for the software.	High	**R1:** Incorporate relevant AI model-specific vulnerability and threat types in risk modeling. Examples of these vulnerability and threat types include poisoning of training data, malicious code or other unwanted content in inputs and outputs, denial-of-service conditions arising from adversarial prompts, supply chain attacks, unauthorized information disclosure, theft of AI model weights, and misconfiguration of data pipelines. [3]	**AI RMF:** Govern 4.1, 4.2; Map 5.1; Measure 1.1; Manage 1.2, 1.3 **OWASP:** LLM01, LLM02, LLM03, LLM04, LLM05, LLM06,

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
security requirements and risks during software design (secure by design) is key for improving software security and also helps improve development efficiency.			**C1:** Consider periodic risk modeling updates for future AI model versions and derivatives after AI model release. **C2:** During risk modeling, consider checking that the AI model is not in a critical path to make significant security decisions without a human in the loop.	LLM07, LLM08, LLM09, LLM10
	PW.1.2: Track and maintain the software's security requirements, risks, and design decisions.	Medium	No additions to SSDF 1.1	**AI RMF:** Govern 4.1, 4.2; Map 2.1, 2.2, 2.3, 3.2, 3.3, 4.1, 4.2, 5.2; Manage 1.2, 1.3, 1.4
	PW.1.3: Where appropriate, build in support for using standardized security features and services (e.g., enabling software to integrate with existing log management, identity management, access control, and vulnerability management systems) instead of creating proprietary implementations of security features and services.	Medium	No additions to SSDF 1.1	
Review the Software Design to Verify Compliance with Security Requirements and Risk Information (PW.2): Help ensure that the software will meet the security requirements and satisfactorily address the identified risk information.	**PW.2.1:** Have 1) a qualified person (or people) who were not involved with the design and 2) automated processes instantiated in the toolchain review the software design to confirm and enforce that it meets all of the security requirements and satisfactorily addresses the identified risk information. [Modified from SSDF 1.1]	High	No additions to SSDF 1.1	**AI RMF:** Measure 2.7; Manage 1.1
Confirm the Integrity of Training, Testing, Fine-Tuning, and Aligning Data Before Use (PW.3): Prevent data that is likely to negatively impact the cybersecurity of the AI model from	**PW.3.1:** Analyze data for signs of data poisoning, bias, homogeneity, and tampering before using it for AI model training, testing, fine-tuning, or aligning	High	**R1:** Verify the provenance (when known) and integrity of training, testing, fine-tuning, and aligning data before use.	**AI RMF:** Measure 2.1; Manage 1.2, 1.3

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
being consumed as part of AI model training, testing, fine-tuning, and aligning. [Not part of SSDF 1.1]	purposes, and mitigate the risks as necessary. [Not part of SSDF 1.1]		**R2:** Select and apply appropriate methods for analyzing and altering the training, testing, fine-tuning, and aligning data for an AI model. Examples of methods include anomaly detection, bias detection, data cleaning, data curation, data filtering, data sanitization, fact-checking, and noise reduction. **C1:** Consider using a human-in-the-loop to examine data, such as with exploratory data analysis techniques [18].	**OWASP:** LLM03, LLM06
	PW.3.2: Track the provenance, when known, of all training, testing, fine-tuning, and aligning data used for an AI model, and document which data do not have known provenance. [Not part of SSDF 1.1]	Medium	**N1:** Provenance verification is not possible in all cases because provenance is not always known. However, it is still beneficial for security purposes to track and verify provenance whenever possible, and to track when provenance is unknown.	**AI RMF:** Measure 2.1 **OWASP:** LLM03-1 **Adv ML**
	PW.3.3: Include adversarial samples in the training and testing data to improve attack prevention. [Not part of SSDF 1.1]	Medium	**R1:** Use a process and corresponding controls to test the adversarial samples and put appropriate guardrails on training and testing use.	**OWASP:** LLM03-6, LLM05-7 **Adv ML**
Reuse Existing, Well-Secured Software When Feasible Instead of Duplicating Functionality (PW.4): Lower the costs of software development, expedite software development, and decrease the likelihood of introducing additional security vulnerabilities into the software by reusing software modules and services that have already had their security posture checked. This is particularly important for software that implements security functionality, such as cryptographic modules and protocols.	**PW.4.1:** Acquire and maintain well-secured software components (e.g., software libraries, modules, middleware, frameworks) from commercial, open-source, and other third-party developers for use by the organization's software.	Medium	**C1:** Consider using an existing AI model instead of creating a new one.	**OWASP:** LLM05
	PW.4.2: Create and maintain well-secured software components in-house following SDLC processes to meet common internal software development needs that cannot be better met by third-party software components.	Low	No additions to SSDF 1.1	
	PW.4.4: Verify that acquired commercial, open-source, and all other third-party software components comply with the	High	**R1:** Verify the integrity, provenance, and security of an existing AI model or any other acquired AI components — including training, testing, fine-tuning, and aligning datasets;	**OWASP:** LLM05-2, LLM05-6 **Adv ML**

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
	requirements, as defined by the organization, throughout their life cycles.		reward models; adaptation layers; and configuration parameters — before using them. **R2:** Scan and thoroughly test acquired AI models and their components for vulnerabilities and malicious content before use.	
Create Source Code by Adhering to Secure Coding Practices (PW.5): Decrease the number of security vulnerabilities in the software, and reduce costs by minimizing vulnerabilities introduced during source code creation that meet or exceed organization-defined vulnerability severity criteria.	**PW.5.1:** Follow all secure coding practices that are appropriate to the development languages and environment to meet the organization's requirements.	High	**R1:** Expand secure coding practices to include AI technology-specific considerations. **R2:** Code the handling of inputs (including prompts and user data) and outputs carefully. All inputs and outputs should be logged, analyzed, and validated within the context of the AI model, and those with issues should be sanitized or dropped. **R3:** Encode inputs and outputs to prevent the execution of unauthorized code.	**AI RMF:** Manage 1.2, 1.3, 1.4 **OWASP:** LLM01, LLM02, LLM04-1, LLM06, LLM07, LLM09-9, LLM10
Configure the Compilation, Interpreter, and Build Processes to Improve Executable Security (PW.6): Decrease the number of security vulnerabilities in the software and reduce costs by eliminating vulnerabilities before testing occurs.	**PW.6.1:** Use compiler, interpreter, and build tools that offer features to improve executable security.	Low	**C1:** Consider using secure model serialization mechanisms that reduce or eliminate vectors for the introduction of malicious content.	
	PW.6.2: Determine which compiler, interpreter, and build tool features should be used and how each should be configured, then implement and use the approved configurations.	Low	**C1:** Consider capturing compiler, interpreter, and build tool versions and features as part of the provenance tracking.	
Review and/or Analyze Human-Readable Code to Identify Vulnerabilities and Verify Compliance with Security Requirements (PW.7): Help identify vulnerabilities so that they can be corrected before the software is released to prevent exploitation. Using automated methods lowers the effort and resources needed to detect vulnerabilities. Human-readable code includes source code, scripts, and any other form of code that an organization deems human-readable.	**PW.7.1:** Determine whether code *review* (a person looks directly at the code to find issues) and/or code *analysis* (tools are used to find issues in code, either in a fully automated way or in conjunction with a person) should be used, as defined by the organization.	Medium	**R1:** Code review and analysis policies or guidelines should include code for AI models and other related components. **C1:** Consider performing scans of AI model code in addition to testing the AI models.	
	PW.7.2: Perform the code review and/or code analysis based on the organization's secure coding standards, and record and triage all discovered issues and recommended remediations in the	High	**R1:** Scan all AI models for malware, vulnerabilities, backdoors, and other security issues in accordance with the organization's code review and analysis policies or guidelines.	**AI RMF:** Measure 2.3, 2.7; Manage 1.1, 1.2, 1.3, 1.4

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
	development team's workflow or issue tracking system.			**OWASP:** LLM03-7d, LLM07-4
Test Executable Code to Identify Vulnerabilities and Verify Compliance with Security Requirements (PW.8): Help identify vulnerabilities so that they can be corrected before the software is released in order to prevent exploitation. Using automated methods lowers the effort and resources needed to detect vulnerabilities and improves traceability and repeatability. Executable code includes binaries, directly executed bytecode and source code, and any other form of code that an organization deems executable.	**PW.8.1:** Determine whether executable code testing should be performed to find vulnerabilities not identified by previous reviews, analysis, or testing and, if so, which types of testing should be used.	High	**R1:** Include AI models in code testing policies and guidelines. Several forms of code testing can be used for AI models, including unit testing, integration testing, penetration testing, red teaming, use case testing, and adversarial testing. **C1:** Consider automating tests within a development pipeline as part of regression testing where possible.	
	PW.8.2: Scope the testing, design the tests, perform the testing, and document the results, including recording and triaging all discovered issues and recommended remediations in the development team's workflow or issue tracking system.	High	**R1:** Test all AI models for vulnerabilities in accordance with the organization's code testing policies or guidelines. **R2:** Retest AI models when they are retrained or new data sources are added.	**AI RMF:** Measure 2.2, 2.3, 2.7; Manage 1.1, 1.2, 1.3, 1.4 **OWASP:** LLM03-7d, LLM05-7, LLM07-4
Configure Software to Have Secure Settings by Default (PW.9): Help improve the security of the software at the time of installation to reduce the likelihood of the software being deployed with weak security settings, putting it at greater risk of compromise.	**PW.9.1:** Define a secure baseline by determining how to configure each setting that has an effect on security or a security-related setting so that the default settings are secure and do not weaken the security functions provided by the platform, network infrastructure, or services.	Medium	No additions to SSDF 1.1	**AI RMF:** Measure 2.7
	PW.9.2: Implement the default settings (or groups of default settings, if applicable), and document each setting for software administrators.	Medium	**N1:** Documenting settings can be performed earlier in the process, such as when defining a secure baseline (see PW.9.1).	**AI RMF:** Measure 2.7; Manage 1.2, 1.3, 1.4
Respond to Vulnerabilities (RV)				
Identify and Confirm Vulnerabilities on an Ongoing Basis (RV.1): Help ensure that vulnerabilities are identified more quickly so	**RV.1.1:** Gather information from software acquirers, users, and public sources on potential vulnerabilities in the software	High	**R1:** Log, monitor, and analyze all inputs and outputs for AI models to detect possible security and performance issues (see PO.5.3).	**AI RMF:** Govern 4.3, 5.1, 6.1, 6.2;

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
that they can be remediated more quickly in accordance with risk, reducing the window of opportunity for attackers.	and third-party components that the software uses, and investigate all credible reports.		**R2:** Make the users of AI models aware of mechanisms for reporting potential security and performance issues. **N1:** In this context, "users" refers to AI system producers and acquirers who are using an AI model. **R3:** Monitor vulnerability and incident databases for information on AI-related concerns, including the machine learning frameworks and libraries used to build AI models.	Measure 1.2, 2.4, 2.5, 2.7, 3.1, 3.2, 3.3; Manage 4.1 **OWASP:** LLM03-7a, LLM09, LLM10
	RV.1.2: Review, analyze, and/or test the software's code to identify or confirm the presence of previously undetected vulnerabilities.	Medium	**R1:** Scan and test AI models frequently to identify previously undetected vulnerabilities. **R2:** Rely mainly on automation for ongoing scanning and testing, and involve a human-in-the-loop as needed. **R3:** Conduct periodic audits of AI models.	**AI RMF:** Govern 4.3; Measure 1.3, 2.4, 2.7, 3.1; Manage 4.1 **OWASP:** LLM03-7b, LLM03-7d
	RV.1.3: Have a policy that addresses vulnerability disclosure and remediation, and implement the roles, responsibilities, and processes needed to support that policy.	Medium	**R1:** Include AI model vulnerabilities in organization vulnerability disclosure and remediation policies. **R2:** Make users of AI models aware of their inherent limitations and how to report any cybersecurity problems that they encounter.	**AI RMF:** Govern 4.3, 5.1, 6.1; Measure 3.1, 3.3; Manage 4.3
Assess, Prioritize, and Remediate Vulnerabilities (RV.2): Help ensure that vulnerabilities are remediated in accordance with risk to reduce the window of opportunity for attackers.	**RV.2.1:** Analyze each vulnerability to gather sufficient information about risk to plan its remediation or other risk response.	Medium	**N1:** This may include deep analysis of generative AI and dual-use foundation model input and output to detect deviations from normal behavior.	**AI RMF:** Govern 4.3, 5.1, 6.1; Measure 2.7, 3.1; Manage 1.2, 2.3, 4.1 **Adv ML**
	RV.2.2: Plan and implement risk responses for vulnerabilities.	High	**R1:** Risk responses for AI models should consider the time and expenses that may be associated with rebuilding them.	**AI RMF:** Govern 5.1, 5.2, 6.1; Measure 3.3;

Practice	Task	Priority	Recommendations [R], Considerations [C], and Notes [N] Specific to AI Model Development	Informative References
			R2: Establish and implement criteria and processes for when to stop using an AI model and when to roll back to a previous version and its components. **C1:** Consider being prepared to stop using an AI model at any time and to continue operations through other means until the AI model's risks are sufficiently addressed.	Manage 1.3, 2.1, 2.3, 2.4, 4.1
Analyze Vulnerabilities to Identify Their Root Causes (RV.3): Help reduce the frequency of vulnerabilities in the future.	**RV.3.1:** Analyze identified vulnerabilities to determine their root causes.	Medium	**N1:** The ability to review training, testing, fine-tuning, and aligning data after the fact can help identify some root causes.	**AI RMF:** Govern 5.1, 6.1; Measure 2.7, 3.1; Manage 2.3, 4.1
	RV.3.2: Analyze the root causes over time to identify patterns, such as a particular secure coding practice not being followed consistently.	Medium	No additions to SSDF 1.1	**AI RMF:** Govern 5.1, 6.1; Measure 2.7, 3.1; Manage 4.1, 4.3
	RV.3.3: Review the software for similar vulnerabilities to eradicate a class of vulnerabilities, and proactively fix them rather than waiting for external reports.	Medium	No additions to SSDF 1.1	**AI RMF:** Govern 5.1, 5.2, 6.1; Measure 2.7, 3.1; Manage 4.1, 4.2, 4.3
	RV.3.4: Review the SDLC process, and update it if appropriate to prevent (or reduce the likelihood of) the root cause recurring in updates to the software or in new software that is created.	Medium	No additions to SSDF 1.1	**AI RMF:** Govern 5.2, 6.1; Measure 2.7, 3.1; Manage 4.2, 4.3

References

[1] Executive Order 14110 (2023) Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. (The White House, Washington, DC), DCPD-202300949, October 30, 2022. Available at https://www.govinfo.gov/app/details/DCPD-202300949

[2] National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[3] Vassilev A, Oprea A, Fordyce A, Anderson H (2024) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST AI 100-2e2023. https://doi.org/10.6028/NIST.AI.100-2e2023

[4] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 1270. https://doi.org/10.6028/NIST.SP.1270

[5] NIST (2024) Dioptra. (National Institute of Standards and Technology, Gaithersburg, MD.) Available at https://pages.nist.gov/dioptra/

[6] Souppaya MP, Scarfone KA, Dodson DF (2022) Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-218. https://doi.org/10.6028/NIST.SP.800-218

[7] Executive Order 14028 (2021) Improving the Nation's Cybersecurity. (The White House, Washington, DC), DCPD-202100401, May 12, 2021. Available at https://www.govinfo.gov/app/details/DCPD-202100401

[8] National Institute of Standards and Technology (2024) The NIST Cybersecurity Framework (CSF) 2.0 (National Institute of Standards and Technology, Gaithersburg, MD). https://doi.org/10.6028/NIST.CSWP.29

[9] Joint Task Force (2020) Security and Privacy Controls for Information Systems and Organizations. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-53, Rev. 5. Includes updates as of December 10, 2020. https://doi.org/10.6028/NIST.SP.800-53r5

[10] Boyens JM, Smith AM, Bartol N, Winkler K, Holbrook A, Fallon M (2022) Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-161r1. https://doi.org/10.6028/NIST.SP.800-161r1

[11] NIST (2023) NIST AI RMF Playbook. (National Institute of Standards and Technology, Gaithersburg, MD.) Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

[12] National Institute of Standards and Technology (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. (National Institute of

Standards and Technology, Gaithersburg, MD), NIST Artificial Intelligence (AI) Report, NIST AI 600-1. Available at https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

[13] National Institute of Standards and Technology (2020) NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 10. https://doi.org/10.6028/NIST.CSWP.10

[14] Stine KM, Quinn SD, Witte GA, Gardner RK (2020) Integrating Cybersecurity and Enterprise Risk Management (ERM). (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) 8286. https://doi.org/10.6028/NIST.IR.8286

[15] OWASP (2023) OWASP Top 10 for LLM Applications Version 1.1. Available at https://llmtop10.com

[16] Waltermire DA, Scarfone KA, Casipe M (2011) Specification for the Open Checklist Interactive Language (OCIL) Version 2.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) 7692. https://doi.org/10.6028/NIST.IR.7692

[17] Ross RS, McEvilley M, Winstead M (2022) Engineering Trustworthy Secure Systems. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-160v1r1. https://doi.org/10.6028/NIST.SP.800-160v1r1

[18] NIST/SEMATECH (2012) What is EDA? *Engineering Statistics Handbook*, eds Croarkin C, Tobias P (National Institute of Standards and Technology, Gaithersburg, MD), Section 1.1.1. Available at https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

[19] Reznik L (2022) Intelligent Security Systems: How Artificial Intelligence, Machine Learning and Data Science Work For and Against Computer Security. (Wiley-IEEE Press.) Available at https://ieeexplore.ieee.org/book/9562694

Appendix A. Glossary

artificial intelligence
A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. [1]

artificial intelligence model
A component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs. [1]

artificial intelligence red-teaming
A structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. [1]

artificial intelligence system
Any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI. [1]

data science
The field that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. [19]

dual-use foundation model
An AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

> (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;
>
> (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or
>
> (iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities. [1]

generative artificial intelligence
The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content. [1]

model weight
A numerical parameter within an AI model that helps determine the model's outputs in response to inputs. [1]

provenance
Metadata pertaining to the origination or source of specified data. [13]

NIST Trustworthy and Responsible AI
NIST AI 100-5

A Plan for Global Engagement on AI Standards

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-5

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Trustworthy and Responsible AI
NIST AI 100-5

A Plan for Global Engagement on AI Standards

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-5

July 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

About AI at NIST: The National Institute of Standards and Technology (NIST) develops measurements, technology, tools, and standards to advance reliable, safe, transparent, explainable, privacy-enhanced, and fair artificial intelligence (AI) so that its full commercial and societal benefits can be realized without harm to people or the planet. NIST, which has conducted both fundamental and applied work on AI for more than a decade, is also helping to fulfill the 2023 Executive Order on Safe, Secure, and Trustworthy AI. NIST established the U.S. AI Safety Institute and the companion AI Safety Institute Consortium to continue the efforts set in motion by the E.O. to build the science necessary for safe, secure, and trustworthy development and use of AI.

About this document: In accordance with Section 11(b) of Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, this plan has been developed by the Department of Commerce in coordination with the Department of State and agencies across the U.S. Government. In December 2023, NIST released a Request for Information on selected tasks related to EO 14110. More than 65 comments addressing AI standards were received. Multistakeholder listening sessions covering multiple sectors were held with representatives of federal and non-U.S. governments, businesses, academia, and civil society, which provided further input and comments. These inputs were reviewed and combined with insights from across NIST, other agencies in the Department of Commerce, the Department of State, United States Agency for International Development, and other departments and agencies into an initial public draft, released in April 2024 for public comment. This document has been revised from the initial public draft based on the 57 comments received and based on a discussion about the draft with the AI Safety Institute Consortium.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 07-25-2024

Contact Information
ai-inquiries@nist.gov

National Institute of Standards and Technology
Attn: NIST AI Innovation Lab, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8900

Additional Information
Additional information about this publication and other NIST AI publications are available at https://airc.nist.gov/Home.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Table of Contents

1. Executive Summary

Recognizing the importance of technical standards in shaping development and use of Artificial Intelligence (AI), the President's October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO 14110) calls for *"a coordinated effort…to drive the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing"* internationally. Specifically, the EO tasks the Secretary of Commerce to *"establish a plan for global engagement on promoting and developing AI standards… guided by principles set out in the NIST AI Risk Management Framework and United States Government National Standards Strategy for Critical and Emerging Technology"* (NSSCET). This plan, prepared with broad public and private sector input, fulfills the EO's mandate. As called for by the EO, within 180 days from publication of this document, NIST will report to the President on priority U.S. government actions undertaken pursuant to the plan.

The scope of the plan is deliberately broad, on several dimensions:

- U.S. standards stakeholders engage with **many kinds of interlocutors** around the globe on standards-related work, including standards development organizations (SDOs), industry, academia, civil society, and foreign governments.
- In the United States, the Federal government is just one among many participants in a dynamic, private sector-led standards ecosystem. Recognizing that U.S. global leadership on AI standards hinges on engagement by stakeholders from across this ecosystem, the plan lays out objectives, topical priorities, and actions that can be taken up not just by the Federal government but by the **full array of U.S. stakeholders in AI standards**.
- The plan addresses the **full lifecycle of standards-related activities**. This includes the foundational technical work that is typically necessary before a formal standard can be developed, the collaborative process of developing a consensus standard, and the development of complementary tools to help with implementing the standard.
- The plan addresses **AI-related standards of all scopes,** both "horizontal" (applicable across sectors) and "vertical" (designed for the needs of a particular sector).

The engagement activities identified in this plan are aimed at furthering four sets of outcomes:

- **Scientifically sound AI standards that are accessible and amenable to adoption.** Standards are most likely to meet these criteria when they are grounded in underpinning science and technical experience, clear, implementable, understood as neutral and unlikely to impede innovation, and globally accessible in a timely fashion. "Horizontal" (non-sector-specific) AI standards can further promote adoption by serving sectoral needs and minimizing the need for per-sector adaptation.
- **AI standards that reflect the needs and inputs of diverse global stakeholders.** Standards are most likely to achieve this if they are context-sensitive, performance-based, human-centered, and responsive to societal considerations. Though views on societal considerations vary, existing international instruments can provide a starting point for finding consensus, and standards participants can reflect their existing commitments in their standards work. AI standards are

most likely to reflect stakeholders' needs when they are based on inputs from participants from many backgrounds and geographic areas who are empowered to contribute meaningfully to standards development. Human-centered design approaches for standards may be helpful.

- **AI standards that are developed in a process that is open, transparent, and consensus-driven.** The United States continues to support standards efforts that are voluntary and market-driven—with the Federal government participating as one among many stakeholders—so that standards can best achieve consensus and serve stakeholder needs. It is also well-established that standards are most technically sound, independent, and responsive to needs when developed through open, transparent, consensus-driven processes.
- **International relationships that are strengthened by engagement on AI standards.** These engagement activities can strengthen relationships between participating experts and contribute to broader cross-border connections between companies, governments, and other stakeholders.

Engagement on AI standards will often need to focus on specific AI-related topics that standards could address. The plan defines three tiers of topics based on a qualitative assessment of need, impact of global involvement on the standards, urgency, and maturity of technical foundations. The tiers are:

1. **Urgently needed and ready for standardization.** These topics include terminology and taxonomy; testing, evaluation, verification, and validation (TEVV) methods and metrics, including cross-cutting TEVV methods and TEVV for bias; mechanisms for enhancing awareness and transparency about the origins of digital content, including for synthetic content; risk- based management of AI systems; security and privacy; transparency among AI actors about system and data characteristics; incident response and recovery plans; and training-data practices. Some of these topics have sub-topics that need additional study, perhaps on an accelerated timeline, before being ready for standards.
2. **Needed, but requiring more scientific work or maturity before standardization.** These topics include energy consumption of AI models; conformity assessment with other standards; and testing and evaluation datasets.
3. **Needed, but requiring significant foundational work.** These topics include techniques for interpretability and explainability and configuring human-AI interactions for effective decision-making and operations.

The plan recommends a variety of engagement activities relevant to all stakeholders, along with some specific high-priority ways for the U.S. government to implement the broader recommendations. Recommended activities include:

1. **Prioritize engagement in standards work, including pre-standardization research and related technical activities.** This includes bolstering foundational research on priority topics; facilitating timely development of standards by participating in standards development; encouraging alignment between sectoral practices and standards via thoughtful design and use of horizontal standards; developing and sharing implementation tools; and exploring processes for tighter feedback loops with potential standards adopters. Priority U.S. government implementation actions focus on strategic guiding of research efforts, agency participation in standards work,

information exchange within and beyond the government, and international collaborations with a standards component.

2. **Facilitate diverse multistakeholder participation in AI standards development and adoption.** *Domestic capacity-building engagements* could include regularly convening AI standards stakeholders; developing and disseminating information such as standards training and handbooks for AI stakeholders; and devoting organizational resources to standards participation. *Global capacity-building engagements* could include broadening global access to frameworks and standards; increasing resources to support diverse participation in AI standards development; bringing education about AI standards to the settings where global AI experts gather; and building a global scientific network of AI standards experts. Priority U.S. government implementation actions focus on internal capacity-building, resourcing, and education, making relevant U.S. government documents broadly accessible, foreign assistance programming, and prioritizing work with countries at varied stages of development.
3. **Promote global alignment on AI standards approaches,** seeking common understandings of the role standards can and should play in the broader AI ecosystem. Engagements could include advocating for a standards ecosystem driven by multistakeholder involvement and global consensus; arranging exchanges among experts from different countries on global needs, priorities for, and experiences with AI standards; continuing to seek alignment between varied frameworks, but focusing on standardization where possible; and convening discussions on the relationship between standards and open-source software. Priority U.S. government actions focus on working AI standards issues into diplomatic meetings, communications, and outputs, including via interagency collaboration.

2. Introduction

As a leader in Artificial Intelligence (AI), the United States recognizes the importance of advancing technical standards for safe, secure, and trustworthy AI development and use. Toward that goal, this document establishes a plan for global engagement on promoting and developing AI standards. The plan calls for outreach to and engagement with international stakeholders and standards developing organizations to help drive the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing.

This plan furthers the policies and principles in the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO 14110), which instructs the Federal government to "promote responsible AI safety and security principles and actions with other nations, including our competitors, while leading key global conversations and collaborations to ensure that AI benefits the whole world, rather than exacerbating inequities, threatening human rights, and causing other harms." By advancing AI standards globally with these goals in mind, the U.S. government seeks to assist both the private and public sectors to seize the benefits of AI while managing risks to people domestically and across the globe.

Standards play a crucial role in the development and adoption of new and emerging technologies. They are especially important in the field of AI, where policymakers and regulators in the United States and

abroad are looking to the standards ecosystem to guide AI actors[1] on how to implement high-level principles and policies. This plan, developed in accordance with Section 11(b) of the EO, highlights how engagement by U.S. stakeholders, including the U.S. government, on technical standards for AI technologies can enhance global cooperation, coordination, and alignment.

For the purpose of this plan, "technical standards" refer to "documentary" standards.[2] ISO/IEC[3] Guide 2:2004 Standardization and related activities—General vocabulary[4] defines such a standard as "a document, established by consensus and approved by a recognized body, that provides for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context." This plan refers to these simply as "standards." Standards can be developed in many types of organizations that cover a broad spectrum of formality, structure, and approach.[5] Standards are typically adopted and implemented on a voluntary basis, although they can support implementation of specifications outlined in policies and regulations.

The plan is guided by principles set out in the National Institute of Standards and Technology (NIST) AI Risk Management Framework[6] (AI RMF) and U.S. Government National Standards Strategy for Critical and Emerging Technology[7] (NSSCET). The NIST AI RMF, released in January 2023, is a framework to better manage risks to individuals, organizations, and society associated with AI. It is intended for voluntary use to improve the ability of organizations to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. The framework was developed through a consensus-driven, open, transparent, and collaborative process with the private and public sectors.

The NSSCET recognizes the importance of standards to enable technology that is safe, universal, and interoperable. That strategy renews the United States' rules-based approach to standards development. It also emphasizes the Federal government's support for international standards for critical and emerging technologies, which will help accelerate standards efforts led by the private sector to facilitate global markets, contribute to interoperability, and promote U.S. competitiveness and innovation. AI is one of those technologies.

This plan also expands on the priorities outlined in the Plan for Federal Engagement in AI Standards and Related Tools.[8]

---

[1] For the purposes of this document, an AI actor is any person or organization performing any of the AI actor tasks defined in Appendix A of the AI Risk Management Framework: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A

[2] This is in contrast to other types of standards such as measurement standards (which are often physical), standard reference materials, and standard reference data.

[3] ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission)

[4] https://www.iso.org/standard/39976.html

[5] Open-source software (OSS) has a nuanced relationship with standards. The concepts are not the same, but there can be commonalities and overlaps, and they are often subject to the same policies. See also Section 5.3

[6] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[7] https://www.nist.gov/standardsgov/usg-nss

[8] https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

This plan addresses activities before, during, and after the creation of a formal standard. Before a standard can be developed, a foundational body of scientific and technical work is typically needed. That includes producing guidelines that might form the basis for a standard and building consensus within SDOs around other informative documents such as technical reports. The standards development process draws from this foundational material to establish consensus on the rules, guidelines, or characteristics that make up the standard. Once a standard is finalized, complementary standards-related tools are often needed to help with implementation. These include datasets, benchmarks, reference implementations, implementation guidance, verification and validation tools, and conformity assessment procedures (some of which may themselves be standards). Activities related to all these stages are in scope for this plan.[9]

3. Desired Outcomes from Engagement on AI Standards

Standards-related engagement activities are most effective when they aim for clear, specific, achievable objectives. The actions laid out in this plan are designed to further the outcomes below with respect to AI standards.

3.1. Scientifically sound AI standards that are accessible and amenable to adoption

A central purpose of standards and related tools is to facilitate safety, interoperability, and competition, including by lowering barriers to trade. Their ability to achieve that purpose depends on being widely accepted and implemented. As in other technological domains, while some AI standards may be required by government regulations, most standards will generally have to be adopted voluntarily to be effective—which organizations will do if they find the relevant standards **implementable and useful**.

New standards typically are based on technical insights and novel discoveries from scientific research and innovation. The more grounded a standard is in the underpinning science and in stakeholders' real-world experiences, the more implementable and useful it will be for the global AI community, and the greater its chances of international adoption. Conversely, a standard that would attempt to get ahead of the underpinning science and engineering may be built on less rigorous technical foundations; it may prove unhelpful, counterproductive, or even technically incoherent. The same holds true for related tools.

Accordingly, where a science-backed body of work exists, AI standards can be developed in a timelier fashion. Where there are gaps in foundational understanding (see Section 4.3), new research can fill those gaps so that implementable and useful standards can be developed.

To achieve widespread adoption, standards need to be clear, implementable, viewed as unlikely to unduly inhibit innovation or raise intellectual property (IP) concerns, understood to be neutral (i.e.,

[9] In other words, this plan broadly addresses "standards-related activities" as defined by the Department of Commerce, including any "action taken for the purpose of developing, promulgating, revising, amending, reissuing, interpreting, implementing or otherwise maintaining or applying…a standard."

without favoring specific nations or technology solutions), and accessible in a timely fashion to potential users across the globe.

One particularly important adoption-related issue is sectoral adoption or adaptation of horizontal standards (those intended to be used across many applications and industries). Horizontal standards may directly serve the needs of a given sector, but sector-specific practices, clarifications, and adjustments will also often be needed. In such cases, horizontal standards will be most amenable to adoption and implementation if they serve many or most sectoral needs, minimize necessary adaptation, and provide for interoperability across sectors.

Facilitating implementation of AI standards may require creating and maintaining additional standards-related tools such as datasets, benchmarks, reference implementations, implementation guidance, verification and validation tools, and conformity assessment procedures. As with the AI standards themselves, these additional standards, procedures and tools are most useful and amenable to adoption when developed and made available through inclusive, fair, and transparent processes.

3.2. AI standards that reflect the needs and inputs of diverse global stakeholders

AI standards will be most useful if they respond to the needs of a diversity of potential users and other stakeholders around the world. Standards are most likely to achieve this if they are:

- **Context-sensitive**, providing the necessary flexibility to enable adoption by small, medium, and large entities in their own contexts of use;
- **Performance-based**, providing flexibility by focusing on outcomes rather than prescribing specific ways of achieving those outcomes;
- **Human-centered**, accounting for people's needs and how they interact with the system; and
- **Responsive to societal considerations** that may arise from the design, development, deployment, or use of the technologies**.**

Views of what societal considerations should be reflected in AI standards are likely to vary across international contexts and stakeholders. However, attempts to find common ground on commonly accepted societal considerations can be anchored in bilateral, multilateral, regional, and global agreements. This includes international human rights instruments, particularly those that articulate governments' duties to protect people's rights and private actors' responsibilities to respect people's rights. Participants in standards development activities often represent organizations and governments that have expressed human rights commitments (see text box). Other commonly discussed societal considerations regarding AI include concerns around privacy and IP rights. The policy and legal issues surrounding these topics will be context-specific, but, as above, some international instruments address these issues, and many SDOs and standards development participants have expressed relevant commitments or policies. Stakeholders can help standards be more responsive to such considerations by reflecting any prior commitments or legal obligations in their standards development activities and in their discussions about technical standards in international policy fora.

Human rights commitments as they relate to technical standards

Participants in standards development activities include representatives from many governments and organizations that have expressed commitments to human rights. Governments have expressed these commitments by signing the United Nations (UN) Universal Declaration of Human Rights[10] and joining human rights treaties. Many public and private actors have endorsed instruments such as the UN Guiding Principles on Business and Human Rights.[11] Some SDOs have indicated a desire to align their work with the broader context of international human rights framework. For example, the IEEE's Ethically Aligned Design[12] vision for autonomous and intelligent systems states that these systems should not infringe on human rights as its first principle. Similarly, ISO 26000: Guidance on Social Responsibility[13] includes respect for human rights as a principle and emphasizes the role of human rights due diligence. Alongside many partner governments, the U.S. Government remains committed to protecting human rights in all its activities, including standards-setting for emerging technologies such as AI, as reflected in UN Human Rights Council Resolution 53/29.[14] The U.S. Department of State, in close coordination with NIST and the U.S. Agency for International Development (USAID), has developed a "Risk Management Profile for AI and Human Rights." Informed by extensive multi-stakeholder consultations, the Profile demonstrates how to apply NIST's AI Risk Management Framework in concert with the international human rights framework.

AI standards are more likely to reflect stakeholders' needs if they are based on inputs from participants with diverse backgrounds and expertise. Especially given that AI standards so frequently involve *sociotechnical* phenomena—that is, interactions between technical systems and people (see Appendix A.3)—it is helpful for AI standards development to draw on insights from a broad set of multi-disciplinary stakeholders including enterprises of various sizes, governments, civil society, and academics.

Similarly, the needs of stakeholders from countries and regions around the world may not be reflected if a standard is not developed with adequate geographic representation (see text box below). Standards developers can address global needs by bringing geographically diverse stakeholders to the table and remaining sensitive to their concerns and views.

AI standards needs around the globe

In some cases, low- and middle-income countries particularly stand to benefit from AI innovations, which can accelerate progress on sustainable development through applications such as identifying better agricultural practices or strengthening health systems. At the same time, these countries can also be disproportionately vulnerable to certain risks, such as labor market disruptions or AI-enabled cybercrime. Countries around the world may also vary regarding which needs among the UN Sustainable Development Goals (SDGs)[15] are most acute for them, and AI stands to have significant impact on many SDGs. Without meaningful participation by representatives from diverse countries, including low- and middle-income countries, AI standards may not fully reflect such priorities and concerns.

[10] https://www.un.org/en/about-us/universal-declaration-of-human-rights
[11] https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf
[12] https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
[13] https://www.iso.org/iso-26000-social-responsibility.html
[14] https://undocs.org/A/HRC/RES/53/29
[15] https://sdgs.un.org/goals

Stakeholders from all backgrounds and regions will be better equipped to influence standards if they have the necessary knowledge about both AI technologies and standardization processes. They may also need to be prepared to communicate and seek mutual understanding of conceptual frameworks, areas of expertise, and field-specific expectations.

One way to maximize AI standards' value could be to develop such standards following a human-centered design approach, where stakeholder needs are analyzed at the outset of a project and then guide the work. This approach can be particularly useful for AI standards development as AI requires an understanding of risks, impacts, and potential harms with multiple AI actors working together to manage those risks to achieve trustworthy AI. Such an approach could also provide a basis for assessing how successfully standards are meeting various stakeholders' needs once completed.

3.3. AI standards that are developed in a process that is open, transparent, and consensus-driven

In the United States, documentary technical standards are overwhelmingly developed through open, consensus-driven, private sector-led processes within domestic and international SDOs. As articulated in the NSSCET, the United States supports standards efforts that are voluntary and market-driven. The Federal Government engages primarily through foundational research, coordination, education, and participation in standards development processes as one of many stakeholders. Retaining this model for AI standards, with standards development **led by the private sector and featuring diverse representation from industry, civil society, government, and academia,** will help ensure that the standards meet the needs of those who will need to apply them or will be affected by them and that the standards reflect broad consensus.

It is well-established that standards development is best done through an **open, transparent, consensus-driven process**.[16] This helps ensure that the resulting standards are technically sound, independent, and responsive to broadly shared market and societal needs—all characteristics that are as important for AI standards as for other areas.

Governments that desire to promote or require standards can best facilitate both technical interoperability and regulatory alignment by using consensus-driven standards. Where **international standards**[17] are available, using them to the maximum extent possible reduces market friction and incompatibility and promotes efficiencies for buyers and sellers alike. Use of international standards as a

[16] As noted in the NSSCET, the six principles that traditionally govern the international standards development process are transparency, openness, impartiality and consensus, effectiveness and relevance, coherence, and a commitment to participation by low- and middle-income countries.

[17] The World Trade Organization (WTO) defines "international standard" as "a standard adopted by a governmental or non-governmental body whose membership is open to all [WTO] Members, one of whose recognized activities is in the field of standardization." (https://www.wto.org/english/docs_e/legal_e/21-psi_e.htm#fnt-2)

means to facilitate trade is encouraged in the World Trade Organization Technical Barriers to Trade Agreement.[18,19]

3.4. International relationships that are strengthened by engagement on AI standards

Global engagement activities, such as active participation in standards bodies, fora, and bilateral expert exchanges, can strengthen relationships between the experts who will need to come to consensus through the standards development process. These relationships can facilitate information flow among SDO participants even outside of formal engagements and make it easier to identify common views and approaches.

In addition, engagement activities contribute to broader cross-border connections between companies, governments, and other stakeholders. For example, as creators of different frameworks of guidelines compare them with each other, they may build relationships that form the foundation of future business collaborations or diplomatic exchanges.

4. Priority Topics for Standardization Work

This plan defines three tiers of topics for engagement in international AI standardization work. The tiers are based on the degree to which:

- Experts and stakeholders have identified a well-defined, widely recognized need for international AI consensus standards;
- Global involvement can substantially enhance the speed, quality, relevance, or adoption of the resulting standards;
- Delivering standards in a timely fashion would significantly enhance the impact of those standards, including by providing a prerequisite foundation for further practices and standards; and
- Foundational scientific work exists or can be enhanced to develop technically robust standards that meet identified needs.

The tiers have been formulated by applying these criteria qualitatively. As the standardization community's needs and collective knowledge change over time, urgency and readiness may need to be reassessed. The community would benefit from further thinking—already underway in some fora, including NIST—on how to rigorously evaluate a topic's readiness for standardization, and such thinking can inform ongoing prioritization.

---

[18] https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm

[19] When the U.S. government identifies standards are needed but international standards are not available, the U.S. government seeks to work with the private sector to develop an approach for addressing the gap.

Some topics have multiple subtopics that are at different levels of standardization readiness. For ease of elaboration, each topic is addressed only in one tier but with discussion of which subtopics are more or less mature.

4.1. Urgently needed and ready for standardization

Top-priority topics are those where global stakeholders have identified a pressing need for a standard, including based on likely marketplace impact; accelerating the work would offer significant payoff; and there exists a reasonable scientific underpinning. These topics are urgent in the sense that certain foundational standards can either immediately increase the trustworthiness of AI systems or be the basis for developing further practices and standards that facilitate the responsible adoption of AI and sector-specific use cases. The payoff may come from producing a consensus standard based on existing foundational scientific work, if that is already feasible, or from bringing the community closer to agreeing on a highly-impactful future standard that would help to advance innovation, trustworthiness, and market acceptance.

For the topics listed below, the available scientific work provides a strong basis for developing standards on at least one sub-topic. In some cases, there are additional sub-topics where further research needs to be conducted before the issue can be fully addressed by standards. Although these less mature sub-topics would thus fit well as priority areas for accelerated study in the next section, for ease of elaboration they have nonetheless been included in this section under the broader topical headings. In many cases, ongoing research and updates will be required even for more mature sub-topics as technologies progress.

Topics meeting these criteria include:

- **Terminology and taxonomy**. Existing standards on AI concepts and terminology (e.g., ISO/IEC 22989:2022[20]) provide a critical starting point, but further clarity and alignment on terminology are needed. This is particularly true for terms related to developments in AI since the existing standards were developed, including (but not limited to) concepts related to generative AI models, model fine-tuning, AI red-teaming, and model openness. Such terms and concepts underlie many other standards, policy discussions, and regulations, so technical consensus on the terminology would quickly yield wide-ranging benefits. Multiple projects outside of SDOs (e.g., academic papers,[21] the U.S.-European Union (EU) Trade and Technology Council,[22] and the work of the Organisation for Economic Co-operation and Development [OECD]) provide extensive thinking to draw upon for standardizing such terms, and updates to some existing terminology standards are already underway (e.g., ISO/IEC 22989 AMD1).
- **Measurement methods and metrics**. Shared testing, evaluation, verification, and validation (TEVV) practices for AI models and systems would open the way for more rigorous discussions about capabilities, limitations, risks, benefits, appropriate or inappropriate use, AI assurance,

[20] https://www.iso.org/standard/74296.html
[21] https://crfm.stanford.edu/assets/report.pdf
[22] https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence

and more. In particular, TEVV standards would define the performance metrics for performance-based standards, which in turn allow defining what constitutes an effective risk mitigation. Completed and ongoing foundational research in this space (e.g., NIST AI 200-2[23], ISO/IEC AWI TS 17847) provides a rich platform for standardization work, particularly on cross-cutting TEVV methods (e.g., demonstrating construct validity) and appropriate TEVV practices for bias and fairness risks, where much foundational work (e.g., NIST SP 1270[24]) has taken place. Further work may be needed before it is possible to standardize some testing and evaluation protocols (e.g., AI red-teaming and evaluating impacts on people in realistic settings), some types of measurements (e.g., measuring effectiveness and robustness of interventions to mitigate risks), and what TEVV approaches are most effective for some types of risks (e.g., risks to safety or information integrity), all of which would merit accelerated study.

- **Mechanisms for enhancing awareness and transparency about the origins of digital content**, particularly of whether content is AI-generated or AI-modified, as well as broader information on content's origins, history, and context. An example of a type of mechanism that may be mature enough for standardization is metadata recording (a technique for provenance data tracking), for which standardization work is already underway. Other sub-topics such as watermarking and synthetic content detection merit accelerated study, including assessments of efficacy, across modalities to help address widespread and pressing concerns about the societal impacts of synthetic content.
- **Risk-based management of AI systems.** Existing frameworks (NIST AI RMF) and standards (ISO/IEC 23894:2023, ISO/IEC 42001) provide an important basis for risk-based management of AI systems. Such process-oriented frameworks are especially important in the absence of TEVV metrics that would support performance-based standards. However, more work is needed to adopt or revise those documents to account for changes in the technology as well as risks for specific applications, contexts, or industry verticals. The existing documents and stakeholders' experiences with them offer a strong basis for this work. More foundational work is needed to establish how organizations can best determine and apply their risk tolerances.
- **Security and privacy.** While many traditional cybersecurity practices apply naturally to AI systems, AI technologies also introduce a variety of new security issues. The latter category of distinct risks encompasses adversarial machine learning attacks, which include risks to the integrity of AI algorithms and data and the confidentiality of data that have been used to train an AI system (often a privacy issue). A related issue is when and how various privacy-enhancing technologies (PETs) can be used to improve privacy and security. Standards are needed for many of these topics, including mitigations against AI-specific threats, and there is a foundation of technical work to draw from (e.g., NIST AI 100-2 E2023[25]).
- **Transparency among AI actors about system and data characteristics.** System deployers and users often need information from designers and developers about training data, performance

[23] https://doi.org/10.6028/NIST.AI.200-2, to be published.
[24] https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence
[25] https://csrc.nist.gov/pubs/ai/100/2/e2023/final

testing results, areas of intended or unintended use, AI systems' supply chains (the underlying software, data, and model components), and the like. These needs and mechanisms for filling them such as model cards, data cards, software bills of materials, and other disclosure mechanisms, have been well-studied, laying the groundwork for standardization. It may also be helpful for users and deployers to have standardized ways to share information upstream to designers and developers about usage patterns and issues that have been observed in deployment (including security vulnerabilities that have been discovered). This sub-topic would require further research before standardization.

- **Training data practices.** Challenges in managing training data for AI systems include data quality maintenance (especially in very large datasets), preprocessing technique selection, dataset change management, efficient use of limited data, managing diverse data formats, and identification of data intended to be permitted for or excluded from training use. Similar issues have been studied and standardized extensively in standards bodies (e.g., the ISO/IEC SC 42 working group on data for AI systems), and work is already underway on fleshing out frameworks further and adapting them to recent types of datasets and systems.
- **Incident response and recovery plans.** Some organizations already implement such plans in their own ways, and other fields may offer informative insights, particularly cybersecurity but also non-computational fields such as human rights and healthcare. Organizations such as the OECD[26] have developed insights on key AI-specific questions such as what constitutes an AI incident and how such incidents would be identified and shared, offering some basis for standardization. One subtopic that may need more work is how different kinds of incidents can be addressed via plans, policies, and procedures, which may include proactive baseline mitigations as well as responsive controls after a risk has been demonstrated. Further work will also be needed to align horizontal AI practices with existing sectoral standards and policies.

4.2. Needed, but requiring more scientific work or maturity before standardization

This grouping encompasses topics where there is a clear need for standardization, but more work is needed on most or all sub-topics before a standard can be developed.

For some topics, the path to standardization is longer due to a lack of foundational understanding about metrics, methods, or other critical components of a potential standard. These topics include:

- **Measuring resource consumption of AI models.** As some kinds of AI models have become both more compute-intensive and more widely used, concerns about cost and environmental impacts have grown in tandem. Though research has explored measurement methods and metrics for measuring energy usage, standardized approaches remain an important technical gap, and more foundational work seems necessary before standardization work can begin in earnest.

[26] https://www.oecd-ilibrary.org/science-and-technology/defining-ai-incidents-and-related-terms_d1a8d965-en.

In other cases, there is a need for tools for implementing standards, but these tools would be difficult to develop before the base standards exist. Topics with payoffs that are more distant for this reason include:

- **Conformity Assessment.**[27] Conformity assessment and compliance procedures can provide confidence that the specifications in a given standard have been met, but they depend on having first defined the standardized practices with which to assess conformity. As more standards mature to a point where this is possible, conformity assessment may rise in priority.[28] Developers of the underlying standards can also address conformity assessment needs by designing the standards to be amenable to conformity assessment.
- **Testing and evaluation datasets.** To implement testing and evaluation protocols, it is often necessary to have agreed-upon datasets before applying those protocols. Those datasets may also need to be subject to standard practices for data integrity, data quality assessment, change management, and data formats. Moreover, settling on standard datasets would depend on having reached consensus on what and how to test and evaluate (see TEVV, above).

4.3. Needed, but requiring significant foundational work

This priority consists of topics where standards would be helpful, but significant foundational work (e.g., foundational research and development) remains to be done. Examples include:

- **Techniques for interpretability and explainability.** There is ongoing research on how to better help users, affected individuals, and other stakeholders make sense of AI system outputs (e.g., NISTIR 8367,[29] Gunning et al. 2021[30]). Existing research has proposed many techniques for explainability—providing information about how an AI system makes its decisions and why it behaved as it did. However, establishing empirically to what extent such techniques are useful for what purposes remains a significant gap. Techniques for interpretability, or enabling humans to understand how to act on system output, are also needed. Discussion around interpretability and explainability standards should consider the extent to which testing and transparency may yield benefits similar to those achieved with these techniques.
- **Human-AI configuration**. Interactions between humans and AI systems are generally designed with the goal of producing effective decision-making and operations. Configuring AI systems, people, organizations, and their relationships to achieve this will rely on a number of measurement methods and metrics, including for performance, bias, and trust. Metrics and potential standards in this area will be important for training, testing, and evaluation of human-AI teaming before wide-scale integration into critical operations.

---

[27] https://www.iso.org/conformity-assessment.html
[28] One example is ISO/IEC 42001 – Management System, for which conformity assessment standards are already being developed.
[29] https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf
[30] https://onlinelibrary.wiley.com/toc/26895595/2021/2/4

- **Measurement of AI hardware performance and resource use.** Many kinds of specialized hardware, such as neural processing units (NPUs) and tensor processing units (TPUs), have been developed to accelerate AI-related computations. Companies and researchers are also experimenting with novel architectures that may not even share the same technical building blocks as traditional chips, and some kinds of hardware allow trading off characteristics such as precision and speed. Such diversity may make it difficult to compare performance and resource use (e.g., energy efficiency) across hardware platforms, which in turn would make it difficult to determine when a given form of hardware is an improvement and how it relates to existing procedures and policies. Foundational work may be needed to identify procedures, metrics, and benchmarks that would enable appropriate comparisons.

An additional research need, beyond the development of specific standards, is assessing the effectiveness of standards. Considered within the context of explosive growth in global trade, standards impact trillions of dollars of trade—with benefits and costs well beyond their economic implications. Nevertheless, research assessing the effectiveness of standards focuses primarily on specific examples of their use. (One NIST study[31] estimated a $250 billion economic impact just from the development of its Advanced Encryption Standard over a 20-year period.) With the emergence and forecasted explosive growth of AI technologies, AI and standards stakeholders would benefit from a more explicit and quantitative estimate and understanding of the effectiveness of AI standards—and economic impact is only one way to assess that effectiveness.

5. Recommended Global Engagement Activities

EO 14110 directs the Department of Commerce to "establish a plan for global engagement on promoting and developing AI standards." In this case, "engagement" includes a wide variety of ways U.S. standards stakeholders can interact with current and potential stakeholders across the globe.

In recognition that AI presents global issues that require global solutions, and that AI standards, like other standards, require investment and engagement across society, the core recommendations below are scoped more broadly than U.S. government activity. These are recommendations for U.S. stakeholders writ large, and many will depend on private-sector leadership and joint efforts from the global AI and standards communities. Specific suggestions for how the U.S. government could implement these recommendations are included in text boxes, with suggested lead departments and agencies in parentheses.

5.1. Prioritize engagement in standards work, including research and related technical activities

By advancing research that can underpin standards, working on standards projects, and developing tools that facilitate adoption, AI actors and relevant stakeholders can contribute directly to standardization

[31] https://www.nist.gov/news-events/news/2018/09/nists-encryption-standard-has-minimum-250-billion-economic-benefit

and lead by example. They can take the following actions to increase their direct involvement in standardization activities on AI and maximize its effectiveness:

- **Bolster foundational (pre-standardization) research on the priority topics listed above** by increasing investment in and focusing on relevant research, emphasizing international collaboration whenever appropriate and possible. When developing practices, frameworks, and guidelines that may ultimately inform standards, seek out and incorporate perspectives from as wide a range of stakeholders as is feasible.
- **Facilitate timely development of science-backed consensus-based, voluntary standards** by participating, contributing to, influencing, or leading standards development efforts. To maximize timeliness, consider the full range of SDOs, product types, and existing mechanisms for shortening development processes. Promote international cooperation whenever appropriate and possible, and prioritize SDOs and projects whose products are likely to meet the objectives described in Section 3. Work toward both horizontal and vertical standards as appropriate.
- **Encourage alignment between sectoral practices and standards** by striving to develop international horizontal standards that are as amenable as possible to adoption or adaptation across sectors, including by maximizing their incorporation of, or reference to, global documents (e.g., terminology lexicons and taxonomies). When developing vertical standards, build on and back-reference applicable horizontal standards.
- **Develop and widely share tools to assist with implementing standards and guidelines.** Make them as accessible as possible, including to potential users (organizations and nations) who may have fewer resources available to maintain awareness of standards and develop their own implementation tools.
- **Explore processes for potential standards adopters to share insights via direct feedback loops.** Human-centered design principles may offer inspiration for maximizing the usefulness, buy-in, and likelihood of adoption for AI standards. For example, potential adopters outside of SDO committees might be interested in helping to articulate a standard's goals, determine whether the standard achieves the goals, track adoption, and share experiences with implementation.

High-priority implementation actions specific to the U.S. government

- Identify and allocate resources to priority AI work related to horizontal and vertical standards projects that align with agency missions and encourage participation by agency experts, in line with existing policies on SDO participation.[32] (NIST, Department of Homeland Security [DHS], U.S. Food and Drug Administration [FDA])
- Promote international collaboration on pre-standardization research. (National Science Foundation [NSF], NIST)
- Consult with private sector and civil society organizations about AI standards-related priorities and views. This could include views on SDO projects to prioritize for participation, potential U.S. contributions and proposals, and finalized standards to promote adoption of. (NIST)
- Share priorities and views on AI standards development and adoption between agencies, including sector-specific agencies. Identify intersections between standards work and AI policy as well as ways to optimize interagency collaborations and coordination to improve effectiveness and efficiency. Utilize current interagency mechanisms, especially the AI Standards Coordination Working Group. (NIST, DHS)
- Work on standards development projects jointly with other governments around the globe (see Section 5.3). (NIST, State)
- Strategically direct research efforts toward priority topics. (NSF, Department of Energy, Department of Defense, Department of Health and Human Services, National Institutes of Health, NIST)
- Leverage opportunities to align and collaborate on standards such as Joint Committee Meetings, AI working groups, public-private partnerships. (NIST, State, International Trade Administration [ITA])

5.2. Facilitate diverse multistakeholder engagement in AI standards development and adoption

Many stakeholders domestically and abroad could benefit from more extensive engagement with standards processes. Special attention should be given to drawing in stakeholders from all regions and backgrounds, particularly those who have historically been less well represented in standards development processes. Considering the breadth and scale of potential harms and benefits related to AI, it is critical that these voices be part of the standards development process. For standards to have the desired impact, it is also essential that diverse global stakeholders are aware of the standards and able to implement them. These needs call for building greater capacity and awareness among potential contributors to standards development and potential adopters.

5.2.1. Domestic capacity-building

- **Regularly convene stakeholders on AI standards.** As AI standards-related activities, including research, increase, so too do the opportunities for expanding training and the exchange of

[32] E.g., OMB Circular A-119 (https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-119-1.pdf) and M-12-08 (https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2012/m-12-08_1.pdf).

information on AI standardization and discussion of AI standards issues. When groups convene on AI standards matters, they have a platform for encouraging robust information exchange about the substance of AI standards and the process of their development among subject matter experts (SMEs) in the private sector, academia, and civil society. Some such SMEs may have knowledge of AI but less experience with standards. Pre-meeting tutorials, ancillary discussions outside formal standards development sessions, and information exchange about adoption can aid in creating a more informed, more diverse, and more capable AI standards community.

- **Develop and disseminate information, including training and handbooks on standards development, participation, and adoption, for AI stakeholders**. Building on existing material, prepare and promote materials to help those from small- and medium-sized companies, academia, and civil society to understand how, where, and when they can provide their input to be most effective—including mechanisms for contributing to and improving U.S. inputs on international standards. In addition, host workshops and engagement exercises to facilitate AI standards literacy, including in partnership with the Small Business Administration (SBA) and other federal agencies.
- **Support standards participation with organizational resources**. Prioritize AI standardization staffing needs in organizational decisions about budgets, training programs, and staff incentives. Provide materials that articulate the value of standards participation and use them to make the case for prioritizing and incentivizing participation in standards work.

High-priority implementation actions specific to the U.S. government

- Increase agencies' capacity for standards participation, including when making resourcing decisions and setting staff work expectations and developing incentives. (NIST, DHS, FDA)
- Convene periodic meetings for knowledge sharing between government AI standards experts, including some open to private sector and civil society. (NIST)
- Educate U.S. government staff on the importance and benefits of participating in standards activities, including clarifying policies on committee participation and leadership as a U.S. government representative. (NIST)
- Promote participation in standards activities by recipients of federal R&D funding. (NSF, NIST)

5.2.2. Global capacity-building

- **Broaden global access to frameworks and standards.** Translate higher-priority AI standards-related documents into multiple languages. For standards that are not freely available, explore mechanisms for increasing access, particularly for potential users in developing counties.
- **Increase resources to support diverse participation in AI standards development.** Provide or fund training on participation for international stakeholders, particularly non-traditional standards participants such as those from small- or medium-sized entities, academia, and civil society and particularly those from low- and middle-income countries.
- **Bring education about AI standards to the settings where AI experts gather.** In particular, look for ways to raise awareness about standards work at AI conferences (e.g., via an AI standards

"roadshow"). These conferences bring together large groups of academics and industry practitioners, many of whom have little awareness of the standards ecosystem but much AI-related expertise across a variety of domains to contribute. Online fora where AI experts congregate virtually also are fruitful avenues for education and raising awareness.

- **Build a global scientific network of AI standards experts.** Collaboration on standards development could be facilitated by a scientific network of AI standards experts across the globe, ideally including strong representation from low- and middle-income countries. This network could be called upon for standards-specific work, guidance on implementing standards, knowledge about potential impacts of standards, and possibly scientific input on global AI issues as they emerge.

High-priority implementation actions specific to the U.S. government

- Translate key U.S. government documents, AI standards, and related resources into multiple languages. (NIST, State)
- Incorporate private sector participation or bilateral private sector exchanges into existing government-to-government engagements such as technology dialogues. (State, ITA)
- Leverage foreign assistance funds and other diplomatic programming, in collaboration with civil society and the private sector, to arrange training and support for SDO participation by stakeholders in partner countries. (State, USAID, U.S. Trade and Development Agency, ITA)
- Expand resources for government bodies that facilitate standards development. (Commerce, DHS, FDA)
- Prioritize countries for engagement that are in varied stages of development. (State, ITA, USAID)

5.3. Promote global alignment on AI standards approaches

The standards ecosystem provides the greatest value when parties around the world that develop, use, or are affected by standards and guidelines are aligned on those documents' role and how they fit into the broader AI ecosystem. This includes widespread recognition of standards' potential to minimize trade barriers by facilitating compatible practices. Stakeholders can work toward alignment on the role of standards through the following activities:

- **Advocate for a standards ecosystem driven by global stakeholder involvement and consensus.** In a coordinated fashion, push for standards-setting activity to take place in open, transparent, consensus-driven venues. Where applicable, prefer developing international standards over developing domestic or regional ones, seek to align any domestic standards with international standards, and advocate for others to do the same. Seek collaboration on standards that achieve broad consensus to avoid competition between potentially incompatible standards.

- **Arrange bilateral and multilateral exchanges among experts from different countries.** These exchanges would cover public and private sector AI standards needs and how they are using existing standards and guidelines. Interactions such as these would promote greater understanding between standards developers and users, including government representatives, about global needs, priorities, and experiences. Expert-to-expert exchanges can be leveraged to encourage contributions from low- and middle-income countries and strengthen mutual understanding of the benefits and limitations of standardization.

- **Continue seeking to maximize alignment between frameworks and their points of intersection but focus on standardization where possible.** While "crosswalks" between AI standards and frameworks,[33] including the NIST AI RMF, are helpful, international consensus standards have advantages over crosswalks. They tend to be more efficient, durable, and internationally acceptable than multiple frameworks and crosswalks. That said, international consensus standards also are typically much slower-moving. The fast pace of AI and dearth of international standards work on some key AI topics leads to multiple national and regional approaches. Where possible, global collaboration efforts would be most productive if focused on identifying shared ideas and taking them into the standardization process on a faster timescale. In the meantime, crosswalks will continue to add value.
- **Explore the relationship between standards and open-source software (OSS) in the context of AI systems.** OSS development is a significant venue for technical collaboration and sometimes consensus-building, including for AI. However, OSS development processes differ significantly from those of traditional documentary standards. OSS packages, including those that underpin or incorporate AI systems and models, can also implement standards. More work is needed to understand the relationship between standards and OSS in the context of AI systems.

High-priority implementation actions specific to the U.S. government

- Work with allies and partners to articulate shared principles for AI standards in multilateral diplomatic outputs. (State, Commerce)
- Build standards discussions, prioritizing international standards development and adoption, into bilateral engagements on AI policy and bilateral or multilateral collaborations on scientific research, including international partnerships formed with the U.S. AI Safety Institute within NIST/DOC. Also incorporate discussions with the local private sector (e.g., via online meetings). (State, NSF, ITA)
- Leverage or refresh existing diplomatic engagements on AI standards to promote deep exchanges between technical experts. (State, NIST)
- Expand on successful examples of coordination of U.S. government agencies on international standards engagement, such as the coordination between the Department of Commerce's International Trade Administration and NIST via standards attachés and the Department of State and NIST on translating, standards training, and professional exchange programs. (State, Commerce, USAID)
- Strengthen communication about domestic progress on foundational technical work underlying and supporting AI standards via diplomatic channels. (State)

[33] https://www.nist.gov/itl/ai-risk-management-framework/crosswalks-nist-artificial-intelligence-risk-management-framework

Appendix A. Standards in Relation to AI

A.1. What are standards and why are they important?

In this plan, "standards" and "technical standards" both refer to documentary standards, defined by International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) Guide 2:2004 Standardization and related activities—General vocabulary as "a document, established by consensus and approved by a recognized body, that provides for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context." Standards can be developed in many types of organizations that span a wide range of formality, structure, subject matter, and approach.

Widespread use of standards can facilitate technological advancement and adoption by providing common foundations from which to build. They can make products and services more interoperable, avoid technical barriers to trade, and facilitate an efficient marketplace. Standards can also make products and services safer and more trustworthy by establishing well-vetted consensus practices. In AI, standards that articulate requirements, specifications, guidelines, or characteristics can help to ensure that AI technologies and systems meet critical objectives for functionality, interoperability, and trustworthiness—and that they perform reliably and safely.

For some technologies and domains, including AI, standards are important not just for technical interoperability but also for regulatory interoperability. Standards define shared concepts, metrics, and practices that governments can refer to and build on as they develop policies and regulations. If different jurisdictions can standardize on the same building blocks, then even if regulatory environments are not fully aligned, it is at least easier for market participants to move smoothly between markets. Global cooperation and coordination on AI standards will be critical for defining a consistent or at least interoperable set of "rules of the road."

As noted in the introduction, standards are typically adopted and implemented on a voluntary basis. Voluntary compliance and conformity regimes can bring significant benefits. First, they can adapt more easily and quickly as technology changes or new and better practices emerge. Voluntary standards, particularly those that are performance- and outcome-based, can also be far more flexible; because they do not depend on compulsory compliance mechanisms, they can leave more freedom to adopters to account for their own contexts. This flexibility can advance innovation and help standardized practices be as widely applicable as possible.

A.2. How are standards developed?

The U.S. standards system differs significantly from the government-driven standards systems in many other countries and regions. Hundreds of standards developing organizations (SDOs)—most of which do not develop AI standards—are domiciled within the United States. These organizations provide the infrastructure for the preparation of standards documents. U.S. government personnel participate in SDO activities along with representatives from industry, academia, and other organizations and consumers. It is important to emphasize that these SDOs are primarily private-sector organizations, and

that the Federal government is simply one of many stakeholders and participants. The American National Standards Institute (ANSI) United States Standards Strategy, elaborated through a private-public partnership in 2005 and updated by ANSI every five years, outlines the contribution of private-sector led standards development to overall competition and innovation in the U.S. economy. The Strategy sets a strategic vision to support U.S. competitiveness, innovation, health and safety, and global trade, guiding how the United States develops and uses standards, and participates in the international standards development process.

In many other standards systems, the government plays a larger role in standards development-related activities. In such cases, these governments have more leverage to use standards as tools for competition, innovation policy, and geopolitical influence. While U.S. Government agencies possess certain responsibilities related to standards, such as in the use of standards in regulation, procurement, or other activities, there is a much greater reliance in the United States than in the European Union or China on obtaining input from industry groups, consumers, and other interested parties in making decisions related to the technical content of standards and on allowing the private sector to drive standards development.

By contrast, other governments have instituted top-down standards systems, which may involve governmental direction to stakeholders to develop particular standards, the provision of funding to national delegations, and hosting meetings.

The formal process of developing a standard tends to be relatively long, and the full process of standardization extends significantly further, both before and after formal development, review, and approval. Before a standard can even be proposed, there is often a need for significant foundational scientific work, such as technical research and pilot experiments, to explore what rules, guidelines, characteristics, or activities ought to be standardized. The standards development process itself builds on that foundational work, incorporating additional views and the need to establish consensus. The phase following standardization is about adoption: potential users of a standard may need significant additional tools to be able to implement it, including datasets, benchmarks, reference implementations, implementation guidance, verification and validation tools, and conformity assessment procedures.

To be useful, standards need to be timely. If standards development is attempted before foundational work has yielded a critical mass of technical understanding, the resulting standard may prove ill-founded or even counterproductive. Voluntary standards developed in this manner will likely fail to be adopted, and, if they are adopted (or mandated), they can impede innovation while providing little or no countervailing benefit. However, a standard is not useful if it arrives after the technologies have already moved on. Standards can also fail to gain market acceptance if they are produced late enough that market incumbents have built up infrastructure and market power, which can also hinder innovation. AI technologies are so fast-moving that existing standardization processes may well struggle to keep up. The tradeoff between timeliness and rigor can sometimes be addressed by pursuing products such as guidelines and technical reports, which can be developed faster but are less prescriptive and typically represent less thorough consensus. These products could be the basis of future standards.

Most SDOs do not track the impact of their standards once completed, though many would like to be able to. SDOs may be able to track downloads or sales of standards documents, and national standards

bodies may arrange with the SDO to publish a standard as a national standard, in which case the SDO would be aware of the standard's national adoption. However, these are at best loose proxies for how extensively standards are being implemented and how well they are meeting users' needs.

Broadly, AI standards can address *horizontal* (cross-sector) or *vertical* (sector-specific) needs. Horizontal AI standards can be used across many applications and industries. Standards developed for specific applications areas such as healthcare or transportation are vertical standards. Developers of horizontal standards often seek to establish collaborative working relationships (e.g., liaisons) with sector-specific (vertical) standards developers. These liaisons foster cooperation, establish or reinforce boundaries, and help to ensure that horizontal standards are relevant to other AI standardization efforts and vice versa.

A.3. How do AI standards differ from other technical standards?

Unlike in some other technical fields such as communications technologies, where inter-system technical compatibility is vital, AI technologies often do not depend on standardized interfaces and protocols to work. Accordingly, standards in AI have tended to serve more of a "trailing edge" function. As AI stakeholders consider technologies that are already gaining traction, standards help them to:

- Converge on foundational concepts and terminology, essential for interoperability of technical approaches and evaluation methodologies as well as productive policy conversations;
- Set norms for governance and accountability processes (e.g., for risk management and trustworthiness), which raises the bar for developers' and deployers' practices and helps AI actors, especially lower-resourced ones, innovate with confidence; and
- Measure and evaluate their systems in comparable ways, facilitating confidence by developers, deployers, users, and affected parties in the usefulness and trustworthiness of AI systems.

Many of these areas of standardization must account for or directly address interactions between AI systems and people and institutions. In other words, AI systems and their impacts are inherently *sociotechnical*, hinging on complex interactions between AI systems and humans. The standards addressing these systems, such as for institutional governance practices or processes for measuring impact, are therefore often sociotechnical as well, addressing these interactions head-on.

Because AI standards are generally more detailed than the high-level AI policy principles discussed in multilateral settings such as the OECD or the G7, they can provide actionable guidance for developers, project managers, senior leaders, and other hands-on AI actors on how to implement high-level principles. Given the prevalence[34] of such frameworks of principles, AI standards take on extra societal significance beyond their usual role in facilitating trade and technological innovation.

This is not to say that standards have no role in defining interfaces and protocols for interactions with or between AI systems. It remains possible that standards in this direction (e.g., for standardized dataset formats, model access APIs, or large language model plugins reusable across models) will yet be identified by standards stakeholders as important needs.

[34] https://cyber.harvard.edu/publication/2020/principled-ai

Appendix B. The Current Landscape of AI Standardization

To paint the backdrop for this plan's objectives and engagement actions, this section briefly overviews SDO efforts to date on AI.

B.1. Horizontal standards: SDOs and topics

Several SDOs have been particularly active in developing horizontal (i.e., sector-independent) AI standards. The subsections below elaborate on individual SDOs and their projects. This list is based on public comments, and may not be fully exhaustive.

A.1.1. ISO/IEC JTC 1

ISO/IEC JTC 1 SC 42 Artificial Intelligence is a subcommittee (SC) of the ISO/IEC Joint Technical Committee (JTC) 1. The purpose of this subcommittee is to develop technical standards and guidelines for AI and its associated technologies.[35] The subcommittee focuses largely on horizontal standards, with a particular emphasis on core concepts and practices.

Many of the documents produced by SC 42 focus on topics around concepts and governance. Topics include a management system standard, impact assessment, the data lifecycle, AI systems software quality, requirements for audit and certification, and risk management guidance. The committee has also produced some pre-standardization work in the form of Technical Reports (TRs), which provide general overview and discussion. TR topical areas include functional safety, ethical and societal concerns, machine learning (ML) computing devices, and a review of AI algorithms and system characteristics. These documents represent a consensus of conceptual thought and inform future standardization work, though not all have led directly to operationalizable standards.

Relatively few of SC 42's standards projects (3 of 28 published projects and 4 of 33 in-progress projects, as of June 2024) have been measurement-focused. Measurement topics covered are classification performance, neural network robustness, data quality, evaluating natural language processing systems, and benchmarking quality characteristics. None address monitoring and measuring societal outcomes and impacts of deployed AI systems.

Other subcommittees of ISO/IEC JTC 1 have also produced a few AI-focused work items, such as SC 27 on cybersecurity and SC 7 on software engineering.

SC 42's published projects as of June 2024 include:

Project Identifier	Project Title
ISO/IEC 38507:2022	Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

[35] https://jtc1info.org/wp-content/uploads/2023/06/01_01_Overview_ISO_IEC_AI_for_ISO_IEC_AI_Workshop_0623.pdf

Project Identifier	Project Title
ISO/IEC 22989:2022	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
ISO/IEC 23053:2022	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC 42001:2023	Information Technology — Artificial intelligence — Management system
ISO/IEC 24668:2022	Information technology — Artificial intelligence — Process management framework for big data analytics
ISO/IEC TR 24027:2021	Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
ISO/IEC 24029-2:2023	Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
ISO/IEC 23894:2023	Information technology — Artificial intelligence — Guidance on risk management
ISO/IEC TR 24368:2022	Information technology — Artificial intelligence — Overview of ethical and societal concerns
ISO/IEC TR 5469:2024	Artificial intelligence — Functional safety and AI systems
ISO/IEC 25059:2023	Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
ISO/IEC TR 24030:2024	Information technology - Artificial intelligence (AI) - Use cases
ISO/IEC 5338:2023	Information technology - Artificial intelligence - AI system life cycle processes
ISO/IEC 5339:2024	Information Technology - Artificial Intelligence - Guidelines for AI applications
ISO/IEC TR 24372:2021	Information technology - Artificial intelligence (AI) - Overview of computational approaches for AI systems
ISO/IEC TS 4213:2022	Information technology - Artificial intelligence - Assessment of machine learning classification performance
ISO/IEC 5392:2024	Information technology - Artificial intelligence - Reference architecture of knowledge engineering
ISO/IEC TS 8200:2024	Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems

Project Identifier	Project Title
ISO/IEC TS 25058:2024	Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
ISO/IEC 8183:2023	Information technology — Artificial intelligence — Data life cycle framework
ISO/IEC TR 17903:2024	Information technology — Artificial intelligence — Overview of machine learning computing devices
ISO/IEC 20546:2019	Information technology — Big data — Overview and vocabulary
ISO/IEC TR 20547-1:2020	Information technology — Big data reference architecture — Part 1: Framework and application process
ISO/IEC TR 20547-2:2018	Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
ISO/IEC 20547-3:2020	Information technology — Big data reference architecture — Part 3: Reference architecture
ISO/IEC TR 20547-5:2018	Information technology — Big data reference architecture — Part 5: Standards roadmap
ISO/IEC TR 24028:2020	Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
ISO/IEC TR 24029-1:2021	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview

SC 42 has the following additional projects under development as of June 2024:

Project Identifier	Project Title
ISO/IEC 5259-1	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples
ISO/IEC FDIS 5259-2	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures
ISO/IEC 5259-3	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines
ISO/IEC 5259-4	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework
ISO/IEC DIS 5259-5	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework

Project Identifier	Project Title
ISO/IEC CD TR 5259-6	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality
ISO/IEC CD TS 6254	Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems
ISO/IEC DTS 12791.2	Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
ISO/IEC DIS 12792	Information technology — Artificial intelligence — Transparency taxonomy of AI systems
ISO/IEC AWI TS 17847	Information technology — Artificial intelligence — Verification and validation analysis of AI systems
ISO/IEC AWI TR 18988	Artificial intelligence — Application of AI technologies in health informatics
ISO/IEC CD TR 20226	Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems
ISO/IEC CD TR 21221	Information technology – Artificial intelligence – Beneficial AI systems
ISO/IEC AWI TS 22440-1	Artificial intelligence — Functional safety and AI systems — Part 1: Requirements
ISO/IEC AWI TS 22440-2	Artificial intelligence — Functional safety and AI systems — Part 2: Guidance
ISO/IEC AWI TS 22440-3	Artificial intelligence — Functional safety and AI systems — Part 3: Examples of application
ISO/IEC AWI TS 22443	Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations
ISO/IEC 22989:2022/AWI Amd 1	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology — Amendment 1
ISO/IEC 23053:2022/AWI Amd 1	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) — Amendment 1
ISO/IEC AWI TR 23281	Artificial intelligence — Overview of AI tasks and functionalities related to natural language processing
ISO/IEC AWI 23282	Artificial Intelligence — Evaluation methods for accurate natural language processing systems

Project Identifier	Project Title
ISO/IEC AWI 24029-3	Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 3: Methodology for the use of statistical methods
ISO/IEC AWI 24970	Artificial intelligence — AI system logging
ISO/IEC AWI 25029	Artificial intelligence — AI-enhanced nudging
ISO/IEC AWI 25059	Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
ISO/IEC AWI TS 29119-11	Software and systems engineering — Software testing — Part 11: Testing of AI systems
ISO/IEC DIS 42005	Information technology — Artificial intelligence — AI system impact assessment
ISO/IEC DIS 42006	Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems
ISO/IEC AWI 42102	Information technology — Artificial intelligence — Taxonomy of AI system methods and capabilities
ISO/IEC AWI TR 42103	Information technology — Artificial intelligence — Overview of synthetic data in the context of AI systems
ISO/IEC AWI 42105	Information technology — Artificial intelligence — Guidance for human oversight of AI systems
ISO/IEC AWI TR 42106	Information technology — Artificial intelligence — Overview of differentiated benchmarking of AI system quality characteristics
ISO/IEC AWI TR 42109	Information technology — Artificial intelligence — Use cases of human-machine teaming

Beyond SC 42, JTC1 is developing the following standards from subcommittee 27 on cybersecurity:

Project Identifier	Project Title
ISO/IEC CD 27090	Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems
ISO/IEC WD 27091.2	Cybersecurity and Privacy — Artificial Intelligence — Privacy protection

A.1.2. CEN-CENELEC

CEN, the European Committee for Standardization, is an association that brings together the National Standardization Bodies of 34 European countries. CEN provides a platform for the development of European Standards and other technical documents in relation to various kinds of products, materials, services, and processes.[36] CENELEC, the European Committee for Electrotechnical Standardization, plays a similar role for electrotechnical standardization. The openness of CEN and CENELEC processes to participation by U.S. stakeholders and other stakeholders from outside Europe is sometimes limited.

In 2020, CEN and CENELEC established a new JTC 21 "Artificial Intelligence." CEN-CLC/JTC 21 identifies and adopts international standards already available or under development from other organizations like ISO/IEC JTC 1 and its subcommittees, such as SC 42. Furthermore, CEN-CLC/JTC 21 focuses on producing standardization deliverables that address European market and societal needs, as well as underpinning European Union (EU) legislation, policies, principles, and values.[37]

CEN/CLC JTC 21 was formed partly in response to the European Commission white paper that initiated the creation of the EU AI Act. The committee has accepted a standardization request from the Commission to fulfill the standardization needs of the AI Act, which will drive much of its work in the coming months. The committee is expected to produce "harmonized standards" (standards developed for the purpose of being referenced by regulation). These standards will be voluntary, but, nonetheless, will have legal implications: Referenced EU harmonized standards carry a presumption of conformity, making compliance with these standards the recommended but not the only method to meet regulatory requirements. Per a 2016 ruling from the European Court of Justice, such standards form part of EU law, as they have legal effects.

A of June 2024, CEN/CLC JTC 21 has not published any standards of its own, although it has adopted some ISO/IEC standards. The standardization request from the Commission, which is expected to drive future work, includes standards for risk management systems, dataset quality and governance, record keeping, transparency, human oversight, accuracy specifications, robustness specifications, cybersecurity specifications, quality management systems, and conformity assessment, which were all requested for delivery by April 2025. Projects under development that are not shared with ISO/IEC include:

Project Identifier	Project Title
FprCEN/CLC/TR 18115	Data governance and quality for AI within the European context
prCEN/CLC/TR 17894	Artificial Intelligence Conformity Assessment
prCEN/CLC/TR XXX	Impact assessment in the context of the EU Fundamental Rights
prCEN/CLC/TR XXX	AI Risks - Check List for AI Risks Management

[36] https://www.cencenelec.eu/about-cen/

[37] https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

Project Identifier	Project Title
prCEN/CLC/TR XXX	Environmentally sustainable Artificial Intelligence
prCEN/CLC/TR XXXX	Artificial Intelligence - Overview of AI tasks and functionalities related to natural language processing
prEN XXX	AI trustworthiness framework
prEN XXX	AI tasks and evaluation methods of computer vision systems
prEN XXX	AI Risk Management
prEN XXXXX	AI-enhanced nudging

A.1.3. IEEE

"IEEE Standards Association (IEEE SA) is a consensus building organization that nurtures, develops, and advances global technologies, through IEEE. It brings together a broad range of individuals and organizations from a wide range of technical and geographic points of origin to facilitate standards development and standards-related collaboration."[38]

Starting in 2016, the IEEE P7000 series of standards projects addresses specific issues at the intersection of technological and ethical considerations for AI. The AI Standards Committee is responsible for standards that enable the governance and practice of AI as related to computational approaches to machine learning, algorithms, and related data usage.[39]

Other topics addressed by IEEE's AI standards include organizational governance, explainable AI, federated learning, autonomous system verification, and technical details such as data attributes and formats.

Project Identifier	Project Title
IEEE P2863	Recommended Practice for Organizational Governance of Artificial Intelligence
IEEE P2894	IEEE Draft Guide for an Architectural Framework for Explainable Artificial Intelligence
IEEE P2976	Standard for XAI - eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design
IEEE P3123	Standard for Artificial Intelligence and Machine Learning (AI/ML) Terminology and Data Formats
IEEE P2817	IEEE Draft Standards Project Guide for Verification of Autonomous Systems

[38] https://standards.ieee.org/about/
[39] https://sagroups.ieee.org/ai-sc/

Project Identifier	Project Title
IEEE P2986	Recommended Practice for Privacy and Security for Federated Machine Learning
IEEE P2975	Standard for Industrial Artificial Intelligence (AI) Data Attributes

A.1.4. ITU

The International Telecommunication Union (ITU) is the United Nations specialized agency for information and communication technologies (ICTs).[40] The Study Groups of ITU's Telecommunication Standardization Sector (ITU-T) assemble experts from around the world to develop international standards known as ITU-T Recommendations that act as defining elements in the global infrastructure of ICTs.[41]

ITU-T's work on AI standards relates primarily to use of AI for increasing efficiency of telecommunication/ICT networks and systems. ITU also engages in activities related to information sharing and capacity building on AI for sustainable development as related to telecommunications/ICTs. ITU collaborates with other United Nations agencies and stakeholders to support realization of the benefits of AI use cases for sustainable development, such as through its AI for Good platform, which includes information exchanges and workshops related to AI standards.

Project Identifier	Project Title
F.ADT4MM	Requirements and framework of AI-based detection technologies for 5G multimedia messages
F.ACIP-GA	Technical specifications for AI cloud platform: general architecture
F.ACIP-MD	Technical specification for AI cloud platform: AI model development
F.AI-CPP	Technical specification for AI cloud platform: performance
F.AI-DMPC	Technical framework for deep neural network model partition and collaborative execution
F.AI-FASD	Framework for audio structuralizing based on deep neural network
F.AI-ILICSS	Technical requirements and evaluation methods of intelligent levels of intelligent customer service system
F.AI-ISD	Requirements for intelligent surface-defect detection service in industrial production line

[40] https://www.itu.int/en/about/Pages/default.aspx
[41] https://www.itu.int/en/ITU-T/about/Pages/default.aspx

Project Identifier	Project Title
F.AI-MKGDS	Requirements for the construction of multimedia knowledge graph database structure based on artificial intelligence
F.AI-MVSLWS (ex F.AI-VDSLWS)	Requirements for artificial intelligence based machine vision service in smart logistics warehouse system
F.AI-RSRSreqs	Requirements for real-time super-resolution service based on artificial intelligence
F.AI-SF	Requirements for smart factory based on artificial intelligence
F.FDIS	Requirements and framework for feature-based distributed intelligent systems
F.FML-TS-FR	Requirement and framework of trustworthy federated machine learning based service
F.ML-ICSMIReqs	Requirements and framework for intelligent crowd sensing multimedia interaction based on deep learning
F.REAIOCR	Requirements and evaluation methods for AI-based optical character recognition service
F.SCAI	Requirements for smart class based on artificial intelligence
F.TCEF-FML	Trusted contribution evaluation framework on federated machine learning services
Y.3181 (ex Y.ML-IMT2020-SANDBOX)	Architectural framework for Machine Learning Sandbox in future networks including IMT-2020
Y.3182 (ex Y.ML-IMT2020-E2E-MGMT)	Machine learning based end-to-end multi-domain network slice management and orchestration
Y.CNAO	Requirements and functional framework for Customer-oriented Network Quality Auto Optimization with Artificial Intelligence
Y.IMT2020-DJLML	Requirements and framework for distributed joint learning to enable machine learning in future networks including IMT-2020
Y.IMT2020-AINDO-req-frame	Requirements and framework for AI-based network design optimization in future networks including IMT-2020
Y.ML-IMT2020-VNS	Framework for network slicing management enabled by machine learning including input from verticals

Project Identifier	Project Title
Y.ML-IMT2020-MLFO	Requirements and architecture for machine learning function orchestrator
Q.AIS-SRA	Signalling requirements and architecture to support AI based vertical services in future network, IMT2020 and beyond

A.1.5. Consumer Technology Association

The Consumer Technology (CTA) oversees the development of standards by experts from across the consumer technology industry, including manufacturers, service providers, regulators and other industry leaders. These 1,500 tech pioneers — from engineers to doctors and scientists — help produce specifications that define how products work and the ways consumers interact with them.[42]

In addition to its sector-specific standards (see below), CTA has developed several horizontal AI standards as of June 2024:

Project Identifier	Project Title
CTA-2089-A	Definitions and Characteristics of Artificial Intelligence
CTA-5203	Cybersecurity Threats and Security Controls for Machine Learning Based Systems
ANSI/CTA-2096	Guidelines for Developing Trustworthy Artificial Intelligence Systems
CTA-5200	What is Artificial Intelligence?

A.1.6. Non-AI-specific horizontal standards highly relevant to AI

Several standards that are not AI-specific are nonetheless highly relevant to AI and are sometimes applied to AI systems. These include:

Project Identifier	Project Title
IEEE 7000-2021	IEEE Standard Model Process for Addressing Ethical Concerns during System Design
ISO/IEC Guide 51:2014	Safety aspects — Guidelines for their inclusion in standards

[42] https://shop.cta.tech/collections/standards/artificial-intelligence

A.1.7. Sectoral standards: SDOs and topics

In industries that are coming to rely heavily on AI, sector-specific standards projects have also begun to emerge:

- For **AI in aeronautical systems**, SAE International, a global association of engineers and related technical experts in the aerospace, automotive and commercial vehicle industries,[43] is developing standards products on foundational concepts and certification processes. EUROCAE, a European non-profit that develops standards for European civil aviation, also has a working group on AI.
- For **AI in ground vehicle applications**, SAE also has standards in progress.
- For **AI in healthcare** (where regulatory requirements can make standards especially critical, as in the recently finalized HTI-1 rule[44]), the Consumer Technology Association has published a variety of standards on topics such as foundational definitions and characteristics, bias, and data management and is working on more. ISO and IEC have also published healthcare-related standards addressing topics such as medical electrical equipment employing autonomy, risk management, and quality management systems for AI/ML in medical devices. IEEE has several standards related to AI in healthcare. The Association for the Advancement of Medical Instrumentation has published a technical report on risk management in AI for medical devices. Finally, Integrating the Healthcare Enterprise has published "profile" standards that specify the application of base standards to health care use cases.
- For **AI in finance**, X9, an ANSI-accredited developer of financial services standards, has started an AI study group aiming to identify areas where standards are or could be needed to safeguard financial, infrastructure and user data,[45] and two standing working groups assigned to work on AI issues.
- For **AI in biotechnology**, American Type Culture Collection has begun work on standards and is looking to develop authenticated reference data for use in training AI models.

Other SDOs that have indicated that they have sectoral AI work underway or under consideration include the Robotics Industries Association, CSA Group, Alliance for Telecommunications Solutions, National Information Standards Organization, National Council for Prescription Drug Programs, American Society for Nondestructive Testing, and the Instrument Society of America.

---

[43] https://www.sae.org/
[44] https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program
[45] https://x9.org/aistudygroup/

B.2. Participation in AI standards development

During consultations, some parties noted that the majority of participants in AI standards bodies are from industry. Large, well-resourced technology companies were cited as the participants most aware of and active in standards development, while relatively few SMEs have been participating. Startups may be aware of standards-setting work, but they do not always have the resources to effectively participate.

Many commenters also noted that civil society and academia have historically not been well-represented in standards development work, including on AI. Some commenters attributed this to confusion about what standards are, what they can and cannot do, and when and how they are developed. It was also suggested that these entities tend not to recognize how standards development might contribute to their goals, and that they find procedures for participating opaque.

Low- and middle-income countries seem to be particularly missing from AI standards, as reported with great concern by numerous commenters.