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

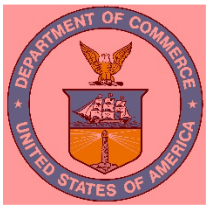
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

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.

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		