Check for updates

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.*

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 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 1
2. Overview of Risks Unique to or Exacerbated by GAI 2
2.1. CBRN Information or Capabilities 5
2.2. Confabulation 6
2.3. Dangerous, Violent, or Hateful Content 6
2.4. Data Privacy 7
2.5. Environmental Impacts 8
2.6. Harmful Bias and Homogenization 8
2.7. Human-AI Configuration 9
2.8. Information Integrity 9
2.9. Information Security 10
2.10. Intellectual Property 11
2.11. Obscene, Degrading, and/or Abusive Content 11
2.12. Value Chain and Component Integration 12
3. Suggested Actions to Manage GAI Risks 12
Appendix A. Primary GAI Considerations 47
Appendix B. References 54

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	**Suggested 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