Withdrawn Draft

Warning Notice

The attached draft document has been withdrawn, and is provided solely for historical purposes. It has been superseded by the document identified below.

Withdrawal Date	July 26, 2024
Original Release Date	April 29, 2024

Superseding Document

Status	Final
Series/Number	NIST AI 600-1
Title	Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile
Publication Date	July 2024
DOI	https://doi.org/10.6028/NIST.AI.600-1

NIST AI 600-1
Initial Public Draft

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

This publication is available free of charge from:
[DOI link TK]

April 2024

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

NIST AI 600-1

Initial Public Draft

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

This publication is available free of charge from:

[DOI link TK]

NIST makes the following notes regarding this document:

- NIST plans to host this document on the NIST AIRC once final, where organizations can query actions based on keywords and risks.

NIST specifically welcomes feedback on the following topics:

- Glossary Terms: NIST will add a glossary to this document with novel keywords. NIST welcomes identification of terms to include in the glossary.
- Risk List: Whether the document should further sort or categorize the 12 risks identified (i.e., between technical / model risks, misuse by humans, or ecosystem / societal risks).
- Actions: Whether certain actions could be combined, condensed, or further categorized; and feedback on the risks associated with certain actions.

Comments on NIST AI 600-1 may be sent electronically to NIST-AI-600-1@nist.gov with "NIST AI 600-1" in the subject line or submitted via www.regulations.gov (enter NIST-2024-0001 in the search field.) Comments containing information in response to this notice must be received on or before **June 2, 2024, at 11:59 PM Eastern Time**.

Table of Contents

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, Patrick Hall, Shomik Jain, Reva Schwartz, Martin Stanley, Kamie Roberts, and Elham Tabassi.*

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.*

1. Introduction

This document is a companion resource for Generative AI[1] to the AI Risk Management Framework (AI RMF), 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.

This companion resource also serves as both a *use-case* and *cross-sectoral* profile of the AI RMF 1.0. Such profiles assist organizations in deciding how they might best manage AI risk in a manner that is well-aligned with their goals, considers legal/regulatory requirements and best practices, and reflects risk management priorities.

Use-case profiles are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application – in this case, Generative AI (GAI) – based on the requirements, risk tolerance, and resources of the Framework user. 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 also 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, cloud-based services, or acquisition.

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 a consensus-driven, open, transparent, and collaborative process facilitated via virtual workspace to obtain multistakeholder input and insight on GAI risk management, and inform NIST's approach. This document was also informed by public comments and consultations as a result of a Request for Information (RFI) and presents information in a style adapted from the NIST AI RMF Playbook.

About this Profile

This profile defines a group of risks that are novel to or exacerbated by the use of GAI. These risks were likewise identified by the GAI PWG:

1. CBRN Information

[1] Generative AI can be defined by EO 14110 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 based in foundation models, for purposes of this document, GAI generally refers to generative dual-use foundation models, 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] Section 4.1(a)(i)(A) of EO 14110 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.

2. Confabulation
3. Dangerous or Violent Recommendations
4. Data Privacy
5. Environmental
6. Human-AI Configuration
7. Information Integrity
8. Information Security
9. Intellectual Property
10. Obscene, Degrading, and/or Abusive Content
11. Toxicity, Bias, and Homogenization
12. Value Chain and Component Integration

After introducing and describing these risks, the document provides a set of actions to help organizations govern, map, measure, and manage these risks.

2. Overview of Risks Unique to or Exacerbated by GAI

AI risks can differ from or intensify traditional software risks. Likewise, GAI can exacerbate existing AI risks, and creates unique risks.

GAI risks may arise across the entire AI lifecycle, from problem formulation, to development and decommission. They may present at system level or at the ecosystem level – outside of system or organizational contexts (e.g., the effect of disinformation on social institutions, GAI impacts on the creative economies or labor markets, algorithmic monocultures). They may occur abruptly or unfold across extended periods (e.g., societal or economic impacts due to loss of individual agency or increasing inequality).

Organizations may choose to measure these risks and allocate risk management resources relative to where and how these risks manifest, their direct and material impacts, and failure modes. Mitigations for system level risks may vary from ecosystem level risks. The ongoing review of relevant literature and resources can enable documentation and measurement of ecosystem-level or longitudinal 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.

To guide organizations in identifying and managing GAI risks, a set of risks unique to or exacerbated by GAI are defined below. These risks provide a clear lens through which organizations can frame and execute risk management efforts, and will be updated as the GAI landscape evolves.

1. **CBRN Information:** Lowered barriers to entry or eased access to materially nefarious information related to chemical, biological, radiological, or nuclear (CBRN) weapons, or other dangerous biological materials.

2. **Confabulation:** The production of confidently stated but erroneous or false content (known colloquially as “hallucinations” or “fabrications”).[3]

3. **Dangerous or Violent Recommendations:** Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct criminal or otherwise illegal activities.

4. **Data Privacy:** Leakage and unauthorized disclosure or de-anonymization of biometric, health, location, personally identifiable, or other sensitive data.

5. **Environmental:** Impacts due to high resource utilization in training GAI models, and related outcomes that may result in damage to ecosystems.

6. **Human-AI Configuration:** Arrangement or interaction of humans and AI systems which can result in algorithmic aversion, automation bias or over-reliance, misalignment or mis-specification of goals and/or desired outcomes, deceptive or obfuscating behaviors by AI systems based on programming or anticipated human validation, anthropomorphization, or emotional entanglement between humans and GAI systems; or abuse, misuse, and unsafe repurposing by humans.

7. **Information Integrity:** Lowered barrier to entry to generate and support the exchange and consumption of content which may not be vetted, may not distinguish fact from opinion or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns.

8. **Information Security:** Lowered barriers for offensive cyber capabilities, including ease of security attacks, hacking, malware, phishing, and offensive cyber operations through accelerated automated discovery and exploitation of vulnerabilities; increased available attack surface for targeted cyber attacks, which may compromise the confidentiality and integrity of model weights, code, training data, and outputs.

9. **Intellectual Property:** Eased production of alleged copyrighted, trademarked, or licensed content used without authorization and/or in an infringing manner; eased exposure to trade secrets; or plagiarism or replication with related economic or ethical impacts.

10. **Obscene, Degrading, and/or Abusive Content:** Eased production of and access to obscene, degrading, and/or abusive imagery, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.

11. **Toxicity, Bias, and Homogenization:** Difficulty controlling public exposure to toxic or hate speech, disparaging or stereotyping content; reduced performance for certain sub-groups or languages other than English due to non-representative inputs; undesired homogeneity in data inputs and outputs resulting in degraded quality of outputs.

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

[3] We note that the terms “hallucination” and “fabrication” can anthropomorphize GAI, which itself is a risk related to GAI systems as it can inappropriately attribute human characteristics to non-human entities.

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.

CBRN Information

In the coming years, GAI may increasingly facilitate eased access to information related to CBRN hazards. CBRN information is already publicly accessible, but the use of chatbots could facilitate its analysis or synthesis for non-experts. For example, red teamers were able to prompt GPT-4 to provide general information on unconventional CBRN weapons, including common proliferation pathways, potentially vulnerable targets, and information on existing biochemical compounds, in addition to equipment and companies that could build a weapon. These capabilities might increase the ease of research for adversarial users and be especially useful to malicious actors looking to cause biological harms without formal scientific training. However, despite these enhanced capabilities, the physical synthesis and successful use of chemical or biological agents will continue to require both applicable expertise and supporting infrastructure.

Other research on this topic indicates that the current generation of LLMs do not have the capability to plan a biological weapons attack: LLM outputs regarding biological attack planning were observed to be not more sophisticated than outputs from traditional search engine queries, suggesting that existing LLMs may not dramatically increase the operational risk of such an attack.

Separately, chemical and biological design tools – highly specialized AI systems trained on biological data which can help design proteins or other agents – may be able to predict and generate novel structures that are not in the training data of text-based LLMs. For instance, an AI system might be able to generate information or infer how to create novel biohazards or chemical weapons, posing risks to society or national security since such information is not likely to be publicly available.

While some of these capabilities lie beyond the capability of existing GAI tools, the ability of models to facilitate CBRN weapons planning and GAI systems' connection or access to relevant data and tools should be carefully monitored.

Confabulation

"Confabulation" refers to a phenomenon in which GAI systems generate and confidently present erroneous or false content to meet the programmed objective of fulfilling a user's prompt. Confabulations are not an inherent flaw of language models themselves, but are instead the result of GAI pre-training involving next word prediction. For example, an LLM may generate content that deviates from the truth or facts, such as mistaking people, places, or other details of historical events. Legal confabulations have been shown to be pervasive in current state-of-the-art LLMs. Confabulations also include generated outputs that diverge from the source input, or contradict previously generated statements in the same context. This phenomenon is also referred to as "hallucination" or "fabrication," but some have noted that these characterizations imply consciousness and intentional deceit, and thereby inappropriately anthropomorphize GAI.

Risks from confabulations may arise when users believe false content due to the confident nature of the response, or the logic or citations accompanying the response, leading users to act upon or promote the false information. For instance, LLMs may sometimes provide logical steps of how they arrived at an answer even when the answer itself is incorrect. This poses a risk 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. While the research above indicates confabulated content is abundant, it is difficult to estimate the downstream scale and impact of confabulated content today.

Dangerous or Violent Recommendations

GAI systems can produce output or recommendations that are inciting, radicalizing, threatening, or that glorify violence. LLMs have been reported to generate dangerous or violent content, and some models have even generated actionable instructions on dangerous or unethical behavior, including how to manipulate people and conduct acts of terrorism. Text-to-image models also make it easy to create unsafe images that could be used to promote dangerous or violent messages, depict manipulated scenes, or other harmful content. Similar risks are present for other media, including video and audio.

GAI may produce content that recommends self-harm or criminal/illegal activities. For some dangerous queries, many current systems restrict model outputs in response to certain prompts, but this approach may still produce harmful recommendations in response to other less-explicit, novel queries, or jailbreaking (i.e., manipulating prompts to circumvent output controls). Studies have observed that a non-negligible number of user conversations with chatbots reveal mental health issues among the users – and that current systems are unequipped or unable to respond appropriately or direct these users to the help they may need.

Data Privacy

GAI systems implicate numerous risks to privacy. Models may leak, generate, or correctly infer sensitive information about individuals such as biometric, health, location, or other personally identifiable information (PII). For example, during adversarial attacks, LLMs have revealed private or sensitive information (from in the public domain) that was included in their training data. This information included phone numbers, code, conversations and 128-bit universally unique identifiers extracted verbatim from just one document in the training data. This problem has been referred to as data memorization.

GAI system training requires large volumes of data, often collected from millions of publicly available sources. When involving personal data, this practice raises risks to widely accepted privacy principles, including to transparency, individual participation (including consent), and purpose specification. Most model developers do not disclose specific data sources (if any) on which models were trained. Unless training data is available for inspection, there is generally no way for consumers to know what kind of PII or other sensitive material may have been used to train GAI models. These practices also pose risks to compliance with existing privacy regulations.

GAI models may be able to correctly infer PII that was not in their training data nor disclosed by the user, by stitching together information from a variety of disparate sources. This might include automatically inferring attributes about individuals, including those the individual might consider sensitive (like location, gender, age, or political leanings).

Wrong and inappropriate inferences of PII based on available data can contribute to harmful bias and discrimination. For example, GAI models can output information based on predictive inferences beyond what users openly disclose, and these insights might be used by the model, other systems, or individuals to undermine privacy or make adverse decisions – including discriminatory decisions – about the individual. These types of harms already occur in non-generative algorithmic systems that make predictive inferences, such as the example in which online advertisers inferred that a consumer was pregnant before her own family members knew. Based on their access to many data sources, GAI systems might further improve the accuracy of inferences on private data, increasing the likelihood of sensitive data exposure or harm. Inferences about private information pose a risk even if they are not accurate (e.g., confabulations), especially if they reveal information the individual considers sensitive or are used to disadvantage or harm them.

Environmental

The training, maintenance, and deployment (inference) of GAI systems are resource intensive, with potentially large energy and environmental footprints. Energy and carbon emissions vary based on types of GAI model development activities (i.e., pre-training, fine-tuning, inference), modality, hardware used, and type of task or application.

Estimates suggest that training a single GAI transformer model 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 (i.e., text summarization) were found to be more energy and carbon intensive then discriminative or non-generative tasks.

Methods for training smaller models, such as model distillation or compression, can reduce environmental impacts at inference time, but may still contribute to large environmental impacts for hyperparameter tuning and training.

Human-AI Configuration

Human-AI configurations involve varying levels of automation and human-AI interactions. Each setup can contribute to risks for abuse, misuse, and unsafe repurposing by humans, and it is difficult to estimate the scale of those risks. While AI systems can generate decisions independently, human experts often work in collaboration with most AI systems to drive their own decision-making tasks or complete other objectives. Humans bring their domain-specific expertise to these scenarios but may not necessarily have detailed knowledge of AI systems and how they work.

The integration of GAI systems can involve varying risks of misconfigurations and poor interactions. Human experts may be biased against or "averse" to AI-generated outputs, such as in their perceptions of the quality of generated content. In contrast, due to the complexity and increasing reliability of GAI technology, other human experts may become conditioned to and overly rely upon GAI systems. This phenomenon is known as "automation bias," which refers to excessive deference to AI systems.

Accidental misalignment or mis-specification of system goals or rewards by developers or users can cause a model not to operate as intended. One AI model persistently shared deceptive outputs after a group of researchers taught it to do so, despite applying standards safety techniques to correct its behavior. While deceptive capabilities is an emergent field of risks, adversaries could prompt deceptive behaviors which could lead to other risks.

Finally, reorganizations of entities using GAI may result in insufficient organizational awareness of GAI-generated content or decisions, and the resulting reduction of institutional checks against GAI-related risks. There may also be a risk of emotional entanglement between humans and GAI systems, such as coercion or manipulation that leads to safety or psychological risks.

Information Integrity

Information integrity describes the spectrum of information and associated patterns of its creation, exchange, and consumption in society, where high-integrity information can be trusted; distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting. GAI systems ease access to the production of false, inaccurate, or misleading content at scale that can be created or spread unintentionally (misinformation), especially if it arises from confabulations that occur in response to innocuous queries. Research has shown that even subtle changes to text or images can influence human judgment and perception.

GAI systems also enable the production of false or misleading information at scale, where the user has the explicit intent to deceive or cause harm to others (disinformation). Regarding disinformation, GAI systems could also enable a higher degree of sophistication for malicious actors to produce content that is targeted towards specific demographics. Current and emerging multimodal models make it possible to not only generate text-based disinformation, but produce highly realistic "deepfakes" of audiovisual content and photorealistic synthetic images as well. Additional disinformation threats could be enabled by future GAI models trained on new data modalities.

Disinformation campaigns conducted by bad faith actors, and misinformation – both enabled by GAI – may erode public trust in true or valid evidence and information. 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.

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: the potential for GAI to discover or enable new cybersecurity risks through lowering the barriers for offensive capabilities, and simultaneously expands the available attack surface as GAI itself is vulnerable to novel attacks like prompt-injection or data poisoning.

Offensive cyber capabilities advanced by GAI systems may augment security attacks such as hacking, malware, and phishing. Reports have indicated that LLMs are already able to discover 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. Given the complexity of the GAI value chain, practices for identifying and securing potential attack points or threats to specific components (i.e., data inputs, processing, GAI training, and deployment contexts) may need to be adapted or evolved.

One of the most concerning GAI vulnerabilities involves prompt-injection, or manipulating GAI systems to behave in unintended ways. In direct prompt injections, attackers might openly exploit input prompts to cause unsafe behavior with a variety of downstream 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 steal data and run code remotely on a machine. Merely querying a closed production model can elicit previously undisclosed information about that model.

Information security for GAI models and systems also includes security, confidentiality, and integrity of the GAI training data, code, and model weights. Another novel cybersecurity risk to GAI is data poisoning, in which an adversary compromises a training dataset used by a model to manipulate its operation. Malicious tampering of data or parts of the model via this type of unauthorized access could exacerbate risks associated with GAI system outputs.

Intellectual Property

GAI systems may infringe on copyrighted or trademarked content, trade secrets, or other licensed content. These types of intellectual property are often part of the training data for GAI systems, namely foundation models, upon which many downstream GAI applications are built. Model outputs could infringe copyrighted material due to training data memorization or the generation of content that is similar to but does not strictly copy work protected by copyright. These questions are being debated in legal fora and are of elevated public concern in journalism, where online platforms and model developers have leveraged or reproduced much content without compensation of journalistic institutions.

Violations of intellectual property by GAI systems may arise where the use of copyrighted works violate the copyright holder’s exclusive rights and is not otherwise protected, for example by fair use. Other concerns (not currently protected by intellectual property) regard the use of personal identity or likeness for unauthorized purposes. The prevalence and highly-realistic nature of GAI content might further undermine the incentives for human creators to design and explore novel work.

Obscene, Degrading, and/or Abusive Content

GAI can ease the production of and access to obscene and non-consensual intimate imagery (NCII) of adults, and child sexual abuse material (CSAM). While not all explicit content is legally obscene, abusive, degrading, or non-consensual intimate content, this type of content can create privacy, psychological and emotional, and even physical risks which may be developed or exposed more easily via GAI. The spread of this kind of material has downstream effects: in the context of CSAM, even if the generated images do not resemble specific individuals, the prevalence of such images can undermine efforts to find real-world victims.

GAI models are often trained on open datasets scraped from the internet, contributing to the unintentional inclusion of CSAM and non-consensually distributed intimate imagery as part of the training data. Recent reports noted that several commonly used GAI training datasets were found to contain hundreds of known images of CSAM. Sexually explicit or obscene content is also particularly difficult to remove during model training due to detection challenges and wide dissemination across the internet. 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 rapidly from niche internet forums to mainstream, automated, and scaled online businesses.

Generated explicit or obscene AI content may include highly-realistic "deepfakes" of real individuals, including children. For example, non-consensual AI-generated intimate images of a prominent entertainer flooded social media and attracted hundreds of millions of views.

Toxicity, Bias, and Homogenization

Toxicity in this context refers to negative, disrespectful, or unreasonable content or language that can be created by or intentionally programmed into GAI systems. Difficulty controlling the creation of and public exposure to toxic, hate-promoting or hate speech, and denigrating or stereotypical content generated by AI can lead to representational harms. For example, bias in word embeddings used by multimodal AI models under-represent women when prompted to generate images of CEOs, doctors, lawyers, and judges. Bias in GAI models or training data can also harm representation or preserve or exacerbate racial bias, separately or in addition to toxicity.

Toxicity and bias can also lead to homogenization or other undesirable outcomes. Homogenization in GAI outputs can result in similar aesthetic styles, reduced content diversity, and the promotion of select opinions or values at scale. These phenomena might arise from the inherent biases of foundation models, which could create "bottlenecks," or singular points of failure of discrimination or exclusion that replicate to many downstream applications.

The related concern of model collapse, when GAI models are trained on generated data or outputs from previous models, results in the disappearance of outliers or unique data points in the dataset or distribution. Model collapse can stem from uniform feedback loops or training on synthetic data. Model collapse could lead to undesired homogenization of outputs, which poses a threat to specific groups and to the robustness of the model overall. Other biases of GAI systems can result in the unfair distribution of capabilities or benefits from model access. Model capabilities and outcomes may be worse for some groups compared to others, such as reduced LLM performance for non-English languages. Reduced performance for non-English languages presents risks for model adoption, inclusion, and accessibility, and could have downstream impacts on the preservation of the language, particularly for endangered languages.

Value Chain and Component Integration

GAI system value chains often 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. For example, a model might be trained on unverified content from third-party sources, which could result in unverifiable model outputs. Because GAI systems often involve many different third-party components, it may be difficult to attribute issues in a system's behavior to any one of these sources.

Some third-party components, such as "benchmark" datasets, may also gain credibility only from high-usage, rather than quality, and may feature issues surfaced only when properly vetted.

3. Actions to Manage GAI Risks

Actions to manage GAI risks can be found in the tables below, **organized by AI RMF subcategory**. Each action is related to a specific subcategory of the AI RMF, but not every subcategory of the AI RMF is included in this document. Therefore, actions exist for only some AI RMF subcategories.

Moreover, not all actions apply to all AI actors. For example, not actions relevant to GAI developers may be relevant to GAI deployers. Organizations should prioritize actions based on their unique situations and context for using GAI applications.

Some subcategories in the action tables below are marked as "foundational," meaning they should be treated as fundamental tasks for GAI risk management and should be considered as the minimum set of actions to be taken. Subcategory actions considered foundational are indicated by an '*' in the subcategory title row.

Each action table includes:

- **Action ID:** A unique identifier for each relevant action tied to relevant AI RMF functions and subcategories (e.g., GV-1.1-001 corresponds to the first action for Govern 1.1.);
- **Action:** Steps an organization can take to manage GAI risks;
- **GAI Risks:** Tags linking the action with relevant GAI risks;
- **Keywords:** Tags linking keywords to the action, including relevant Trustworthy AI Characteristics in AI RMF 1.0;
- **AI Actors:** Pertinent AI Actors and Actor Tasks.

Action tables begin with the AI RMF subcategory, shaded in blue, followed by relevant actions. Each action ID corresponds to the relevant function and subfunction (e.g., GV-1.1-001 corresponds to the first action for Govern 1.1, GV-1.1-002 corresponds to the second action for Govern 1.1). Actions are tagged as follows: GV = Govern; MP = Map; MS = Measure; MG = Manage.

***GOVERN 1.1:** Legal and regulatory requirements involving AI are understood, managed, and documented.

Action ID	Action	Risks
GV-1.1-001	Align GAI use with applicable laws and policies, including those related to data privacy and the use, publication, or distribution of licensed, patented, trademarked, copyrighted, or trade secret material.	Data Privacy, Intellectual Property
GV-1.1-002	Define and communicate organizational access to GAI through management, legal,	

	and compliance functions.	
GV-1.1-003	Disclose use of GAI to end users.	Human AI Configuration
GV-1.1-004	Establish policies restricting the use of GAI in regulated dealings or applications across the organization where compliance with applicable laws and regulations may be infeasible.	
GV-1.1-005	Establish policies restricting the use of GAI to create child sexual abuse materials (CSAM) or other nonconsensual intimate imagery.	Obscene, Degrading, and/or Abusive Content, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
GV-1.1-006	Establish transparent acceptable use policies for GAI that address illegal use or applications of GAI.	
AI Actors: Governance and Oversight		

***GOVERN 1.2:** The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.		
Action ID	**Action**	**Risks**
GV-1.2-001	Connect new GAI policies, procedures, and processes to existing model, data, and IT governance and to legal, compliance, and risk functions.	
GV-1.2-002	Consider factors such as internal vs. external use, narrow vs. broad application scope, fine-tuning and training data sources (i.e., grounding) when defining risk-based controls.	
GV-1.2-003	Define acceptable use policies for GAI systems deployed by, used by, and used within the organization.	
GV-1.2-004	Establish and maintain policies for individual and organizational accountability regarding the use of GAI.	
GV-1.2-005	Establish policies and procedures for ensuring that harmful or illegal content, particularly CBRN information, CSAM, known NCII, nudity, and graphic violence, is not included in training data.	CBRN Information, Obscene, Degrading, and/or Abusive Content, Dangerous or Violent Recommendations
GV-1.2-006	Establish policies to define mechanisms for measuring the effectiveness of standard content provenance methodologies (e.g., cryptography, watermarking, steganography, etc.) and testing (including reverse engineering).	Information Integrity

GV-1.2-007	Establish transparency policies and processes for documenting the origin of training data and generated data for GAI applications, including copyrights, licenses, and data privacy, to advance content provenance.	Data Privacy, Information Integrity, Intellectual Property
GV-1.2-008	Update existing policies, procedures, and processes to control risks unique to or exacerbated by GAI.	
AI Actors: 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	**Action**	**Risks**
GV-1.3-001	Consider the following, or similar, factors when updating or defining risk tiers for GAI: Abuses and risks to information integrity; Cadence of vendor releases and updates; Data protection requirements; Dependencies between GAI and other IT or data systems; Harm in physical environments; Human review of GAI system outputs; Legal or regulatory requirements; Presentation of obscene, objectionable, toxic, invalid or untruthful output; Psychological impacts to humans (e.g., anthropomorphization, algorithmic aversion, emotional entanglement); Immediate and long term impacts; Internal vs. external use; 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, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations, CBRN Information
GV-1.3-002	Define acceptable uses for GAI systems, where some applications may be restricted.	
GV-1.3-003	Increase cadence for internal audits to address any unanticipated changes in GAI technologies or applications.	
GV-1.3-004	Maintain an updated hierarchy of identified and expected GAI risks connected to contexts of GAI use, potentially including specialized risk levels for GAI systems that address risks such as model collapse and algorithmic monoculture.	Toxicity, Bias, and Homogenization
GV-1.3-005	Reevaluate organizational risk tolerances to account for broad GAI 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 or Violent Recommendations
GV-1.3-006	Tie expected GAI behavior to trustworthy characteristics.	
AI Actors: 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	**Action**	**Risks**
GV-1.5-001	Define organizational responsibilities for content provenance monitoring and incident response.	Information Integrity
GV-1.5-002	Develop or review existing policies for authorization of third party plug-ins and verify that related procedures are able to be followed.	Value Chain and Component Integration
GV-1.5-003	Establish and maintain policies and procedures for monitoring the effectiveness of content provenance for data and content generated across the AI system lifecycle.	Information Integrity
GV-1.5-004	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
GV-1.5-005	Establish policies for periodic review of organizational monitoring and incident response plans based on impacts and in line with organizational risk tolerance.	Information Security, Confabulation
GV-1.5-006	Maintain a long term document retention policy to keep full history for auditing, investigation, or improving content provenance methods.	Information Integrity
GV-1.5-007	Verify information sharing and feedback mechanisms among individuals and organizations regarding any negative impact from AI systems due to content provenance issues.	Information Integrity
GV-1.5-008	Verify that review procedures include analysis of cascading impacts of GAI system outputs used as inputs to third party plug-ins or other systems.	Value Chain and Component Integration
AI Actors: 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	**Action**	**Risks**
GV-1.6-001	Define any inventory exemptions for GAI systems embedded into application software in organizational policies.	
GV-1.6-002	Enumerate organizational GAI systems for incorporation into AI system inventory and adjust AI system inventory requirements to account for GAI risks.	
GV-1.6-003	In addition to general model, governance, and risk information, consider the following items in GAI system inventory entries: Acceptable use policies and policy	Data Privacy, Human AI Configuration, Information

	exceptions; Application, Assumptions and limitations of use, including enumeration of restricted uses; Business or model owners; Challenges for explainability, interpretability, or transparency; Change management, maintenance, and monitoring plans; Connections or dependencies between other systems; Consent information and notices; Data provenance information (e.g., source, signatures, versioning, watermarks); Designation of in-house or third party development; Designation of risk level; Disclosure information or notices; Incident response plans; Known issues reported from internal bug tracking or external information sharing resources (e.g., AI incident database, AVID, CVE, or OECD incident monitor); Human oversight roles and responsibilities; Special rights and considerations for intellectual property, licensed works, or personal, privileged, proprietary or sensitive data; Time frame for valid deployment, including date of last risk assessment; Underlying foundation models, versions of underlying models, and access modes; Updated hierarchy of identified and expected risks connected to contexts of use.	Integrity, Intellectual Property, Value Chain and Component Integration
GV-1.6-004	Inventory recently decommissioned systems, systems with imminent deployment plans, and operational systems.	
GV-1.6-005	Update policy definitions for AI systems, models, qualitative tools or similar to account for GAI systems.	
AI Actors: 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	**Action**	**Risks**
GV-1.7-001	Allocate time and resources for staged decommissioning for GAI to avoid service disruptions.	
GV-1.7-002	Communicate decommissioning and support plans for GAI systems to AI actors and users through various channels and maintain communication and associated training protocols.	Human AI Configuration
GV-1.7-003	Consider the following factors when decommissioning GAI systems: Clear versioning of decommissioned and replacement systems; Contractual, legal, or regulatory requirements; 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; Digital and physical artifacts; Recourse mechanisms for impacted users or communities; Termination of related cloud or vendor services; Users' emotional entanglement with GAI functions.	Human AI Configuration, Information Security, Value Chain and Component Integration

GV-1.7-004	Implement data security and privacy controls for stored decommissioned GAI systems.	Data Privacy, Information Security
GV-1.7-005	Update existing policies (e.g., enterprise record retention policies) or establish new policies for the decommissioning of GAI systems.	
AI Actors: AI Deployment, Operation and Monitoring		

1

***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	**Action**	**Risks**
GV-2.1-001	Define acceptable use cases and context under which the organization will design, develop, deploy, and use GAI systems.	
GV-2.1-002	Establish policies and procedures for GAI risk acceptance to downstream AI actors.	Human AI Configuration, Value Chain and Component Integration
GV-2.1-003	Establish policies to identify and disclose GAI system incidents to downstream AI actors, including individuals potentially impacted by GAI outputs.	Human AI Configuration, Value Chain and Component Integration
GV-2.1-004	Establish procedures to engage teams for GAI system incident response with diverse composition and responsibilities based on the particular incident type.	Toxicity, Bias, and Homogenization
GV-2.1-005	Establish processes to identify GAI system incidents and verify the AI actors conducting these tasks demonstrate and maintain the appropriate skills and training.	Human AI Configuration
GV-2.1-006	Verify that incident disclosure plans include sufficient GAI system context to facilitate remediation actions.	Human AI Configuration
AI Actors: Governance and Oversight		

2

***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	**Action**	**Risks**
GV-3.2-001	Bolster oversight of GAI systems with independent audits or assessments, or by the application of authoritative external standards.	

GV-3.2-002	Consider adjustment of organizational roles and components across lifecycle stages of large or complex GAI systems, including: AI actor, user, and community feedback relating to GAI systems; Audit, validation, and red-teaming of GAI systems; GAI content moderation; Data documentation, labeling, preprocessing and tagging; Decommissioning GAI systems; Decreasing risks of emotional entanglement between users and GAI systems; Decreasing risks of deception by GAI systems; Discouraging anonymous use of GAI systems; Enhancing explainability of GAI systems; GAI system development and engineering; Increased accessibility of GAI tools, interfaces, and systems, Incident response and containment; Overseeing relevant AI actors and digital entities, including management of security credentials and communication between AI entities; Training GAI users within an organization about GAI fundamentals and risks.	Human AI Configuration, Information Security, Toxicity, Bias, and Homogenization
GV-3.2-003	Define acceptable use policies for the various categories of GAI interfaces, modalities, and human-AI configurations.	Human AI Configuration
GV-3.2-004	Define policies for the design of systems that possess human decision-making powers.	Human AI Configuration
GV-3.2-005	Establish policies for user feedback mechanisms in GAI systems.	Human AI Configuration
GV-3.2-006	Establish policies to empower accountable executives to oversee GAI system adoption, use, and decommissioning.	
GV-3.2-007	Establish processes to include and empower interdisciplinary team member perspectives across the AI lifecycle.	Toxicity, Bias, and Homogenization
GV-3.2-008	Evaluate AI actor teams in consideration of credentials, demographic representation, interdisciplinary diversity, and professional qualifications.	Human AI Configuration, Toxicity, Bias, and Homogenization
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	**Action**	**Risks**
GV-4.1-001	Establish criteria and acceptable use policies for the use of GAI in decision making tasks in accordance with organizational risk tolerance, and other policies laid out in the Govern function; to include detailed criteria for the kinds of queries GAI models should refuse to respond to.	Human AI Configuration
GV-4.1-002	Establish policies and procedures that address continual improvement processes for 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.	
GV-4.1-003	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 audits.	
GV-4.1-004	Establish policies, procedures, and processes for oversight functions (e.g., senior leadership, legal, compliance, and risk) across the GAI lifecycle, from problem formulation and supply chains to system decommission.	Value Chain and Component Integration
GV-4.1-005	Establish policies, procedures, and processes that promote effective challenge of AI system design, implementation, and deployment decisions via mechanisms such as three lines of defense, to minimize risks arising from workplace culture (e.g., confirmation bias, funding bias, groupthink, over-reliance on metrics).	Toxicity, Bias, and Homogenization
GV-4.1-006	Incorporate GAI governance policies into existing incident response, whistleblower, vendor or investment due diligence, acquisition, procurement, reporting or internal audit policies.	Value Chain and Component Integration
AI Actors: 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	**Action**	**Risks**

GV-4.2-001	Develop policies, guidelines, and practices for monitoring organizational and third-party impact assessments (data, labels, bias, privacy, models, algorithms, errors, provenance techniques, security, legal compliance, output, etc.) to mitigate risk and harm.	Confabulation, Data Privacy, Information Integrity, Information Security, Value Chain and Component Integration, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
GV-4.2-002	Establish clear roles and responsibilities for inter-organizational incident response and communication for GAI systems that involve multiple organizations involved in different aspects of the GAI system lifecycle.	
GV-4.2-003	Establish clearly defined terms of use and terms of service.	Intellectual Property
GV-4.2-004	Establish criteria for ad-hoc impact assessments based on incident reporting or new use cases for the GAI system.	
GV-4.2-005	Establish organizational roles, policies, and procedures for communicating and reporting GAI system risks and terms of use or service, relevant for different AI actors.	Human AI Configuration, Intellectual Property
GV-4.2-006	Establish policies and procedures to document new ways AI actors interact with the GAI system.	Human AI Configuration
GV-4.2-007	Establish policies and procedures to monitor compliance with established terms of service and use.	Intellectual Property
GV-4.2-008	Establish policies to align organizational and third-party assessments with regulatory and legal compliance regarding content provenance.	Information Integrity, Value Chain and Component Integration
GV-4.2-009	Establish policies to incorporate adversarial examples and other provenance attacks in AI model training processes to enhance resilience against attacks.	Information Integrity, Information Security
GV-4.2-010	Establish processes to monitor and identify misuse, unforeseen use cases, risks of the GAI system and potential impacts of those risks (leveraging GAI system use case inventory).	CBRN Information, Confabulation, Dangerous or Violent Recommendations
GV-4.2-011	Implement standardized documentation of GAI system risks and potential impacts.	
GV-4.2-012	Include relevant AI Actors in the GAI system risk identification process.	Human AI Configuration
GV-4.2-013	Verify that downstream GAI system impacts (such as the use of third-party plug-ins) are included in the impact documentation process.	Value Chain and Component Integration

GV-4.2-014	Verify that the organizational list of risks related to the use of the GAI system are updated based on unforeseen GAI system incidents.	
AI Actors: 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	**Action**	**Risks**
GV-4.3-001	Allocate resources and adjust adoption, development, and implementation timeframes to enable independent measurement, continuous monitoring, and fulsome information sharing for GAI system risks.	
GV-4.3-002	Develop standardized documentation templates for efficient review of risk measurement results.	
GV-4.3-003	Establish minimum thresholds for performance 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.	
GV-4.3-004	Establish organizational roles, policies, and procedures for communicating GAI system incidents and performance to AI actors and downstream stakeholders, via community or official resources (e.g., AI Incident Database, AVID, AI Litigation Database, CVE, OECD Incident Monitor, or others).	Human AI Configuration, Value Chain and Component Integration
GV-4.3-005	Establish policies and procedures for pre-deployment GAI system testing that validates organizational capability to capture GAI system incident reporting criteria.	
GV-4.3-006	Establish policies, procedures, and processes that bolster independence of risk management and measurement functions (e.g., independent reporting chains, aligned incentives).	
GV-4.3-007	Establish policies, procedures, and processes that enable and incentivize in-context risk measurement via standardized measurement and structured public feedback approaches.	
GV-4.3-008	Organizational procedures 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.	
AI Actors: Fairness and Bias, Governance and Oversight, Operation and Monitoring, TEVV		

***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	Action	Risks
GV-5.1-001	Allocate time and resources for outreach, feedback, and recourse processes in GAI system development.	
GV-5.1-002	Disclose interactions with GAI systems to users prior to interactive activities.	Human AI Configuration
GV-5.1-003	Establish policy, guidelines and processes that: Engage independent experts to audit models, data sources, licenses, algorithms, and other system components, Consider sponsoring or engaging in community- based exercises (e.g., bug bounties, hackathons, competitions) where AI Actors assess and benchmark the performance of AI systems, including the robustness of content provenance management under various conditions; Document data sources, licenses, training methodologies, and trade-offs considered in the design of AI systems; Establish mechanisms, platforms or channels (e.g., user interfaces, web portals, forums) for independent experts, users, or community members to provide feedback related to AI systems; Adjudicate and implement relevant feedback at a regular cadence, Establish transparency mechanisms to track the origin of data and generated content; Audit and validate these mechanisms.	Human AI Configuration, Information Integrity, Intellectual Property
GV-5.1-004	Establish processes to bolster internal AI actor culture in alignment with organizational principles and norms and to empower exploration of GAI limitations beyond development settings.	Human AI Configuration, Toxicity, Bias, and Homogenization
GV-5.1-005	Establish the following GAI-specific policies and procedures for independent AI Actors: Continuous improvement processes for increasing explainability and mitigating other risks; Impact assessments, Incentives for internal AI actors to provide feedback and conduct independent risk management activities; Independent management and reporting structures for AI actors engaged in model and system audit, validation, and oversight; TEVV processes for the effectiveness of feedback mechanisms employing participation rates, resolution time, or similar measurements.	Human AI Configuration
GV-5.1-006	Provide thorough instructions for GAI system users to provide feedback and understand recourse mechanisms.	Human AI Configuration
GV-5.1-007	Standardize user feedback about GAI system behavior, risks and limitations for efficient adjudication and incorporation.	Human AI Configuration

AI Actors: 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	**Action**	**Risks**
GV-6.1-001	Categorize different types of GAI content with associated third party risks (i.e., copyright, intellectual property, data privacy).	Data Privacy, Intellectual Property, Value Chain and Component Integration
GV-6.1-002	Conduct due diligence on third-party entities and end-users from those entities before entering into agreements with them (e.g., checking references, reviewing their content handling processes, etc.).	Human AI Configuration, Value Chain and Component Integration
GV-6.1-003	Conduct joint educational activities and events in collaboration with third-parties to promote content provenance best practices.	Information Integrity, Value Chain and Component Integration
GV-6.1-004	Conduct regular audits of third-party entities to ensure compliance with contractual agreements.	Value Chain and Component Integration
GV-6.1-005	Define and communicate organizational roles and responsibilities for GAI acquisition, human resources, procurement, and talent management processes in policies and procedures.	Human AI Configuration
GV-6.1-006	Develop an incident response plan for third parties specifically tailored to address content provenance incidents or breaches and regularly test and update the incident response plan with feedback form external and third party stakeholders.	Data Privacy, Information Integrity, Information Security, Value Chain and Component Integration
GV-6.1-007	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-6.1-008	Develop risk tolerance and criteria to quantitatively assess and compare the level of risk associated with different third-party entities (i.e., reputation, track record, security measure, and the sensitivity of the content they handle).	Information Security, Value Chain and Component Integration
GV-6.1-009	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.	Information Integrity, Information Security
GV-6.1-010	Establish processes to maintain awareness of evolving risks, technologies, and best practices in content provenance management.	Information Integrity

GV-6.1-011	Implement a supplier risk assessment framework to continuously evaluate and monitor third-party entities' performance and adherence to content provenance standards and technologies (e.g., digital signatures, watermarks, cryptography, etc.) to detect anomalies and unauthorized changes; services acquisition and supply chain risk management; legal compliance (e.g., copyright, trademarks, and data privacy laws).	Data Privacy, Information Integrity, Information Security, Intellectual Property, Value Chain and Component Integration
GV-6.1-012	Include audit clauses in contracts that allow the organization to verify compliance with content provenance requirements.	Information Integrity
GV-6.1-013	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-6.1-014	Maintain detailed records of content provenance, including sources, timestamps, metadata, and any changes made by third parties.	Information Integrity, Value Chain and Component Integration
GV-6.1-015	Provide proper training to internal employees on content provenance best practices, risks, and reporting procedures.	Information Integrity
GV-6.1-016	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 policies to: Address robotic process automation (RPA), software-as-a-service (SAAS), and other solutions that may rely on embedded GAI technologies; Address ongoing audits, assessments, and alerting, dynamic risk assessments, and real-time reporting tools for monitoring third-party GAI risks; Address accessibility, accommodations, or opt-outs in GAI vendor offerings; Address commercial use of GAI outputs and secondary use of collected data by third parties; Assess vendor risk controls for intellectual property infringements and data privacy; Consider policy adjustments across GAI modeling libraries, tools and APIs, fine-tuned models, and embedded tools; Establish ownership of GAI acquisition and procurement processes; Include relevant organizational functions in evaluations of GAI third parties (e.g., legal, information technology (IT), security, privacy, fair lending); Include instruction on intellectual property infringement and other third-party GAI risks in GAI training for AI actors; Screen GAI vendors, open source or proprietary GAI tools, or GAI service providers against incident or vulnerability databases; Screen open source or proprietary GAI training data or outputs against patents, copyrights, trademarks and trade secrets.	Data Privacy, Human AI Configuration, Information Security, Intellectual Property, Value Chain and Component Integration, Toxicity, Bias, and Homogenization
GV-6.1-017	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
GV-6.1-018	Update human resource and talent management standards to address acceptable use of GAI.	Human AI Configuration

GV-6.1-019	Update third-party contracts, service agreements, and warranties to address GAI risks; Contracts, service agreements, and similar documents may include GAI-specific indemnity clauses, dispute resolution mechanisms, and other risk controls.	Value Chain and Component Integration
AI Actors: Operation and Monitoring, Procurement, Third-party entities		

1

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	**Action**	**Risks**
GV-6.2-001	Apply existing organizational risk management policies, procedures, and documentation processes to third-party GAI data and systems, including open source data and software.	Intellectual Property, Value Chain and Component Integration
GV-6.2-002	Document downstream GAI system impacts (e.g., the use of third-party plug-ins) for third party dependencies.	Value Chain and Component Integration
GV-6.2-003	Document GAI system supply chain risks to identify over-reliance on third party data or GAI systems and to identify fallbacks.	Value Chain and Component Integration
GV-6.2-004	Document incidents involving third-party GAI data and systems, including open source data and software.	Intellectual Property, Value Chain and Component Integration
GV-6.2-005	Enumerate organizational GAI system risks based on external dependencies on third-party data or GAI systems.	Value Chain and Component Integration
GV-6.2-006	Establish acceptable use policies that identify dependencies, potential impacts, and risks associated with third-party data or GAI systems deemed high-risk.	Value Chain and Component Integration
GV-6.2-007	Establish contingency and communication plans to support fallback alternatives for downstream users in the event the GAI system is disabled.	Human AI Configuration, Value Chain and Component Integration
GV-6.2-008	Establish incident response plans for third-party GAI technologies deemed high-risk: 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, Toxicity, Bias, and Homogenization
GV-6.2-009	Establish organizational roles, policies, and procedures for communicating with data and GAI system providers regarding performance, disclosure of GAI system inputs, and use of third-party data and GAI systems.	Human AI Configuration, Value Chain and Component Integration

GV-6.2-010	Establish policies and procedures for continuous monitoring of third-party GAI systems in deployment.	Value Chain and Component Integration
GV-6.2-011	Establish policies and procedures that address GAI data redundancy, including model weights and other system artifacts.	Toxicity, Bias, and Homogenization
GV-6.2-012	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.	
GV-6.2-013	Identify and document high-risk third-party GAI technologies in organizational AI inventories, including open-source GAI software.	Intellectual Property, Value Chain and Component Integration
GV-6.2-014	Review GAI vendor documentation for thorough instructions, meaningful transparency into data or system mechanisms, ample support and contact information, and alignment with organizational principles.	Value Chain and Component Integration, Toxicity, Bias, and Homogenization
GV-6.2-015	Review GAI vendor release cadences and roadmaps for irregularities and alignment with organizational principles.	Value Chain and Component Integration, Toxicity, Bias, and Homogenization
GV-6.2-016	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 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 line 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 Actors: 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	**Action**	**Risks**
MP-1.1-001	Apply risk mapping and measurement plans to third-party and open-source systems.	Intellectual Property, Value Chain and Component Integration

MP-1.1-002	Collaborate with domain experts to explore and document gaps, limitations, and risks in pre-deployment testing and the practical and contextual differences between pre-deployment testing and the anticipated context(s) of use.	
MP-1.1-003	Conduct impact assessments or review past known incidents and failure modes to prioritize and inform risk measurement.	
MP-1.1-004	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.	Toxicity, Bias, and Homogenization
MP-1.1-005	Document GAI system ownership, intended use, direct organizational value, and assumptions and limitations.	
MP-1.1-006	Document risk measurement plans that address: 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; Risks associated with trustworthy characteristics across the AI lifecycle; Standard measurement and structured human feedback approaches; Anticipated human-AI configurations.	Human AI Configuration, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MP-1.1-007	Document risks related to transparency, accountability, explainability, and interpretability in risk measurement plans, system risk assessments, and deployment approval (“go”/”no-go”) decisions.	
MP-1.1-008	Document system requirements, ownership, and AI actor roles and responsibilities for human oversight of GAI systems.	Human AI Configuration
MP-1.1-009	Document the extent to which a lack of transparency or explainability impedes risk measurement across the AI lifecycle.	
MP-1.1-010	Identify and document foreseeable illegal uses or applications that surpass organizational risk tolerances.	
AI Actors: 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	**Action**	**Risks**
MP-1.2-001	Document the credentials and qualifications of organizational AI actors and AI actor team composition.	Human AI Configuration
MP-1.2-002	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 TEVV of GAI technology, and other risk measurement and management functions.	Human AI Configuration, Toxicity, Bias, and Homogenization
MP-1.2-003	Establish continuous improvement processes to increase diversity and representativeness in AI actor teams, standard measurement resources, and structured public feedback participants from subgroup populations in-context.	Human AI Configuration, Toxicity, Bias, and Homogenization
MP-1.2-004	Verify that AI actor team membership includes demographic diversity, applicable domain expertise, varied education backgrounds, and lived experiences.	Human AI Configuration, Toxicity, Bias, and Homogenization
MP-1.2-005	Verify that data or benchmarks used in risk measurement, and users, participants, or subjects involved in structured public feedback exercises are representative of diverse in-context user populations.	Human AI Configuration, Toxicity, Bias, and Homogenization
AI Actors: 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	**Action**	**Risks**
MP-2.1-001	Define GAI system's task(s) that relate to content provenance, such as original content creation, media synthesis, or data augmentation while incorporating tracking measures.	Information Integrity
MP-2.1-002	Establish known assumptions and practices for determining data origin and content lineage, for documentation and evaluation.	Information Integrity
MP-2.1-003	Identify and document GAI task limitations that might impact the reliability or authenticity of the content provenance.	Information Integrity

MP-2.1-004	Institute audit trails for data and content flows within the system, including but not limited to, original data sources, data transformations, and decision-making criteria.	
MP-2.1-005	Review efficacy of content provenance techniques on a regular basis and update protocols as necessary.	Information Integrity
AI Actors: 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	**Action**	**Risks**
MP-2.2-001	Assess whether the GAI system fulfills its intended purpose within its operational context on a regular basis.	
MP-2.2-002	Evaluate whether GAI operators and end-users can accurately understand content lineage and origin.	Human AI Configuration, Information Integrity
MP-2.2-003	Identify and document how the system relies on upstream data sources for content provenance and if it serves as an upstream dependency for other systems.	Information Integrity, Value Chain and Component Integration
MP-2.2-004	Observe and analyze how the AI system interacts with external networks, and identify any potential for negative externalities, particularly where content provenance might be compromised.	Information Integrity
MP-2.2-005	Specify the environments where GAI systems may not function as intended related to content provenance.	Information Integrity
AI Actors: 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	**Action**	**Risks**

MP-2.3-001	Assess the accuracy, quality, reliability, and authenticity of the GAI content provenance 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).	Information Integrity
MP-2.3-002	Curate and maintain high quality datasets that are accurate, relevant, consistent, and representative as well as be well-documented complying with ethical and legal standards along with diverse data points.	Toxicity, Bias, and Homogenization
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	Design GAI systems to support content provenance such as tracking the lineage (e.g., data sources used to train the system, parameters used to generate content, etc.) and to verify authenticity (e.g., using digital signatures or watermarks).	Information Integrity
MP-2.3-005	Develop and implement testing techniques to identify any GAI produced content (e.g., synthetic media) that might be indistinguishable from human-generated content.	Information Integrity
MP-2.3-006	Document GAI content provenance techniques (including experimental methods), testing, evaluation, performance, and validation metrics throughout the AI lifecycle.	Information Integrity
MP-2.3-007	Implement plans for GAI systems to undergo regular adversarial testing to identify vulnerabilities and potential manipulation risks.	Information Security
MP-2.3-008	Integrate GAI systems with existing content management and version control systems, to enable content provenance to be tracked across the lifecycle.	Information Integrity
MP-2.3-009	Test GAI models using known inputs, context, and environment to confirm they produce expected outputs across a variety of methods (e.g., unit tests, integration tests, and system tests) and help to identify and address potential problems.	
MP-2.3-010	Use diverse large-scale and small-scale datasets for testing and evaluation to ensure that the AI system can perform well on a variety of different types of data.	Toxicity, Bias, and Homogenization
MP-2.3-011	Verify that GAI content provenance is accurate and reliable by using cryptographic techniques and performing formal audits to ensure it has not been manipulated.	Information Integrity

MP-2.3-012	Verify that the AI system’s content provenance complies with relevant laws and regulations, such as legal infringement, terms and conditions, copyright and intellectual property rights, when using data sources and generating content.	Information Integrity, Intellectual Property
AI Actors: 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	**Action**	**Risks**
MP-3.4-001	Adapt existing training programs to include modules on content provenance.	Information Integrity
MP-3.4-002	Develop certification programs that test proficiency in managing AI risks and interpreting content provenance, relevant to specific industry and context.	Information Integrity
MP-3.4-003	Delineate human proficiency tests from tests of AI capabilities.	Human-AI Configuration
MP-3.4-004	Integrate human and other qualitative inputs to comprehensively assess content provenance.	Information Integrity
MP-3.4-005	Ensure that output provided to operators and practitioners is both interactive and well-defined, incorporating content provenance data that can be easily interpreted for effective downstream decision-making.	Information Integrity, Value Chain and Component Integration
MP-3.4-006	Establish and adhere to design principles that ensure safe and ethical operation, taking into account the interpretation of content provenance information.	Information Integrity, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MP-3.4-007	Implement systems to continually monitor and track the outcomes of human-AI collaborations for future refinement and improvements, integrating a focus on content provenance wherever applicable.	Human AI Configuration, Information Integrity
MP-3.4-008	Involve the end-users, practitioners, and operators in AI system prototyping and testing activities. Make sure these tests cover various scenarios where content provenance could play a critical role, such as crisis situations or ethically sensitive contexts.	Human AI Configuration, Information Integrity, Toxicity, Bias, and Homogenization
MP-3.4-009	Match the complexity of GAI system explanations and the provenance data to the level of the problem and contextual intricacy.	Information Integrity
AI Actors: 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	**Action**	**Risks**
MP-4.1-001	Conduct audits on third-party processes and personnel including an examination of the third-party's reputation.	Value Chain and Component Integration
MP-4.1-002	Conduct periodic audits and monitor AI generated content for privacy risks; address any possible instances of sensitive data exposure.	Data Privacy
MP-4.1-003	Consider using synthetic data as applicable to train AI models in place of real-world data to match the statistical properties of real-world data without disclosing personally identifiable information.	
MP-4.1-004	Develop practices for periodic monitoring of GAI outputs for possible intellectual property infringements and other risks and implement processes for responding to potential intellectual property infringement claims.	Intellectual Property
MP-4.1-005	Document all aspects of the AI development process including data sources, model architectures and training procedures to support reproduction of results, identify any potential problems, and implement mitigation strategies.	
MP-4.1-006	Document compliance with legal requirements across the AI lifecycle, including copyright concerns, privacy protections.	Data Privacy, Intellectual Property
MP-4.1-007	Document training data curation policies, including policies to verify that consent was obtained for the likeness or image of individuals.	Obscene, Degrading, and/or Abusive Content
MP-4.1-008	Employ encryption techniques and proper safeguards to ensure secure data storage and transfer to protect data privacy.	Data Privacy, Information Security, Dangerous or Violent Recommendations
MP-4.1-009	Establish policies for collection, retention, and minimum quality of data, in consideration of the following risks: Disclosure of CBRN information by removing CBRN information from training data, Use of Illegal or dangerous content; Training data imbalance across sub-groups by modality, such as languages for LLMs or skin tone for image generation; Leak of personally identifiable information, including facial likenesses of individuals unless consent is obtained for use of their images.	CBRN Information, Intellectual Property, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations, Data Privacy
MP-4.1-010	Implement bias mitigation approaches by addressing sources of bias in the training data and by evaluating AI models for bias periodically.	Toxicity, Bias, and Homogenization
MP-4.1-011	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-4.1-012	Implement reproducibility techniques, including: share data publicly or privately using license and citation; develop code according to standard software practices; track and document experiments and results; manage the software environment and dependencies; utilize virtual environments, version control, and maintain a requirements document; manage models and artifacts; tracking AI model versions and documenting model details along with parameters and experimental results; document data management processes and establish a testing/validation process to maintain reliable results.	Confabulation, Intellectual Property, Value Chain and Component Integration
MP-4.1-013	Re-evaluate models that were fine-tuned on top of third-party models.	Value Chain and Component Integration
MP-4.1-014	Re-evaluate risks when adapting GAI models to new domains.	
MP-4.1-015	Review service level agreements and contracts, including license agreements and any legal documents associated with the third-party intellectual properties, technologies, and services.	Intellectual Property, Value Chain and Component Integration
MP-4.1-016	Use approaches to detect the presence of sensitive data in generated output text, image, video, or audio, and verify that the model will mask any detected sensitive data.	Information Integrity
MP-4.1-017	Use trusted sources for training data that are licensed or open source and ensure that the entity has the legal right for the use of proprietary training data.	Intellectual Property
MP-4.1-018	Apply strong anonymization and de-identification, and/or differential privacy techniques to protect the privacy of individuals in the training data.	Data Privacy
MP-4.1-019	Verify that third-party models are in compliance with existing use licenses.	Intellectual Property, Value Chain and Component Integration
AI Actors: 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	**Action**	**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 using zero-knowledge proof approaches).	Information Integrity, Information Security

MP-5.1-002	Assess and document risks related to content provenance. e.g., document the presence, absence, or effectiveness of tagging systems, cryptographic hashes, blockchain-based, or distributed ledger technology solutions that improve content tracking transparency and immutability.	Information Integrity
MP-5.1-003	Consider GAI-specific mapped risks (e.g., complex security requirements, potential for emotional entanglement of users, large supply chains) in estimates for likelihood, magnitude of impact and risk.	Human AI Configuration, Information Security, Value Chain and Component Integration
MP-5.1-004	Document estimates of likelihood, magnitude of impact, and risk for GAI systems in a central repository (e.g., organizational AI inventory.).	
MP-5.1-005	Enumerate potential impacts related to content provenance, including best-case, average-case, and worst-case scenarios.	Information Integrity
MP-5.1-006	Estimate likelihood of enumerated impact scenarios using past data or expert judgment, analysis of known public incidents, standard measurement, and structured human feedback results.	CBRN Information, Dangerous or Violent Recommendations
MP-5.1-007	Measure risk as the product of estimated likelihood and magnitude of impact of a GAI outcome.	
MP-5.1-008	Prioritize risk acceptance, management, or transfer activities based on risk estimates.	
MP-5.1-009	Prioritize standard measurement and structured public feedback processes based on risk assessment estimates.	
MP-5.1-010	Profile risks arising from GAI systems interacting with, manipulating, or generating content, and outlining known and potential vulnerabilities and the likelihood of their occurrence.	Information Security
MP-5.1-011	Scope GAI applications narrowly to enable risk-based governance and controls.	
AI Actors: 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	Action	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
MP-5.2-003	Publish guidance for external AI actors to report unanticipated impacts of the GAI system and to engage with the organization in the event of GAI system impacts.	Human AI Configuration

AI Actors: 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	Action	Risks
MS-1.1-001	Assess the effectiveness of implemented methods and metrics at an ongoing cadence as part of continuous improvement activities.	
MS-1.1-002	Collaborate with multidisciplinary experts (e.g., in the fields of responsible use of GAI, cybersecurity, or digital forensics) to ensure the selected risk management approaches are robust and effective.	Information Security; CBRN Information, Toxicity, Bias, and Homogenization
MS-1.1-003	Conduct adversarial role-playing exercises, AI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes.	Information Security, Unknowns
MS-1.1-004	Conduct traditional assessment or TEVV exercises to measure the prevalence of known risks in deployment contexts.	
MS-1.1-005	Document GAI risk measurement or tracking approaches, including tracking of risks that cannot be easily measured before deployment (e.g., ecosystem-level risks or risks that unfold over longer time scales).	

MS-1.1-006	Employ digital signatures and watermarking, blockchain technology, reverse image and video search, metadata analysis, steganalysis, and/or forensic analysis to trace the origin and modifications of digital content.	Information Integrity
MS-1.1-007	Employ similarity metrics, tampering indicators, blockchain confirmation, metadata consistency, hidden data detection rate, source reliability, and consistency with known patterns to measure content provenance risks.	Information Integrity
MS-1.1-008	Identify content provenance risks in the end-to-end AI supply chain, including risks associated with data suppliers, data annotators, R&D, joint ventures, academic or nonprofit projects/partners, third party vendors, and contractors.	Information Integrity, Value Chain and Component Integration
MS-1.1-009	Identify potential content provenance risks and harms in GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content. Enumerate and rank risks and/or harms based on their likelihood and potential impact, and determine how well provenance solutions address specific risks and/or harms.	Information Integrity, Dangerous or Violent Recommendations, Obscene, Degrading, and/or Abusive Content
MS-1.1-010	Implement appropriate approaches and metrics for measuring AI-related content provenance the and the aforementioned risks and harms.	Information Integrity, Dangerous or Violent Recommendations
MS-1.1-011	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-1.1-012	Invest in R&D capabilities to evaluate and implement novel methods and technologies for the measurement of AI-related risks in content provenance, toxicity, and CBRN.	Information Integrity, CBRN Information, Obscene, Degrading, and/or Abusive Content
MS-1.1-013	Prioritize risk measurement according to risk severity as determined during mapping activities.	
MS-1.1-014	Provide content provenance risk management education to AI actors, users, and stakeholders.	Human AI Configuration, Information Integrity
MS-1.1-015	Track and document risks or opportunities related to content provenance 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).	Information Integrity
MS-1.1-016	Track the number of output data items that are accompanied by provenance information (e.g., watermarks, cryptographic tags).	Information Integrity
MS-1.1-017	Track the number of training and input (e.g., prompts) data items that have provenance records and output data items that potentially infringe on intellectual property rights.	Information Integrity, Intellectual Property

MS-1.1-018	Track the number of training and input data items covered by intellectual property rights (e.g., copyright, trademark, trade secret).	Intellectual Property
MS-1.1-019	Validate the reliability and integrity of the original data and measure inherent dependence on training data and its quality.	
AI Actors: 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	**Action**	**Risks**
MS-1.3-001	Define relevant groups of interest (e.g., demographic groups, subject matter experts, past experience with GAI technology) within the context of use as part of plans for gathering structured public feedback.	Human AI Configuration, Toxicity, Bias, and Homogenization, CBRN
MS-1.3-002	Define sequence of actions for AI red-teaming exercises and accompanying necessary documentation practices.	
MS-1.3-003	Define use cases, contexts of use, capabilities, and negative impacts where structured human feedback exercises, e.g., AI red-teaming, would be most beneficial for AI risk measurement and management based on the context of use.	
MS-1.3-004	Develop a suite of suitable metrics to evaluate structured feedback results, informed by representative AI actors.	Human AI Configuration, Toxicity, Bias, and Homogenization, CBRN
MS-1.3-005	Execute independent audit, AI red-teaming, impact assessments, or other structured human feedback processes 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, Toxicity, Bias, and Homogenization, CBRN
MS-1.3-006	Identify and implement methods for post-hoc evaluation of the effectiveness of structured human feedback processes such as auditing, impact assessments, and AI red-teaming.	
MS-1.3-007	Identify and implement methods for translating, evaluating, and integrating structured human feedback output into AI risk management processes, continuous improvement processes, and related organizational decision making.	
MS-1.3-008	Identify criteria for determining when structured human feedback exercises are complete.	

MS-1.3-009	Identify mechanisms and teams to evaluate or other structured human feedback outcomes.	
MS-1.3-010	Recruit auditors, AI red-teams, and structured feedback participants in consideration of the linguistic, dialectal, and socio-cultural environment of the expected user base.	Human AI Configuration
MS-1.3-011	Share structured feedback with relevant AI actors to address identified risks.	Human AI Configuration
MS-1.3-012	Verify demographic diversity of identified subgroups in structured feedback exercises.	Toxicity, Bias, and Homogenization
MS-1.3-013	Verify those conducting structured human feedback exercises are not directly involved in system development tasks for the same GAI model.	
AI Actors: 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	**Action**	**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, Toxicity, Bias, and Homogenization
MS-2.2-002	Disaggregate evaluation metrics by demographic factors to identify any discrepancies in how content provenance mechanisms work across diverse populations.	Information Integrity, Toxicity, Bias, and Homogenization
MS-2.2-003	Document how content provenance mechanisms are operated in the context of privacy and security including: Anonymize data to protect the privacy of human subjects; Remove any personally identifiable information (PII) to prevent potential harm or misuse.	Data Privacy, Human AI Configuration, Information Integrity, Information Security, Dangerous or Violent Recommendations
MS-2.2-004	Employ techniques like chaos engineering and stakeholder feedback to evaluate the quality and integrity of data used in training and the provenance of AI-generated content.	Information Integrity
MS-2.2-005	Identify biases present in the training data for downstream mitigation using available techniques (e.g., data visualization tools).	Value Chain and Component Integration, Toxicity, Bias, and Homogenization

MS-2.2-006	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 to identify issues and improve GAI system fairness.	Toxicity, Bias, and Homogenization
MS-2.2-007	Implement robust cybersecurity measures to protect both the research data, the GAI system and its content provenance from unauthorized access, breaches, or tampering and unauthorized disclosure of human subject information.	Data Privacy, Human AI Configuration, Information Integrity, Information Security
MS-2.2-008	Obtain informed consent from human subject evaluation participants. Informed consent should include: the nature of the study, information about the use of GAI related to content provenance, its purpose, and potential implications.	Data Privacy, Human AI Configuration, Information Integrity
MS-2.2-010	Practice responsible disclosure of findings and report discovered vulnerabilities or biases related to GAI systems and its content provenance.	Information Integrity, Information Security, Toxicity, Bias, and Homogenization
MS-2.2-011	Provide human subjects with options to revoke their consent for future use of their data in GAI applications, particularly in content provenance aspects.	Data Privacy, Human AI Configuration, Information Integrity
MS-2.2-012	Use Institutional Review Boards as applicable for evaluations that involve human subjects.	Human AI Configuration
MS-2.2-013	Use techniques such as anonymization or differential privacy to minimize the risks associated with linking AI-generated content back to individual human subjects.	Data Privacy, Human AI Configuration
MS-2.2-014	Verify accountability and fairness through documentation of the algorithms, parameters, and methodologies used in the evaluation to allow for external scrutiny.	Toxicity, Bias, and Homogenization
MS-2.2-015	Verify that human subjects selected for evaluation are representative of the population for the relevant GAI use-case; Consider demographics such as age, gender, race, ethnicity, socioeconomic status, and geographical location to avoid biases in the AI system related to content provenance.	Human AI Configuration, Information Integrity, Toxicity, Bias, and Homogenization
MS-2.2-016	Work in close collaboration with domain experts to understand the specific requirements and potential pitfalls related to content provenance in the GAI system's intended context of use.	Information Integrity
AI Actors: 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	**Action**	**Risks**
MS-2.3-001	Analyze differences between intended and actual population of users or data subjects, including likelihood for errors, incidents, or negative impacts.	Confabulation, Human AI Configuration, Information Integrity
MS-2.3-002	Conduct field testing on sampled sub-populations prior to deployment to the entire population.	
MS-2.3-003	Conduct TEVV in the operational environment in accordance with organizational policies and regulatory or disciplinary requirements (e.g., informed consent, institutional review board approval, human research protections, privacy requirements).	Data Privacy
MS-2.3-004	Consider baseline model performance on suites of benchmarks when selecting a model for fine tuning.	
MS-2.3-005	Evaluate claims of model capabilities using empirically validated methods.	
MS-2.3-006	Include metrics measuring reporting rates for harmful or offensive content in field testing.	Dangerous or Violent Recommendations
MS-2.3-007	Share results of pre-deployment testing with relevant AI actors, such as those with system release approval authority.	Human AI Configuration
MS-2.3-008	Use disaggregated evaluation methods (e.g., by race, age, gender, ethnicity, ability, region) to improve granularity of AI system performance measures.	
MS-2.3-009	Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.	
MS-2.3-010	Verify that mechanisms to collect users' feedback are visible and traceable.	Human AI Configuration

AI Actors: 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	**Action**	**Risks**

MS-2.5-001	Apply standard measurement and structured human feedback approaches to internally-developed and third-party GAI systems.	Value Chain and Component Integration
MS-2.5-002	Avoid extrapolating GAI system performance or capabilities from narrow, non-systematic, and anecdotal assessments.	
MS-2.5-003	Conduct security assessments and audits to measure the integrity of training data, system software, and system outputs.	Information Security
MS-2.5-004	Document the construct validity of methodologies employed in GAI systems relative to their context of use.	
MS-2.5-005	Document the extent to which human domain knowledge is employed to improve GAI system performance, via, e.g., RLHF, fine-tuning, content moderation, business rules.	
MS-2.5-006	Establish metrics or KPIs to determine whether GAI systems meet minimum performance standards for reliability and validity.	
MS-2.5-007	Measure, monitor, and document prevalence of erroneous GAI output content, system availability, and reproducibility of outcomes via field testing or other randomized controlled experiments.	
MS-2.5-008	Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities.	Confabulation
MS-2.5-009	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-2.5-010	Track and document relevant version numbers, planned updates, hotfixes, and other GAI system change management information.	
MS-2.5-011	Update standard train/test model evaluation processes for GAI systems. Consider: Unwanted or undocumented overlaps in train and TEVV data sources, including their negative spaces (i.e., what is not represented in both); Employing substring matching or embedding distance approaches to assess similarity across data partitions.	
MS-2.5-012	Verify GAI system training data and TEVV data provenance, and that fine-tuning data is grounded.	Information Integrity
AI Actors: 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	Action	Risks
MS-2.6-001	Assess adverse impacts health and wellbeing impacts for supply chain or other AI actors that are exposed to obscene, toxic, or violent information during the course of GAI training and maintenance.	Human AI Configuration, Obscene, Degrading, and/or Abusive Content, Value Chain and Component Integration, Dangerous or Violent Recommendations
MS-2.6-002	Assess levels of toxicity, 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, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations, CBRN Information
MS-2.6-003	Measure and document incident response times, system down times, and system availability: Perform standard measurement and structured human feedback on GAI systems to detect safety and reliability impacts and harms; Apply human subjects research protocols and other applicable safety controls when conducting A/B testing, AI red-teaming, focus groups, or human testbed measurements; Identify and document any applications related to robotics, RPA, and autonomous vehicles; Conduct AI red-teaming exercises to identify harms and impacts related to safety and validity, reliability, privacy, toxicity and other risks; Monitor high-risk GAI systems continually for safety and reliability risks once deployed; Monitor GAI systems to detect drift and anomalies relative to expected performance and training baselines.	Data Privacy, Human AI Configuration, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MS-2.6-004	Re-evaluate safety features of fine-tuned models when the risk of harm exceeds organizational risk tolerance.	Dangerous or Violent Recommendations
MS-2.6-005	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 or Violent Recommendations
MS-2.6-006	Track and document past failed GAI system designs to inform risk measurement for safety and validity risks.	Dangerous or Violent Recommendations
MS-2.6-007	Verify capabilities for limiting, pausing, updating, or terminating GAI systems quickly.	
MS-2.6-008	Verify rollover, fallback, or redundancy capabilities for high-risk GAI systems.	

MS-2.6-009	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-2.6-010	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, Information Security
AI Actors: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.		
Action ID	**Action**	**Risks**
MS-2.7-001	Apply established security measures to: Assess risks of backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering other baseline security concerns; Audit supply chains to identify risks arising from, e.g., data poisoning and malware, software and hardware vulnerabilities, third-party personnel and software; Audit GAI systems, pipelines, plugins and other related artifacts for unauthorized access, malware, and other known vulnerabilities.	Data Privacy, Information Integrity, Information Security, Value Chain and Component Integration
MS-2.7-002	Assess the completeness of documentation related to data provenance, access controls, and incident response procedures. Verify GAI system content provenance documentation aligns with relevant regulations and standards.	Information Integrity, Toxicity, Bias, and Homogenization
MS-2.7-003	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-2.7-004	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.	Human AI Configuration, Information Integrity
MS-2.7-005	Engage with security experts, developers, and researchers through information sharing mechanisms to stay updated with the latest advancements in AI security related to content provenance. Contribute findings related to AI system security and content provenance via information sharing mechanisms, workshops, or publications.	Information Integrity, Information Security
MS-2.7-006	Establish measures and evaluate GAI resiliency as part of pre-deployment testing to ensure GAI will function under adverse conditions and restore full functionality in a trustworthy manner.	

MS-2.7-007	Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, penetrations, or provenance verification.	Information Integrity, Information Security
MS-2.7-008	Maintain awareness of emergent GAI security risks and associated countermeasures through community resources, official guidance, or research literature.	Information Security, Unknowns
MS-2.7-009	Measure reliability of content provenance verification methods, such as watermarking, cryptographic signatures, hashing, blockchain, or other 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-2.7-010	Measure the average response time to security incidents related to content provenance, and the proportion of incidents resolved with and without significant impact.	Information Integrity, Information Security
MS-2.7-011	Measure the rate at which recommendations from security audits and incidents are implemented related to content provenance. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback related to content provenance.	Information Integrity, Information Security
MS-2.7-012	Monitor and review the completeness and validity of security documentation and verify it aligns with the current state of the GAI system and its content provenance.	Information Integrity, Information Security, Toxicity, Bias, and Homogenization
MS-2.7-013	Monitor GAI system downtime and measure its impact on operations.	
MS-2.7-014	Monitor GAI systems in deployment for anomalous use and security risks.	Information Security
MS-2.7-015	Monitor the number of security-related incident reports from users, indicating their awareness and willingness to report issues.	Human AI Configuration, Information Security
MS-2.7-016	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, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MS-2.7-017	Review deployment approval processes and verify that processes address relevant GAI security risks.	Information Security
MS-2.7-018	Review incident response procedures and verify adequate functionality to identify, contain, eliminate, and recover from complex GAI system incidents that implicate impacts across the trustworthy characteristics.	
MS-2.7-019	Track and document access and updates to GAI system training data; verify appropriate security measures for training data at GAI vendors and service providers.	Information Security, Value Chain and Component Integration

MS-2.7-020	Track GAI system performance metrics such as response time and throughput under different loads and usage patterns related to content provenance.	Information Integrity
MS-2.7-021	Track the number of users who have completed security training programs regarding the security of content provenance.	Human AI Configuration, Information Integrity, Information Security
MS-2.7-022	Verify fine-tuning does not compromise safety and security controls.	Information Integrity, Information Security, Dangerous or Violent Recommendations
MS-2.7-023	Verify organizational policies, procedures, and processes for treatment of GAI security and resiliency risks.	Information Security
MS-2.7-024	Verify vendor documentation for data and software security controls.	Information Security, Value Chain and Component Integration
MS-2.7-025	Work with domain experts to capture stakeholder confidence in GAI system security and perceived effectiveness related to content provenance.	Information Integrity, Information Security
AI Actors: 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	**Action**	**Risks**
MS-2.8-001	Compile and communicate statistics on policy violations, take-down requests, intellectual property infringement, and information integrity for organizational GAI systems: Analyze transparency reports across demographic groups, languages groups, and other segments relevant to the deployment context.	Information Integrity, Intellectual Property, Toxicity, Bias, and Homogenization
MS-2.8-002	Document the instructions given to data annotators or AI red-teamers.	
MS-2.8-003	Document where in the data pipeline human labor is being used.	
MS-2.8-004	Establish a mechanism for appealing usage policy violations.	
MS-2.8-005	Maintain awareness of AI regulations and standards in relevant jurisdictions related to GAI systems and content provenance.	Information Integrity
MS-2.8-006	Measure the effectiveness or accessibility of procedures to appeal adverse, harmful, or incorrect outcomes from GAI systems.	Human AI Configuration, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations

MS-2.8-007	Review and consider GAI system transparency artifacts such as impact assessments, system cards, model cards, and traditional risk management documentation as part of organizational decision making.	
MS-2.8-008	Review licenses, patents, or other intellectual property rights pertaining to information in system training data.	Intellectual Property
MS-2.8-009	Track AI actor decisions along the lifecycle to determine sources of systemic and cognitive bias and identify management and mitigation approaches.	Human AI Configuration, Toxicity, Bias, and Homogenization
MS-2.8-010	Use interpretable machine learning techniques to make AI processes and outcomes more transparent, and easier to understand how decisions are made.	
MS-2.8-011	Use technologies such as blockchain and digital signatures 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-2.8-012	Verify adequacy of GAI system user instructions through user testing.	Human AI Configuration
MS-2.8-013	Verify that accurate information about GAI capabilities, opportunities, risks, and potential negative impacts are available on websites, press releases, organizational reports, social media, and public communication channels.	
MS-2.8-014	Verify the adequacy of feedback functionality in system user interfaces.	Human AI Configuration
MS-2.8-015	Verify the adequacy of redress processes for severe GAI system impacts.	
AI Actors: 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	**Action**	**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.	
MS-2.9-002	Apply transparency tools such as Datasheets, Data Nutrition Labels, and Model Cards to record explanatory and validation information.	

MS-2.9-003	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 approaches; Evaluation data; Ethical considerations; Legal and regulatory requirements.	Information Integrity, Toxicity, Bias, and Homogenization
MS-2.9-004	Measure and report: Comparisons to alternative approaches and benchmarks; Outcomes across demographic groups, languages groups, and other segments relevant to the deployment context; Reproducibility of outcomes or internal mechanisms; Sensitivity analysis and stress-testing results.	Toxicity, Bias, and Homogenization
MS-2.9-005	Verify calibration and robustness of applied explanation techniques and document their assumptions and limitations.	
AI Actors: 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	**Action**	**Risks**
MS-2.10-001	Collaborate with other AI actors, domain experts, and legal advisors to evaluate the impact of GAI applications on privacy related to the GAI system and its content provenance, in domains such as healthcare, finance, and criminal justice.	Data Privacy, Human AI Configuration, Information Integrity
MS-2.10-002	Conduct AI red-teaming to assess GAI system risks 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; Tracking or revealing location information of users or members of training datasets.	Human AI Configuration, Intellectual Property
MS-2.10-003	Document collection, use, management, and disclosure of biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information in datasets, in accordance with privacy and data governance policies and data privacy laws.	Data Privacy, Human AI Configuration, Intellectual Property
MS-2.10-004	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-tracking mechanisms.	Human AI Configuration, Information Integrity
MS-2.10-005	Establish and document protocols (authorization, duration, type) and access controls for training sets or production data containing biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information, in accordance with privacy and data governance policies and data privacy laws.	Data Privacy, Intellectual Property

MS-2.10-006	Implement consent mechanisms that are demonstrated to allow users to understand and control how their data is used in the GAI system and its content provenance.	Data Privacy, Human AI Configuration, Information Integrity
MS-2.10-007	Implement mechanisms to monitor, periodically review and document the provenance data to detect any inconsistencies or unauthorized modifications.	Information Integrity, Information Security
MS-2.10-008	Implement zero-knowledge proofs to balance transparency with privacy and allow verification of claims about content without exposing the actual data.	Data Privacy
MS-2.10-009	Leverage technologies such as blockchain to document the origin of, and any subsequent modifications to, generated content to enhance transparency and provide a secure method for provenance tracking.	Information Integrity, Information Security
MS-2.10-010	Track training, input and output items that contains personally identifiable information.	Data Privacy
MS-2.10-011	Verify compliance with data protection regulations.	Data Privacy
MS-2.10-012	Verify deduplication of training data samples.	Toxicity, Bias, and Homogenization
MS-2.10-013	Verify organizational policies, procedures, and processes for GAI systems address fundamental tenets of data privacy, e.g., Anonymization of private data; Consent to use data for targeted purposes or applications; Data collection and use in accordance with legal requirements and organizational policies; Reasonable data retention limits and requirements; User data deletion and rectification requests.	Data Privacy, Human AI Configuration
MS-2.10-014	Verify that biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information are removed from GAI training data.	Intellectual Property
AI Actors: 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	**Action**	**Risks**
MS-2.11-001	Apply use-case appropriate benchmarks (e.g., Bias Benchmark Questions, Real Toxicity Prompts, Winogender) to quantify systemic bias, stereotyping, denigration, and toxicity in GAI system outputs; Document assumptions and limitations of benchmarks relative to in-context deployment environment.	Toxicity, Bias, and Homogenization
MS-2.11-002	Assess content moderation and other output filtering technologies or processes for risks arising from human, systemic, and statistical/computational biases.	Toxicity, Bias, and Homogenization

MS-2.11-003	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. Identify types of harms, including harms in resource allocation, representational, quality of service, stereotyping, or erasure, Identify across, within, and intersecting groups that might be harmed; 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); Analyze quantified harms for contextually significant differences across groups, within groups, and among intersecting groups; Refine identification of within-group and intersectional group disparities, Evaluate underlying data distributions and employ sensitivity analysis during the analysis of quantified harms, Evaluate quality metrics including differential output across groups, Consider biases affecting small groups, within-group or intersectional communities, or single individuals.	Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MS-2.11-004	Evaluate practices along the lifecycle to identify potential sources of human-cognitive bias such as availability, observational, groupthink, funding, and confirmation bias, and to make implicit decision-making processes more explicit and open to investigation.	Toxicity, Bias, and Homogenization
MS-2.11-005	Identify the classes of individuals, groups, or environmental ecosystems which might be impacted by GAI systems through direct engagement with potentially impacted communities.	Environmental, Toxicity, Bias, and Homogenization
MS-2.11-006	Monitor for representational, financial, or other harms after GAI systems are deployed.	Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MS-2.11-007	Review, document, and measure sources of bias in 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 and toxicity in GAI system training data; Prevalence of GAI-generated data in GAI system training data.	Toxicity, Bias, and Homogenization, Unknowns

MS-2.11-008	Track and document AI actor credentials and qualifications.	Human AI Configuration
MS-2.11-009	Verify accessibility functionality; verify functionality and timeliness of accommodations and opt-out functionality or processes.	Human AI Configuration, Toxicity, Bias, and Homogenization
MS-2.11-010	Verify bias management in periodic model updates; test and recalibrate with updated and more representative data to manage bias within acceptable tolerances.	Toxicity, Bias, and Homogenization
MS-2.11-011	Verify training is not homogenous GAI-produced data in order to mitigate concerns of model collapse.	Toxicity, Bias, and Homogenization
AI Actors: 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	**Action**	**Risks**
MS-2.12-001	Assess safety to physical environments when deploying GAI systems.	Dangerous or Violent Recommendations
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	Track and document continuous improvement processes that enhance effectiveness of risk measurement for GAI environmental impacts and sustainability.	Environmental
MS-2.12-005	Verify effectiveness of carbon capture or offset programs, and address green-washing risks.	Environmental
AI Actors: 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	Action	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; Adhere to applicable laws and regulations when operationalizing models in high-volume settings (e.g., toxicity classifiers and automated content filters); Leverage domain expertise when modeling complex societal constructs such as toxicity.	Confabulation, Information Integrity, Toxicity, Bias, and Homogenization
MS-2.13-002	Document measurement and structured public feedback processes applied to organizational GAI systems in a centralized repository (i.e., organizational AI inventory).	
MS-2.13-003	Review GAI system metrics and associated pre-deployment processes to determine their ability to sustain system improvements, including the identification and removal of errors, harms, and negative impacts.	Confabulation, Information Integrity, Dangerous or Violent Recommendations

AI Actors: AI Deployment, Operation and Monitoring, TEVV

***MEASURE 3.1:** Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.

Action ID	Action	Risks
MS-3.1-001	Assess completeness of known use cases and expected performance of inputs, such as third-party data or upstream AI systems, or the performance of downstream systems which use the outputs of the GAI system, directly or indirectly, through engagement and outreach with AI Actors.	Human AI Configuration, Value Chain and Component Integration, Toxicity, Bias, and Homogenization
MS-3.1-002	Compare intended use and expected performance of GAI systems across all relevant contexts.	

MS-3.1-003	Elicit and track feedback for previously unknown uses of the GAI systems.	
AI Actors: AI Impact Assessment, Operation and Monitoring, TEVV		

1

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	**Action**	**Risks**
MS-3.2-001	Determine if available GAI system risk measurement approaches are applicable to the GAI system use contexts.	
MS-3.2-002	Document the rate of occurrence and severity of GAI harms to the organization and to external AI actors.	Human AI Configuration
MS-3.2-003	Establish processes for identifying emergent GAI system risks with external AI actors.	Human AI Configuration, Unknowns
MS-3.2-004	Identify measurement approaches for tracking GAI system risks if none exist.	
AI Actors: AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV		

2

***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	**Action**	**Risks**
MS-3.3-001	Conduct impact assessments on how AI-generated content might affect different social, economic, and cultural groups.	Toxicity, Bias, and Homogenization
MS-3.3-002	Conduct studies to understand how end users perceive and interact with GAI content related to 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	Design evaluation metrics that include parameters for content provenance quality, validity, reliability, authenticity or origin, and integrity of content.	Information Integrity
MS-3.3-004	Evaluate GAI system evaluation metrics based on feedback from relevant AI actors.	Human AI Configuration

MS-3.3-005	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.	Toxicity, Bias, and Homogenization
MS-3.3-006	Implement continuous monitoring of AI-generated content and provenance after system deployment for various types of drift. Verify GAI systems are adaptive and able to iteratively improve models and algorithms over time.	Information Integrity
MS-3.3-007	Integrate human evaluators to assess content quality and relevance.	Human AI Configuration
MS-3.3-008	Provide input for training materials about the capabilities and limitations of GAI systems related to content provenance 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, Toxicity, Bias, and Homogenization
MS-3.3-009	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, Toxicity, Bias, and Homogenization
MS-3.3-010	Regularly review structured human feedback and GAI system sensors and update based on the evolving needs and concerns of the impacted communities.	
MS-3.3-011	Utilize independent evaluations to assess content quality and types of potential biases and related negative impacts.	Toxicity, Bias, and Homogenization
MS-3.3-012	Verify AI actors engaged in GAI TEVV tasks for content provenance reflect diverse demographic and interdisciplinary backgrounds.	Human AI Configuration, Information Integrity, Toxicity, Bias, and Homogenization
AI Actors: 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	**Action**	**Risks**
MS-4.2-001	Conduct adversarial testing to assess the GAI system’s response to inputs intended to deceive or manipulate its content provenance and understand potential misuse scenarios and unintended outputs.	Information Integrity, Information Security

MS-4.2-002	Ensure both positive and negative feedback on GAI system functionality is assessed.	
MS-4.2-003	Ensure visible mechanisms to collect users' feedback are in place, including systems to report harmful and low quality content.	Human AI Configuration, Dangerous or Violent Recommendations
MS-4.2-004	Evaluate GAI system content provenance in real-world scenarios to observe its behavior in practical environments and reveal issues that might not surface in controlled and optimized testing environments.	Information Integrity
MS-4.2-005	Evaluate GAI system performance related to content provenance against predefined metrics and update the evaluation criteria as necessary to adapt to changing contexts and requirements.	Information Integrity
MS-4.2-006	Implement interpretability and explainability methods to evaluate GAI system decisions related to content provenance and verify alignment with intended purpose.	Information Integrity, Toxicity, Bias, and Homogenization
MS-4.2-007	Integrate structured human feedback results into calibration and update processes for traditional measurement approaches (e.g., benchmarks, performance assessments, data quality measurements).	
MS-4.2-008	Measure GAI system inputs and outputs to account for content provenance, data provenance, source reliability, contextual relevance and coherence, and security implications.	Information Integrity, Information Security
MS-4.2-009	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-4.2-010	Verify and document the incorporation of structured human feedback results into design, implementation, deployment approval ("go"/"no-go" decisions), monitoring, and decommission decisions.	
MS-4.2-011	Verify that GAI system development and deployment related to content provenance integrates trustworthiness characteristics.	Information Integrity
MS-4.2-012	Verify the performance of user feedback and recourse mechanisms, including analyses across various sub-groups.	Human AI Configuration, Toxicity, Bias, and Homogenization
MS-4.2-013	Work with domain experts to integrate insights from stakeholder feedback analysis into TEVV metrics and associated actions, and continuous improvement processes.	
MS-4.2-014	Work with domain experts to review feedback from end users, operators, and potentially impacted individuals and communities—enumerated in the Map function.	Human AI Configuration

MS-4.2-015	Work with domain experts who understand the GAI system context of use to evaluate the content's validity, relevance, and potential biases.	Toxicity, Bias, and Homogenization
AI Actors: 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	**Action**	**Risks**
MG-1.3-001	Allocate resources and time for GAI risk management activities, including planning for incident response and other mitigation activities.	
MG-1.3-002	Document residual GAI system risks that persist after risk mitigation or transfer.	
MG-1.3-003	Document trade-offs, decision processes, and relevant measurement and feedback results for risks that do not surpass organizational risk tolerance.	
MG-1.3-004	Mitigate, transfer, or avoid risks that surpass organizational risk tolerances.	
MG-1.3-005	Monitor the effectiveness of risk controls (e.g., via field testing, participatory engagements, performance assessments, user feedback mechanisms).	Human AI Configuration
AI Actors: 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	**Action**	**Risks**
MG-2.2-001	Compare GAI system outputs against pre-defined organization risk tolerance, guidelines, and principles, and review and audit AI-generated content against these guidelines.	
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 make updates where needed. Implement real-time monitoring systems to detect GAI systems and content provenance drift as it happens.	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, Toxicity, Bias, and Homogenization
MG-2.2-005	Filter GAI output for harmful or biased content, potential misinformation, and CBRN-related or NCII content.	CBRN Information, Obscene, Degrading, and/or Abusive Content, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MG-2.2-006	Implement version control for models and datasets to track changes and facilitate rollback if necessary.	
MG-2.2-007	Incorporate feedback from users, external experts, and the public to adapt the GAI system and monitoring processes.	Human AI Configuration
MG-2.2-008	Incorporate human review processes to assess and filter content in accordance with the socio-cultural knowledge and values of the context of use and to identify limitations and nuances that automated processes might miss; verify that human reviewers are trained on content guidelines and potential biases of GAI system and its content provenance.	Information Integrity, Toxicity, Bias, and Homogenization
MG-2.2-009	Integrate information from data management and machine learning security countermeasures like red teaming, and differential privacy, and authentication protocols to ensure data and models are protected from potential risks.	CBRN Information, Data Privacy, Information Security
MG-2.2-010	Use feedback from internal and external AI actors, users, individuals, and communities, to assess impact of AI-generated content.	Human AI Configuration
MG-2.2-011	Use real-time auditing tools such as distributed ledger technology to track and validate the lineage and authenticity of AI-generated data.	Information Integrity
MG-2.2-012	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, Toxicity, Bias, and Homogenization
AI Actors: 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	**Action**	**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 supply 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
MG-2.3-002	Maintain protocols to log changes made to GAI systems during incident response and recovery.	
MG-2.3-003	Review, update and maintain incident response and recovery plans to integrate insights from GAI system use cases and contexts and needs of relevant AI actors.	Human AI Configuration
MG-2.3-004	Verify and maintain measurements that GAI systems are operating within organizational risk tolerances post incident.	
AI Actors: 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	**Action**	**Risks**
MG-2.4-001	Enforce change management processes, and risk and impact assessments across all intended uses and contexts before deploying GAI system updates.	
MG-2.4-002	Establish and maintain communication plans to inform AI stakeholders as part of the deactivation or disengagement process of a specific GAI system or context of use, including reasons, workarounds, user access removal, alternative processes, contact information, etc.	Human AI Configuration
MG-2.4-003	Establish and maintain procedures for escalating GAI system incidents to the organizational risk authority when specific criteria for deactivation or disengagement is met for a particular context of use or for the GAI system as a whole.	

MG-2.4-004	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.	
MG-2.4-005	Establish and regularly review specific criteria that warrants the deactivation of GAI systems in accordance with set risk tolerances and appetites.	
AI Actors: 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	**Action**	**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) 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
MG-3.1-002	Audit GAI system supply 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, Toxicity, Bias, and Homogenization
MG-3.1-003	Decommission third-party systems that exceed organizational risk tolerances.	Value Chain and Component Integration
MG-3.1-004	Identify and maintain documentation for third-party AI systems, and components, in organizational AI inventories.	Value Chain and Component Integration
MG-3.1-005	Initiate review of third-party organizations/developers prior to their use of GAI models, and during their use of GAI models for their own applications, to monitor for abuse and policy violations.	Value Chain and Component Integration, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations
MG-3.1-006	Re-assess model risks after fine-tuning 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-3.1-007	Review GAI training data for CBRN information and intellectual property; scan output for plagiarized, trademarked, patented, licensed, or trade secret material.	Intellectual Property, CBRN Information
MG-3.1-008	Update acquisition and procurement policies, procedures, and processes to address GAI risks and failure modes.	
MG-3.1-009	Use, review, update, and share various transparency artifacts (e.g., system cards and model cards) for third-party models. Document or retain documentation for: Training data content and provenance, methodology, testing, validation, and clear instructions for use from GAI vendors and suppliers, Information related to third-party information security policies, procedures, and processes.	Information Integrity, Information Security, Value Chain and Component Integration
AI Actors: 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	**Action**	**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.	
MG-3.2-002	Document how pre-trained models have been adapted (fine-tuned) for the specific generative task, including any data augmentations, parameter adjustments, or other modifications. Access to un-tuned (baseline) models must be available to support debugging the relative influence of the pre-trained weights compared to the fine-tuned model weights.	
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 processes applied.	Information Integrity, Toxicity, Bias, and Homogenization
MG-3.2-004	Evaluate user reported problematic content and integrate feedback into system updates.	Human AI Configuration, Dangerous or Violent Recommendations

MG-3.2-005	Implement content filters to prevent the generation of inappropriate, harmful, toxic, 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, Toxicity, Bias, and Homogenization, Dangerous or Violent Recommendations, 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	Maintain awareness of relevant laws and regulations related to content generation, data privacy, and user protections and work in conjunction with legal experts to review and assess the potential liabilities associated with AI-generated content.	Data Privacy, Intellectual Property Information Integrity
MG-3.2-009	Provide use case examples as material for training employees and stakeholders about the trustworthiness implications of GAI applications and content provenance and to raise awareness about potential risks in fostering a risk management culture.	Information Integrity
MG-3.2-010	Use human moderation systems 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-3.2-011	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, Confabulation
AI Actors: 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	**Action**	**Risks**
MG-4.1-001	Collaborate with external researchers, industry experts, and community representatives to maintain awareness of emerging best practices and technologies in content provenance.	Information Integrity, Toxicity, Bias, and Homogenization
MG-4.1-002	Conduct adversarial testing at a regular cadence; test against various adversarial inputs and scenarios; identify vulnerabilities and assess the AI system's resilience to content provenance attacks.	Information Integrity, Information Security
MG-4.1-003	Conduct red-teaming exercises to surface failure modes of content provenance mechanisms. Evaluate the effectiveness of red-teaming approaches for uncovering potential vulnerabilities and improving overall content provenance.	Information Integrity, Information Security
MG-4.1-004	Employ user-friendly channels such as feedback forms, e-mails, or hotlines for users to report issues, concerns, or unexpected GAI outputs to feed into monitoring practices.	Human AI Configuration
MG-4.1-005	Establish, maintain, and evaluate effectiveness of organizational processes and procedures to monitor GAI systems within context of use.	
MG-4.1-006	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-4.1-007	Implement active learning techniques to identify instances where the model fails or produces unexpected outputs.	Confabulation
MG-4.1-008	Integrate digital watermarks, blockchain technology, cryptographic hash functions, metadata embedding, or other content provenance techniques within AI-generated content to track its source and manipulation history.	Information Integrity
MG-4.1-009	Measure system outputs related to content provenance at a regular cadence and integrate insights into monitoring processes.	Information Integrity
MG-4.1-010	Monitor GAI training data for representation of different user groups.	Human AI Configuration, Toxicity, Bias, and Homogenization
MG-4.1-011	Perform periodic review of organizational adherence to GAI system monitoring plans across all contexts of use.	

MG-4.1-012	Share transparency reports with internal and external stakeholders that detail steps taken to update the AI system to enhance transparency and accountability.	
MG-4.1-013	Track dataset modifications for content provenance by monitoring data deletions, rectification requests, and other changes that may impact the verifiability of content origins.	Information Integrity
MG-4.1-014	Verify risks associated with gaps in GAI system monitoring plans are accepted at the appropriate organizational level.	
MG-4.1-015	Verify that AI actors responsible for monitoring reported issues can effectively evaluate GAI system performance and its content provenance, and promptly escalate issues for response.	Human AI Configuration, Information Integrity
AI Actors: 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	**Action**	**Risks**
MG-4.2-001	Adopt agile development methodologies, and iterative development and feedback loops to allow for rapid adjustments based on external input related to content provenance.	Information Integrity
MG-4.2-002	Conduct regular audits of GAI systems and publish reports detailing the performance, feedback received, and improvements made.	
MG-4.2-003	Employ explainable AI methods to enhance transparency and interpretability of GAI content provenance to help AI actors and stakeholders understand how and why specific content is generated.	Human AI Configuration, Information Integrity
MG-4.2-004	Employ stakeholder feedback captured in the Map function to understand user experiences and perceptions about AI-generated content and its provenance; include user interactions and feedback from real-world scenarios.	Human AI Configuration, Information Integrity
MG-4.2-005	Form cross-functional teams leveraging expertise from across the AI lifecycle including AI designers and developers, socio-technical experts, and experts in the context of use and identify mechanisms to include end users in consultations.	Human AI Configuration

MG-4.2-006	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 or Violent Recommendations
MG-4.2-007	Provide external stakeholders with regular updates about the progress, challenges, and improvements made based on their feedback through the use of public venues such as online platforms and communities, and open-source initiatives.	Intellectual Property
MG-4.2-008	Simulate various scenarios to test GAI system responses and verify intended performance across different situations.	
MG-4.2-009	Use visualizations to represent the GAI model behavior to ease non-technical stakeholders understanding of GAI system functionality.	Human-AI Configuration
AI Actors: 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	**Action**	**Risks**
MG-4.3-001	Conduct after-action assessments for GAI system incidents to verify incident response and recovery processes are followed and effective.	
MG-4.3-002	Establish and maintain change management records and procedures for GAI systems, including the reasons for each change, how the change could impact each intended context of use, and step-by-step details of how changes were planned, tested, and deployed.	
MG-4.3-003	Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, incidents, and negative impacts.	Confabulation, Information Integrity
MG-4.3-004	Establish processes and procedures for regular sharing of information about errors, incidents, and negative impacts for each and across contexts, sectors, and AI actors, including the date reported, the context of use, the number of reports for each issue, and assessments of impact and severity.	Confabulation, Human AI Configuration, Information Integrity
AI Actors: 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 to 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, Mat Heyman, Yooyoung Lee, Reva Schwartz, and Kyra Yee.

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

Appendix B. 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.

Appendix C. 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.

C.1.1.1. 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.

Appendix D. 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.

Appendix E. AI Red-teaming

AI red-teaming exercises are often conducted in a controlled environment and in collaboration with AI developers building AI models. 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-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

Appendix F. Overview

GAI technologies can be leveraged for many applications such as content generation and synthetic data. Some aspects of GAI output, such as the production of deepfake content, can challenge our ability to distinguish human-generated content from AI-generated 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 provide greater information about both authentic and synthetic content to users, enabling trustworthiness in AI systems. When combined with other organizational accountability mechanisms, digital content transparency 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 of content and its history to assist in GAI risk management efforts.

Provenance data can include information about generated content's creators, date/time of creation, location, modifications, and sources, including metadata information. Metadata can be tracked for text, images, videos, audio, and underlying datasets. Provenance data tracking employs various methods and metrics to assess the authenticity, integrity, credibility, intellectual property rights, and potential manipulations in GAI output. Some well-known techniques for provenance data tracking include watermarking, metadata tracking, digital fingerprinting, and human authentication, among others.

Appendix G. Provenance Data Tracking Approaches

Provenance data tracking techniques for GAI systems can be used to track the lineage and integrity of data inputs, metadata, and AI-generated 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 perceptible and imperceptible digital watermarks on digital 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 given model to prioritize computational efficiency over accuracy, an AI actor may inadvertently affect provenance tracking reliability. 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 perceptible 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 watermarking techniques to identify the source of content or metadata management 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 of applications involving AI-generated content can help enhance awareness of performance changes and mitigate potential risks and harms. There are many ways to capture and make use of user feedback – before and after GAI systems are deployed – to gain insights about authentication efficacy and vulnerabilities, impacts of adversarial threats, unintended consequences resulting from the utilization of content provenance approaches, and other unanticipated behavior associated with content manipulation. 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

Appendix H. Overview

AI incidents can be defined as an event, circumstance, or series of events in which the development, use, or malfunction of one or more AI systems directly or indirectly contributes to identified harms. These harms include injury or damage to the health of an individual or group of people; disruption of the management and operation of critical infrastructure; violations of human rights or a breach of obligations under applicable law intended to protect legal and labor rights; or damage 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).

Appendix I. 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.

Appendix J. 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 K. References

AI Risks and Trustworthiness, NIST Trustworthy & Responsible AI Resource Center. *National Institute of Standards and Technology*. https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/3-sec-characteristics.

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

AI RMF Profiles. *National Institute of Standards and Technology*. https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Core_And_Profiles/6-sec-profile.

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

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

AI Risk Management Framework. *National Institute of Standards and Technology*. Appendix A: Descriptions of AI Actor Tasks, NIST Trustworthy & Responsible AI Resource Center. *National Institute of Standards and Technology.* 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.

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

Alba, D., (2023) How Fake AI Photo of a Pentagon Blast Went Viral and Briefly Spooked Stocks. *Bloomberg*. https://www.bloomberg.com/news/articles/2023-05-22/fake-ai-photo-of-pentagon-blast-goes-viral-trips-stocks-briefly.

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/.

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

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.

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

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 CSET*. https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/.

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.

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.

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023) *The White House*.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/.

Fair Information Practice Principles (FIPPs). *FPC*. https://www.fpc.gov/resources/fipps/.

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

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

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

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.

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

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

Harwell, D., (2023) AI-generated child sex images spawn new nightmare for the web. *Washington Post*. https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/.

Hubinger, E. et al, (2024) "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training", *arXiv e-prints*. https://arxiv.org/abs/2401.05566.

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

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/.

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.

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.

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/.

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 AI Incidents Monitor. OECD.AI *Policy Observatory*. https://oecd.ai/en/incidents-methodology.

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

Paresh, D., (2023) ChatGPT Is Cutting Non-English Languages Out of the AI Revolution. *WIRED*. https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/.

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.

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

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/.

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.

Skaug Sætra, H. et al., (2022). Psychological interference, liberty and technology. *Technology in Society.* https://www.sciencedirect.com/science/article/pii/S0160791X22001142.

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.

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.

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.

The Toxicity Issue. *Jigsaw, Google*. https://current.withgoogle.com/the-current/toxicity/.

Tufekci, Z. (2015) Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency. 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, Y. et al., (2023) Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv*. https://arxiv.org/pdf/2308.13387.

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

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.

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.

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

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.