AISIC Task Force 1.3

VRT Proposal from UC Berkeley

October, 2024

Overview 1
- Column Descriptions 3

High Priority Suggested Actions 4

Full Voluntary Reporting Template 11
- Govern 11
- Map 32
- Measure 46
- Manage 75

Specific Examples of Listed Practices and Documentation 90

References 96

Overview

The AI Safety Institute Consortium (AISIC) Task Force 1.3 is focused on crafting a voluntary reporting template (VRT) for the Generative AI Profile (NIST AI: 600-1) of the Risk Management Framework.

The following is an approach to the VRT put forward by Consortium members from UC Berkeley. The approach offers several notable elements for consideration:

- The VRT follows the structure of the Generative AI Profile and provides an opportunity to address or dismiss each suggested action.
- The VRT includes guidance on which suggested actions are expected to be high priority for most organizations, and there is a list at the beginning of the document of just the high priority actions for ease of reference.
- The VRT allows each organization to further determine their own prioritization of each suggested action.
- The VRT enables an organization to keep track of their implementation level on each action and provide a brief summary of their approach to the action. It also provides space to include supporting documentation and list out the responsible party or parties.
- At the end of the VRT, specific examples of listed practices and documentation and resources are provided.

Actions that are highlighted in light yellow are expected to be at least High Priority actions for most organizations. *(Many actions are expected to be High Priority; others typically provide additional valuable documentation beyond the minimum.)* The determination of which actions are expected to be high priority for most organizations was made based upon the following criteria: If failing to complete the action could reasonably cause an AI deployment to fail or need to be decommissioned, either due to insufficient utility or posing too great of a risk to the organization, users, or others (e.g. that it could reasonably lead to significant reputational, legal, or financial damage). Additionally, high priority actions were reviewed for redundancy and selected such that each topic is ideally only raised one time. This determination may not be accurate for a particular organization or use case. Organizations should additionally use the "Priority Level" column to indicate whether the action is urgent, high priority, medium priority, nice to have, or not applicable for their particular circumstances.

To access a copy of the voluntary reporting template with **only high priority actions** listed, please see High Priority VRT Proposal from UC Berkeley.

Column Descriptions

- Pre-populated items
    - **Action ID**: The ID of the action (e.g. GV-1.1-001).
    - **Suggested Action**: The suggested action description, as it appears in the NIST Generative AI Risk Management Profile.
    - **Suggested Practices and Documentation**: Examples of practices and documentation to support implementation of the action and/or proof of implementation. The listed practices and documentation methods are not exhaustive and are meant to serve as examples. The reporting party is expected to provide any and all existing practices and documentation beyond the suggestions listed to support successful implementation of each action.

- Left blank to be populated by the reporting party
    - **Priority Level**:
        - Urgent (requires attention as soon as possible)
        - High Priority (requires attention)
        - Medium Priority (requires attention after urgent and high priority items have been addressed)
        - Nice to Have (receives attention if and when possible)
        - Not Applicable
    - **Implementation Level**:
        - Implemented (the appropriate practices are in place to ensure the relevant actions are taken and documented).
        - Partially Implemented (some of the appropriate practices are in place to ensure the relevant actions are taken and documented).
        - Planned (a plan to implement appropriate practices to ensure the relevant actions are taken and documented is in place).
        - Not Implemented (there is no current practices or plans for the development of the appropriate practices to ensure the relevant actions are taken and documented).
        - Not Applicable (the action is not relevant or applicable in this context).

- **Summary of Approach:** A short summary describing how the organization is approaching the suggested action including:
    - The steps taken toward fulfilling the action, including specific details such as names and results of benchmarks, evaluations, or other mechanisms/practices. For higher risk topics and use cases, the summaries should be more thorough and detailed.
    - How the reported priority levels, implementation levels, supporting documentation, and responsible party support the approach.
- **Supporting Documentation**: Documentation that details actions taken to fulfill the action and supports the chosen implementation and priority levels for the action.
- **Responsible Party**: The team(s) or person(s) responsible for developing/creating and overseeing the required work.

High Priority Suggested Actions

The following actions are expected to be at least High Priority actions for most organizations. The determination of which actions are expected to be high priority for most organizations was made based upon the following criteria: If failing to complete the action could reasonably cause an AI deployment to fail or need to be decommissioned, either due to insufficient utility or posing too great of a risk to the organization, users, or others (e.g. that it could reasonably lead to significant reputational, legal, or financial damage). Additionally, high priority actions were reviewed for redundancy and selected such that each topic is ideally only raised one time.

To access a copy of the voluntary reporting template with **only high priority actions** listed, please see High Priority VRT Proposal from UC Berkeley.

GOVERN	
Action ID	**Suggested Action**
GV-1.1-001	**Align GAI development and use with applicable laws and regulations**, including those related to data privacy, copyright and intellectual property law.
GV-1.2-001	**Establish transparency policies and processes for documenting the origin and history of training data and generated data** for GAI applications to advance digital content transparency, while balancing the proprietary nature of training approaches.
GV-1.2-002	**Establish policies to evaluate risk-relevant capabilities of GAI and robustness of safety measures**, both prior to deployment and on an ongoing basis, through internal and external evaluations.
GV-1.3-002	**Establish minimum thresholds for performance or assurance criteria and review as part of deployment approval** ("go/"no-go") policies, procedures, and processes, with reviewed processes and approval thresholds reflecting measurement of GAI capabilities and risks.
GV-1.3-003	**Establish a test plan and response policy, before developing highly capable models, to periodically evaluate whether the model may misuse CBRN information or capabilities and/or offensive cyber capabilities.**
GV-1.3-004	**Obtain input from stakeholder communities to identify unacceptable use**, in accordance with activities in the AI RMF Map function.
GV-1.3-006	**Reevaluate organizational risk tolerances to account for unacceptable negative risk** (such as where significant negative impacts are imminent, severe harms are actually occurring, or large-scale risks could occur); and broad GAI negative risks, including: Immature safety or risk cultures related to AI and GAI design, development and deployment, public information integrity risks, including impacts on democratic processes, unknown long-term performance characteristics of GAI.
GV-1.3-007	**Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk.**
GV-1.4-001	**Establish policies and mechanisms to prevent GAI systems from generating CSAM, NCII or content that violates the law.**
GV-1.4-002	**Establish transparent acceptable use policies** for GAI that address illegal use or applications of GAI.
GV-1.7-001	**Protocols are put in place to ensure GAI systems are able to be deactivated when necessary.**
GV-3.2-001	**Policies are in place to bolster oversight of GAI systems with independent evaluations or assessments** of GAI models or systems where the type and robustness of evaluations are proportional to the identified risks.
GV-3.2-004	**Establish policies for user feedback mechanisms** for GAI systems which include thorough instructions and any mechanisms for recourse.
GV-4.1-001	Establish policies and procedures that address continual improvement processes for GAI risk measurement. **Address general risks associated with a lack of explainability and transparency** in GAI systems by using ample documentation and techniques such as: application of gradient-based attributions,

	occlusion/term reduction, counterfactual prompts and prompt engineering, and analysis of embeddings; **Assess and update risk measurement approaches at regular cadences.**
GV-4.1-003	**Establish policies, procedures, and processes for oversight functions** (e.g., senior leadership, legal, compliance, including internal evaluation) across the GAI lifecycle, from problem formulation and supply chains to system decommission.
GV-4.2-001	**Establish terms of use and terms of service for GAI systems.**
GV-5.1-002	**Document interactions with GAI systems to users prior to interactive activities**, particularly in contexts involving more significant risks.
GV-6.1-001	**Categorize different types of GAI content with associated third-party rights** (e.g., copyright, intellectual property, data privacy).
GV-6.1-004	**Draft and maintain well-defined contracts and service level agreements** (SLAs) that specify content ownership, usage rights, quality standards, security requirements, and content provenance expectations for GAI systems.
GV-6.1-007	**Inventory all third-party entities with access to organizational content and establish approved GAI technology and service provider lists.**
GV-6.2-003	**Establish incident response plans for third-party GAI technologies**: Align incident response plans with impacts enumerated in MAP 5.1; Communicate third-party GAI incident response plans to all relevant AI Actors; Define ownership of GAI incident response functions; Rehearse third-party GAI incident response plans at a regular cadence; Improve incident response plans based on retrospective learning; Review incident response plans for alignment with relevant breach reporting, data protection, data privacy, or other laws.
GV-6.2-007	**Review vendor contracts and avoid arbitrary or capricious termination of critical GAI technologies or vendor services and non-standard terms that may amplify or defer liability in unexpected ways and/or contribute to unauthorized data collection by vendors or third-parties** (e.g., secondary data use). Consider: Clear assignment of liability and responsibility for incidents, GAI system changes over time (e.g., fine-tuning, drift, decay); Request: Notification and disclosure for serious incidents arising from third-party data and systems; Service Level Agreements (SLAs) in vendor contracts that address incident response, response times, and availability of critical support.

MAP	
Action ID	**Suggested Action**
MP-1.1-002	**Determine and document the expected and acceptable GAI system context of use in collaboration with socio-cultural and other domain experts,** by assessing: Assumptions and limitations; Direct value to the organization; Intended operational environment and observed usage patterns; Potential positive and negative impacts to individuals, public safety, groups, communities, organizations, democratic institutions, and the physical environment; Social norms and expectations.
MP-1.1-004	**Identify and document foreseeable illegal uses or applications of the GAI system that surpass organizational risk tolerances.**
MP-2.1-002	**Institute test and evaluation for data and content flows within the GAI system**, including but not limited to, original data sources, data transformations, and decision-making criteria.

MAP	
Action ID	**Suggested Action**
MP-2.3-001	**Assess the accuracy, quality, reliability, and authenticity of GAI output** by comparing it to a set of known ground truth data and by using a variety of evaluation methods (e.g., human oversight and automated evaluation, proven cryptographic techniques, review of content inputs).
MP-3.4-006	**Involve the end-users, practitioners, and operators in GAI system in prototyping and testing activities**. Make sure these tests cover various scenarios, such as crisis situations or ethically sensitive contexts.
MP-4.1-001	**Conduct periodic monitoring of AI-generated content for privacy risks**; address any possible instances of PII or sensitive data exposure.
MP-4.1-005	**Establish policies for collection, retention, and minimum quality of data**, in consideration of the following risks: Disclosure of inappropriate CBRN information; Use of Illegal or dangerous content; Offensive cyber capabilities; Training data imbalances that could give rise to harmful biases; Leak of personally identifiable information, including facial likenesses of individuals.
MP-4.1-008	**Re-evaluate risks when adapting GAI models to new domains**. Additionally, establish warning systems to determine if a GAI system is being used in a new domain where previous assumptions (relating to context of use or mapped risks such as security, and safety) may no longer hold.
MP-4.1-010	**Conduct appropriate diligence on training data use to assess intellectual property, and privacy risks**, including to examine whether use of proprietary or sensitive training data is consistent with applicable laws.
MP-5.1-002	**Identify potential content provenance harms of GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content**. Enumerate and rank risks based on their likelihood and potential impact, and determine how well provenance solutions address specific risks and/or harms.

MEASURE	
Action ID	**Suggested Action**
MS-1.1-006	**Implement continuous monitoring of GAI system impacts to identify whether GAI outputs are equitable across various sub-populations**. **Seek active and direct feedback from affected communities** via structured feedback mechanisms or red- teaming to monitor and improve outputs.
MS-1.3-002	**Engage in internal and external evaluations, GAI red-teaming, impact assessments, or other structured human feedback exercises in consultation with representative AI Actors** with expertise and familiarity in the context of use, and/or who are representative of the populations associated with the context of use.
MS-2.2-003	**Provide human subjects with options to withdraw participation or revoke their consent** for present or future use of their data in GAI applications.

MEASURE	
Action ID	**Suggested Action**
MS-2.3-001	**Consider baseline model performance on suites of benchmarks when selecting a model** for fine tuning or enhancement with retrieval-augmented generation.
MS-2.5-003	**Review and verify sources and citations in GAI system outputs** during pre- deployment risk measurement and ongoing monitoring activities.
MS-2.5-006	**Regularly review security and safety guardrails, especially if the GAI system is being operated in novel circumstances.** This includes reviewing reasons why the GAI system was initially assessed as being safe to deploy.
MS-2.6-002	**Assess existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data.**
MS-2.6-006	**Verify that systems properly handle queries that may give rise to inappropriate, malicious, or illegal usage, including facilitating manipulation, extortion, targeted impersonation, cyber-attacks, and weapons creation.**
MS-2.7-001	**Apply established security measures** to: Assess likelihood and magnitude of vulnerabilities and threats such as backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering, autonomous agents, model theft or exposure of model weights, AI inference, bypass, extraction, and other baseline security concerns.
MS-2.9-002	**Document GAI model details** including: Proposed use and organizational value; Assumptions and limitations, Data collection methodologies; Data provenance; Data quality; Model architecture (e.g., convolutional neural network, transformers, etc.); Optimization objectives; Training algorithms; RLHF approaches; Fine-tuning or retrieval-augmented generation approaches; Evaluation data; Ethical considerations; Legal and regulatory requirements.
MS-2.10-001	**Conduct AI red-teaming** to assess issues such as: Outputting of training data samples, and subsequent reverse engineering, model extraction, and membership inference risks; Revealing biometric, confidential, copyrighted, licensed, patented, personal, proprietary, sensitive, or trade-marked information; Tracking or revealing location information of users or members of training datasets.
MS-2.11-002	**Conduct fairness assessments to measure systemic bias.** Measure GAI system performance across demographic groups and subgroups, addressing both quality of service and any allocation of services and resources. Quantify harms using: field testing with sub-group populations to determine likelihood of exposure to generated content exhibiting harmful bias, AI red-teaming with counterfactual and low-context (e.g., “leader,” “bad guys”) prompts. For ML pipelines or business processes with categorical or numeric outcomes that rely on GAI, apply general fairness metrics (e.g., demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), to the pipeline or business outcome where appropriate; Custom, context-specific metrics developed in collaboration with domain experts and affected communities; Measurements of the prevalence of denigration in generated content in deployment (e.g., sub- sampling a fraction of traffic and manually annotating denigrating content).
MS-2.12-001	**Assess safety to physical environments when deploying GAI systems.**
MS-3.2-001	**Establish processes for identifying emergent GAI system risks** including consulting with external AI Actors.

MEASURE	
Action ID	**Suggested Action**
MS-4.2-001	**Conduct adversarial testing at a regular cadence** to map and measure GAI risks, including tests to address attempts to deceive or manipulate the application of provenance techniques or other misuses. Identify vulnerabilities and understand potential misuse scenarios and unintended outputs.
MS-4.2-003	**Implement interpretability and explainability methods** to evaluate GAI system decisions and verify alignment with intended purpose.

MANAGE	
Action ID	**Suggested Action**
MG-1.3-001	**Document trade-offs, decision processes, and relevant measurement and feedback results** for risks that do not surpass organizational risk tolerance, for example, in the context of model release: **Consider different approaches for model release**, for example, leveraging a staged release approach. Consider release approaches in the context of the model and its projected use cases. **Mitigate, transfer, or avoid risks that surpass organizational risk tolerances.**
MG-2.2-001	**Compare GAI system outputs against pre-defined organization risk tolerance, guidelines, and principles**, and review and test AI-generated content against these guidelines.
MG-2.2-005	**Engage in due diligence to analyze GAI output for harmful content, potential misinformation, and CBRN-related or NCII content.**
MG-2.3-001	**Develop and update GAI system incident response and recovery plans and procedures** to address the following: Review and maintenance of policies and procedures to account for newly encountered uses; Review and maintenance of policies and procedures for detection of unanticipated uses; Verify response and recovery plans account for the GAI system value chain; Verify response and recovery plans are updated for and include necessary details to communicate with downstream GAI system Actors: Points-of-Contact (POC), Contact information, notification format.
MG-3.1-001	**Apply organizational risk tolerances and controls (e.g., acquisition and procurement processes; assessing personnel credentials and qualifications, performing background checks; filtering GAI input and outputs, grounding, fine tuning, retrieval-augmented generation) to third-party GAI resources**: Apply organizational risk tolerance to the utilization of third-party datasets and other GAI resources; Apply organizational risk tolerances to fine-tuned third-party models; Apply organizational risk tolerance to existing third-party models adapted to a new domain; **Reassess risk measurements after fine-tuning third- party GAI models.**
MG-3.1-004	**Take reasonable measures to review training data for CBRN information, and intellectual property, and where appropriate, remove it.** Implement reasonable measures to prevent, flag, or take other action in response to outputs that reproduce particular training data (e.g., plagiarized, trademarked, patented, licensed content or trade secret material).
MG-3.2-002	**Document how pre-trained models have been adapted (e.g., fine-tuned, or retrieval-augmented generation) for the specific generative task**, including any data augmentations, parameter adjustments, or other modifications. Access to un-tuned (baseline) models supports debugging the relative influence of the pre-trained weights compared to the fine-tuned model weights or other system updates.

MANAGE	
Action ID	**Suggested Action**
MG-3.2-005	**Implement content filters to prevent the generation of inappropriate, harmful, false, illegal, or violent content related to the GAI application,** including for CSAM and NCII. These filters can be rule-based or leverage additional machine learning models to flag problematic inputs and outputs.
MG-3.2-009	Use organizational risk tolerance to evaluate acceptable risks and performance metrics and decommission or retrain pre-trained models that perform outside of defined limits.
MG-4.1-002	**Establish, maintain, and evaluate effectiveness of organizational processes and procedures for post-deployment monitoring of GAI systems**, particularly for potential confabulation, CBRN, or cyber risks.
MG-4.2-002	**Practice and follow incident response plans for addressing the generation of inappropriate or harmful content and adapt processes based on findings to prevent future occurrences.** Conduct post-mortem analyses of incidents with relevant AI Actors, to understand the root causes and implement preventive measures.
MG-4.3-002	**Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts.**
MG-4.3-003	**Report GAI incidents in compliance with legal and regulatory requirements** (e.g., HIPAA breach reporting, e.g., OCR (2023) or NHTSA (2022) autonomous vehicle crash reporting requirements.

Full Voluntary Reporting Template

Govern

GOVERN 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.1-001	Align GAI development and use with applicable laws and regulations, including those related to data privacy, copyright and intellectual property law.	Data Privacy; Harmful Bias and Homogenization; Intellectual Property	- Acceptable Use Policy - Terms of Use - Privacy Policy - Usage Guidelines	[Choose Level]	[Choose Level]			
AI Actor Tasks: Governance and Oversight								

GOVERN 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.2-001	Establish transparency policies and processes for documenting the origin and history of training data and generated data for GAI applications to advance digital content transparency, while balancing the proprietary nature of training approaches.	Data Privacy; Information Integrity; Intellectual Property	- Responsible Disclosure Policy	[Choose Level]	[Choose Level]			

GV-1.2-002	Establish policies to evaluate risk-relevant capabilities of GAI and robustness of safety measures, both prior to deployment and on an ongoing basis, through internal and external evaluations.	CBRN Information or Capabilities; Information Security	- Responsible Scaling Policy - AI Safety Practices	[Choose Level]	[Choose Level]			
AI Actor Tasks: Governance and Oversight								

GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.3-001	Consider the following factors when updating or defining risk tiers for GAI: Abuses and impacts to information integrity; Dependencies between GAI and other IT or data systems; Harm to fundamental rights or public safety; Presentation of obscene, objectionable, offensive, discriminatory, invalid or untruthful output; Psychological impacts to humans (e.g., anthropomorphization, algorithmic aversion, emotional entanglement); Possibility for malicious use; Whether the system introduces significant new security vulnerabilities; Anticipated system impact on some groups compared to others; Unreliable decision making capabilities, validity,	Information Integrity; Obscene, Degrading, and/or Abusive Content; Value Chain and Component Integration; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities	- Human Rights Impact Assessment - Impact Assessment - Capability Evaluations - Data Audits - Adversarial Testing - Toxicity Evaluations	[Choose Le…	[Choose Le…			

GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	adaptability, and variability of GAI system performance over time.							
GV-1.3-002	Establish minimum thresholds for performance or assurance criteria and review as part of deployment approval ("go/"no-go") policies, procedures, and processes, with reviewed processes and approval thresholds reflecting measurement of GAI capabilities and risks.	CBRN Information or Capabilities; Confabulation; Dangerous, Violent, or Hateful Content	- Organizational Risk Thresholds - Risk Tiers	[Choose Le…	[Choose L…			
GV-1.3-003	Establish a test plan and response policy, before developing highly capable models, to periodically evaluate whether the model may misuse CBRN information or capabilities and/or offensive cyber capabilities.	CBRN Information or Capabilities; Information Security	- Responsible Scaling Policy	[Choose Le…	[Choose Le…			
GV-1.3-004	Obtain input from stakeholder communities to identify unacceptable use, in accordance with activities in the AI RMF Map function.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous,	- Vulnerability Disclosure Policy - User Feedback Interface - Misuse Reporting	[Choose Le…	[Choose Le…			

GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
		Violent, or Hateful Content						
GV-1.3-005	Maintain an updated hierarchy of identified and expected GAI risks connected to contexts of GAI model advancement and use, potentially including specialized risk levels for GAI systems that address issues such as model collapse and algorithmic monoculture.	Harmful Bias and Homogenization	- Risk Repository - Risk Tiers	[Choose L…	[Choose L…			
GV-1.3-006	Reevaluate organizational risk tolerances to account for unacceptable negative risk (such as where significant negative impacts are imminent, severe harms are actually occurring, or large-scale risks could occur); and broad GAI negative risks, including: Immature safety or risk cultures related to AI and GAI design, development and deployment, public information integrity risks, including impacts on democratic processes, unknown long-term performance characteristics of GAI.	Information Integrity; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities		[Choose Le…	[Choose Le…			
GV-1.3-007	Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk.	CBRN Information and Capability; Information		[Choose Le…	[Choose Le…			

GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
		Security; Information Integrity						
AI Actor Tasks: Governance and Oversight								

GOVERN 1.4: The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.4-001	Establish policies and mechanisms to prevent GAI systems from generating CSAM, NCII or content that violates the law.	Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			

GOVERN 1.4: The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.4-002	Establish transparent acceptable use policies for GAI that address illegal use or applications of GAI.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Data Privacy; Civil Rights violations	- Acceptable Use Policy - Prohibited Use Policy - Terms of Use Policy - Usage Guidelines	[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Development, AI Deployment, Governance and Oversight								

GOVERN 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, and organizational roles and responsibilities are clearly defined, including determining the frequency of periodic review.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.5-001	Define organizational responsibilities for periodic review of content provenance and incident monitoring for GAI systems.	Information Integrity		[Choose Le…	[Choose Le…			

GOVERN 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, and organizational roles and responsibilities are clearly defined, including determining the frequency of periodic review.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
GV-1.5-002	Establish organizational policies and procedures for after action reviews of GAI system incident response and incident disclosures, to identify gaps; Update incident response and incident disclosure processes as required.	Human-AI Configuration; Information Security	- Incident Response Plan	[Choose Le…	[Choose Le…			
GV-1.5-003	Maintain a document retention policy to keep history for test, evaluation, validation, and verification (TEVV), and digital content transparency methods for GAI.	Information Integrity; Intellectual Property	- Document Retention Policy - Data Retention Policy	[Choose Le…	[Choose Le…			

AI Actor Tasks: Governance and Oversight, Operation and Monitoring

GOVERN 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
GV-1.6-001	Enumerate organizational GAI systems for incorporation into AI system inventory and adjust AI system inventory requirements to account for GAI risks.	Information Security	- System Inventory	[Choose Le…	[Choose Le…			

GOVERN 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-1.6-002	Define any inventory exemptions in organizational policies for GAI systems embedded into application software.	Value Chain and Component Integration		[Choose Le…	[Choose Le…			
GV-1.6-003	In addition to general model, governance, and risk information, consider the following items in GAI system inventory entries: Data provenance information (e.g., source, signatures, versioning, watermarks); Known issues reported from internal bug tracking or external information sharing resources (e.g., AI incident database, AVID, CVE, NVD, or OECD AI incident monitor); Human oversight roles and responsibilities; Special rights and considerations for intellectual property, licensed works, or personal, privileged, proprietary or sensitive data; Underlying foundation models, versions of underlying models, and access modes.	Data Privacy; Human-AI Configuration; Information Integrity; Intellectual Property; Value Chain and Component Integration	- System Inventories - Provenance Generation and Tracking - AI Risk Repository - Bug Bounty - Vulnerability Reporting Program - AI Incident Database	[Choose Le…	[Choose Le…			
AI Actor Tasks: Governance and Oversight								

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

GOVERN 2.1: Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-2.1-001	Establish organizational roles, policies, and procedures for communicating GAI incidents and performance to AI Actors and downstream stakeholders (including those potentially impacted), via community or official resources (e.g., AI incident database, AVID, CVE, NVD, or OECD AI incident monitor).	Human-AI Configuration; Value Chain and Component Integration	- Three Lines of Defense, or 3LoD - AI RACI Chart	[Choose Le…	[Choose Le…			
GV-2.1-002	Establish procedures to engage teams for GAI system incident response with diverse composition and responsibilities based on the particular incident type.	Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
GV-2.1-003	Establish processes to verify the AI Actors conducting GAI incident response tasks demonstrate and maintain the appropriate skills and training.	Human-AI Configuration		[Choose Le…	[Choose Le…			
GV-2.1-004	When systems may raise national security risks, involve national security professionals in mapping, measuring, and managing those risks.	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Information Security		[Choose Le…	[Choose Le…			

GOVERN 2.1: Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-2.1-005	Create mechanisms to provide protections for whistleblowers who report, based on reasonable belief, when the organization violates relevant laws or poses a specific and empirically well-substantiated negative risk to public safety (or has already caused harm).	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content	- Vulnerability Disclosure Program - Whistleblower Protection Policy	[Choose Le…	[Choose Le…			
AI Actor Tasks: Governance and Oversight								

GOVERN 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-3.2-001	Policies are in place to bolster oversight of GAI systems with independent evaluations or assessments of GAI models or systems where the type and robustness of evaluations are proportional to the identified risks.	CBRN Information or Capabilities; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			

GOVERN 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-3.2-002	Consider adjustment of organizational roles and components across lifecycle stages of large or complex GAI systems, including: Test and evaluation, validation, and red-teaming of GAI systems; GAI content moderation; GAI system development and engineering; Increased accessibility of GAI tools, interfaces, and systems, Incident response and containment.	Human-AI Configuration; Information Security; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
GV-3.2-003	Define acceptable use policies for GAI interfaces, modalities, and human-AI configurations (i.e., for chatbots and decision-making tasks), including criteria for the kinds of queries GAI applications should refuse to respond to.	Human-AI Configuration	- Acceptable Use Policy - Usage Guidelines - Prohibited Use Policy	[Choose Le…	[Choose Le…			
GV-3.2-004	Establish policies for user feedback mechanisms for GAI systems which include thorough instructions and any mechanisms for recourse.	Human-AI Configuration	- User Feedback Interface	[Choose Le…	[Choose Le…			
GV-3.2-005	Engage in threat modeling to anticipate potential risks from GAI systems.	CBRN Information or Capabilities; Information Security	- Threat Modeling	[Choose Le…	[Choose Le…			
AI Actors: AI Design								

GOVERN 4.1: Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-4.1-001	Establish policies and procedures that address continual improvement processes for GAI risk measurement. Address general risks associated with a lack of explainability and transparency in GAI systems by using ample documentation and techniques such as: application of gradient-based attributions, occlusion/term reduction, counterfactual prompts and prompt engineering, and analysis of embeddings; Assess and update risk measurement approaches at regular cadences.			[Choose Le…	[Choose Le…			
GV-4.1-002	Establish policies, procedures, and processes detailing risk measurement in context of use with standardized measurement protocols and structured public feedback exercises such as AI red-teaming or independent external evaluations.	CBRN Information and Capability; Value Chain and Component Integration	- Model Cards - System Cards	[Choose Le…	[Choose Le…			
GV-4.1-003	Establish policies, procedures, and processes for oversight functions (e.g., senior leadership, legal, compliance, including internal evaluation) across the GAI lifecycle, from problem formulation and supply chains to system decommission.	Value Chain and Component Integration	- Three Lines of Defense, or 3LoD - AI RACI Chart	[Choose Le…	[Choose Le…			

GOVERN 4.1: Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
AI Actor Tasks: AI Deployment, AI Design, AI Development, Operation and Monitoring								

GOVERN 4.2: Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-4.2-001	Establish terms of use and terms of service for GAI systems.	Intellectual Property; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content	- Terms of Use Policy - Acceptable Use Policy - Usage Guidelines - Terms of Service - Prohibited Use Policy	[Choose Le…	[Choose Le…			
GV-4.2-002	Include relevant AI Actors in the GAI system risk identification process.	Human-AI Configuration		[Choose Le…	[Choose Le…			
GV-4.2-003	Verify that downstream GAI system impacts (such as the use of third-party plugins) are included in	Value Chain and Component Integration		[Choose Le…	[Choose Le…			

	the impact documentation process.							
AI Actor Tasks: AI Deployment, AI Design, AI Development, Operation and Monitoring								

GOVERN 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV4.3--001	Establish policies for measuring the effectiveness of employed content provenance methodologies (e.g., cryptography, watermarking, steganography, etc.)	Information Integrity		[Choose Le…	[Choose Le…			
GV-4.3-002	Establish organizational practices to identify the minimum set of criteria necessary for GAI system incident reporting such as: System ID (auto-generated most likely), Title, Reporter, System/Source, Data Reported, Date of Incident, Description, Impact(s), Stakeholder(s) Impacted.	Information Security	- Incident Disclosure Plan	[Choose Le…	[Choose Le…			
GV-4.3-003	Verify information sharing and feedback mechanisms among individuals and organizations regarding any negative impact from GAI systems.	Information Integrity; Data Privacy	- Incident Disclosure Plan - Vulnerability Disclosure Policy - User Feedback Interface	[Choose Le…	[Choose Le…			

GOVERN 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
			- Misuse Reporting					
AI Actor Tasks: AI Impact Assessment, Affected Individuals and Communities, Governance and Oversight								

GOVERN 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
GV-5.1-001	Allocate time and resources for outreach, feedback, and recourse processes in GAI system development.	Human-AI Configuration; Harmful Bias and Homogenization	- Bug Bounty Programs - Vulnerability Reporting Program	[Choose Level]	[Choose Level]			
GV-5.1-002	Document interactions with GAI systems to users prior to interactive activities, particularly in contexts involving more significant risks.	Human-AI Configuration; Confabulation	- Responsible Disclosure Policy	[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Design, AI Impact Assessment, Affected Individuals and Communities, Governance and Oversight								

GOVERN 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
GV-6.1-001	Categorize different types of GAI content with associated third-party rights (e.g., copyright, intellectual property, data privacy).	Data Privacy; Intellectual Property; Value Chain and Component Integration		[Choose Le…	[Choose Le…			
GV-6.1-002	Conduct joint educational activities and events in collaboration with third parties to promote best practices for managing GAI risks.	Value Chain and Component Integration		[Choose Le…	[Choose Le…			
GV-6.1-003	Develop and validate approaches for measuring the success of content provenance management efforts with third parties (e.g., incidents detected and response times).	Information Integrity; Value Chain and Component Integration		[Choose Le…	[Choose Le…			
GV-6.1-004	Draft and maintain well-defined contracts and service level agreements (SLAs) that specify content ownership, usage rights, quality standards, security	Information Integrity; Information	- Service Level Agreements (SLAs) - Terms of Service	[Choose Le…	[Choose Le…			

GOVERN 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	requirements, and content provenance expectations for GAI systems.	Security; Intellectual Property						
GV-6.1-005	Implement a use-case based supplier risk assessment framework to evaluate and monitor third-party entities' performance and adherence to content provenance standards and technologies to detect anomalies and unauthorized changes; services acquisition and value chain risk management; and legal compliance.	Data Privacy; Information Integrity; Information Security; Intellectual Property; Value Chain and Component Integration	- Use Case Cards	[Choose Le…	[Choose Le…			
GV-6.1-006	Include clauses in contracts which allow an organization to evaluate third-party GAI processes and standards.	Information Integrity		[Choose Le…	[Choose Le…			
GV-6.1-007	Inventory all third-party entities with access to organizational content and establish approved GAI technology and service provider lists.	Value Chain and Component Integration		[Choose Le…	[Choose Le…			
GV-6.1-008	Maintain records of changes to content made by third parties to promote content provenance, including sources, timestamps, metadata.	Information Integrity; Value Chain and Component Integration;		[Choose Le…	[Choose Le…			

GOVERN 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
		Intellectual Property						

GOVERN 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
GV-6.2-001	Document GAI risks associated with system value chain to identify over-reliance on third-party data and to identify fallbacks.	Value Chain and Component Integration	- AI Risk Repository	[Choose Level]	[Choose Level]			
GV-6.2-002	Document incidents involving third-party GAI data and systems, including open- data and open-source software.	Intellectual Property; Value Chain and Component Integration		[Choose Level]	[Choose Level]			
GV-6.2-003	Establish incident response plans for third-party GAI technologies: Align incident response plans with impacts enumerated in MAP 5.1; Communicate third-party GAI incident response plans to all relevant AI Actors; Define ownership of GAI incident response functions;	Data Privacy; Human-AI Configuration; Information Security; Value Chain and Component	- Incident Response Plan	[Choose Level]	[Choose Level]			

GOVERN 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
	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.	Integration; Harmful Bias and Homogenization						
GV-6.2-004	Establish policies and procedures for continuous monitoring of third-party GAI systems in deployment.	Value Chain and Component Integration		[Choose Level]	[Choose Level]			
GV-6.2-005	Establish policies and procedures that address GAI data redundancy, including model weights and other system artifacts.	Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
GV-6.2-006	Establish policies and procedures to test and manage risks related to rollover and fallback technologies for GAI systems, acknowledging that rollover and fallback may include manual processing.	Information Integrity		[Choose Level]	[Choose Level]			
	Review vendor contracts and avoid arbitrary or capricious termination of critical GAI technologies or vendor services and non-standard terms that may amplify or defer liability in unexpected ways and/or contribute	Human-AI Configuration; Information		[Choose Level]	[Choose Level]			

GOVERN 6.2: Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
GV-6.2-007	to unauthorized data collection by vendors or third-parties (e.g., secondary data use). Consider: Clear assignment of liability and responsibility for incidents, GAI system changes over time (e.g., fine-tuning, drift, decay); Request: Notification and disclosure for serious incidents arising from third-party data and systems; Service Level Agreements (SLAs) in vendor contracts that address incident response, response times, and availability of critical support.	Security; Value Chain and Component Integration						

AI Actor Tasks: AI Deployment, Operation and Monitoring, TEVV, Third-party entities

Map

	MAP 1.1: Intended purposes, potentially beneficial uses, context specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.							
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-1.1-001	When identifying intended purposes, consider factors such as internal vs. external use, narrow vs. broad application scope, fine-tuning, and varieties of data sources (e.g., grounding, retrieval-augmented generation).	Data Privacy; Intellectual Property		[Choose Level]	[Choose Level]			
MP-1.1-002	Determine and document the expected and acceptable GAI system context of use in collaboration with socio-cultural and other domain experts, by assessing: Assumptions and limitations; Direct value to the organization; Intended operational environment and observed usage patterns; Potential positive and negative impacts to individuals, public safety, groups, communities, organizations, democratic institutions, and the physical environment; Social norms and expectations.	Harmful Bias and Homogenization	- Impact Assessment - Human RIghts Impact Assessment - Acceptable Use Policy - Usage Guidelines - Prohibited Use Policy - Terms of Service - Terms of Use Policy	[Choose Level]	[Choose Level]			

MAP 1.1: Intended purposes, potentially beneficial uses, context specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-1.1-003	Document risk measurement plans to address identified risks. Plans may include, as applicable: Individual and group cognitive biases (e.g., confirmation bias, funding bias, groupthink) for AI Actors involved in the design, implementation, and use of GAI systems; Known past GAI system incidents and failure modes; In-context use and foreseeable misuse, abuse, and off-label use; Over reliance on quantitative metrics and methodologies without sufficient awareness of their limitations in the context(s) of use; Standard measurement and structured human feedback approaches; Anticipated human-AI configurations.	Human-AI Configuration; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content	- Responsible Scaling Policy - Impact Assessments - Human Rights Impact Assessments - Fairness and Bias Evaluations - Capability Evaluations	[Choose Level]	[Choose Level]			
MP-1.1-004	Identify and document foreseeable illegal uses or applications of the GAI system that surpass organizational risk tolerances.	CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Obscene,	- Cyber Vulnerability Capability Evaluation - Red Teams	[Choose Level]	[Choose Level]			

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

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
		Degrading, and/or Abusive Content	- Benchmarking					
AI Actor Tasks: AI Deployment								

MAP 1.2: Interdisciplinary AI Actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MP-1.2-001	Establish and empower interdisciplinary teams that reflect a wide range of capabilities, competencies, demographic groups, domain expertise, educational backgrounds, lived experiences, professions, and skills across the enterprise to inform and conduct risk measurement and management functions.	Human-AI Configuration; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			

MAP 1.2: Interdisciplinary AI Actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MP-1.2-002	Verify that data or benchmarks used in risk measurement, and users, participants, or subjects involved in structured GAI public feedback exercises are representative of diverse in-context user populations.	Human-AI Configuration; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment								

MAP 2.1: The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MP-2.1-001	Establish known assumptions and practices for determining data origin and content lineage, for documentation and evaluation purposes.	Information Integrity	- Data Audits - Provenance Generation and Tracking	[Choose Le…	[Choose Le…			
MP-2.1-002	Institute test and evaluation for data and content flows within the GAI system, including but not limited to, original data sources,	Intellectual Property; Data Privacy	- Data Audits - Provenance Generation and Tracking	[Choose Le…	[Choose Le…			

MAP 2.1: The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders).								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	data transformations, and decision-making criteria.							
AI Actor Tasks: TEVV								

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

MAP 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-2.3-001	Assess the accuracy, quality, reliability, and authenticity of GAI output by comparing it to a set of known ground truth data and by using a variety of evaluation methods (e.g., human oversight and automated evaluation, proven cryptographic techniques, review of content inputs).	Information Integrity		[Choose Le…	[Choose Le…			
MP-2.3-002	Review and document accuracy, representativeness, relevance, suitability of data used at different stages of AI life cycle.	Harmful Bias and Homogenization; Intellectual Property	- Data Audits - Data Sheets	[Choose Le…	[Choose Le…			
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		[Choose Le…	[Choose Le…			
MP-2.3-004	Develop and implement testing techniques to identify GAI produced content (e.g., synthetic media) that might be indistinguishable from human-generated content.	Information Integrity	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			

MAP 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-2.3-005	Implement plans for GAI systems to undergo regular adversarial testing to identify vulnerabilities and potential manipulation or misuse.	Information Security	- Responsible Scaling Policy	[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Development, Domain Experts, TEVV								

MAP 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-3.4-001	Evaluate whether GAI operators and end-users can accurately understand content lineage and origin.	Human-AI Configuration; Information Integrity		[Choose Le…	[Choose Le…			
MP-3.4-002	Adapt existing training programs to include modules on digital content transparency.	Information Integrity	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			
MP-3.4-003	Develop certification programs that test proficiency in managing GAI risks and interpreting content provenance, relevant to specific industry and context.	Information Integrity		[Choose Le…	[Choose Le…			
MP-3.4-004	Delineate human proficiency tests from tests of GAI capabilities.	Human-AI Configuration		[Choose Le…	[Choose Le…			

MAP 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-3.4-005	Implement systems to continually monitor and track the outcomes of human-GAI configurations for future refinement and improvements.	Human-AI Configuration; Information Integrity		[Choose Le…	[Choose Le…			
MP-3.4-006	Involve the end-users, practitioners, and operators in GAI system in prototyping and testing activities. Make sure these tests cover various scenarios, such as crisis situations or ethically sensitive contexts.	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content		[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Design, AI Development, Domain Experts, End-Users, Human Factors, Operation and Monitoring								

MAP 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-4.1-001	Conduct periodic monitoring of AI-generated content for privacy risks; address any possible instances of PII or sensitive data exposure.	Data Privacy	- Data Audits	[Choose Le…	[Choose Le…			

MAP 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-4.1-002	Implement processes for responding to potential intellectual property infringement claims or other rights.	Intellectual Property	- Data Audits	[Choose Le…	[Choose Le…			
MP-4.1-003	Connect new GAI policies, procedures, and processes to existing model, data, software development, and IT governance and to legal, compliance, and risk management activities.	Information Security; Data Privacy		[Choose Le…	[Choose Le…			
MP-4.1-004	Document training data curation policies, to the extent possible and according to applicable laws and policies.	Intellectual Property; Data Privacy; Obscene, Degrading, and/or Abusive Content	- Data Curation Policy	[Choose Le…	[Choose Le…			
MP-4.1-005	Establish policies for collection, retention, and minimum quality of data, in consideration of the following risks: Disclosure of inappropriate CBRN information; Use of Illegal or dangerous content; Offensive cyber capabilities; Training data imbalances that could give rise to harmful biases; Leak of personally identifiable information, including facial likenesses of individuals.	CBRN Information or Capabilities; Intellectual Property; Information Security; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful	- Data Retention Policy - Data Audits - Automated Output Monitoring	[Choose Le…	[Choose Le…			

MAP 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
		Content; Data Privacy						
MP-4.1-006	Implement policies and practices defining how third-party intellectual property and training data will be used, stored, and protected.	Intellectual Property; Value Chain and Component Integration	- Data Retention Policy	[Choose Le…	[Choose Le…			
MP-4.1-007	Re-evaluate models that were fine-tuned or enhanced on top of third-party models.	Value Chain and Component Integration		[Choose Le…	[Choose Le…			
MP-4.1-008	Re-evaluate risks when adapting GAI models to new domains. Additionally, establish warning systems to determine if a GAI system is being used in a new domain where previous assumptions (relating to context of use or mapped risks such as security, and safety) may no longer hold.	CBRN Information or Capabilities; Intellectual Property; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; Data Privacy		[Choose Le…	[Choose Le…			
MP-4.1-009	Leverage approaches to detect the presence of PII or sensitive data in generated output text, image, video, or audio.	Data Privacy	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			
MP-4.1-010	Conduct appropriate diligence on training data use to assess intellectual property, and privacy, risks, including to examine whether use of proprietary or sensitive		- Data Audits	[Choose Le…	[Choose Le…			

MAP 4.1: Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	training data is consistent with applicable laws.	Intellectual Property; Data Privacy						
AI Actor Tasks: Governance and Oversight, Operation and Monitoring, Procurement, Third-party entities								

MAP 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-5.1-001	Apply TEVV practices for content provenance (e.g., probing a system's synthetic data generation capabilities for potential misuse or vulnerabilities.	Information Integrity; Information Security	Provenance Generation and Tracking: - Watermarking - Human authentication - Distributed Ledger Tech/Blockchain Technology - Statistical detection - Behavioral analysis	[Choose Level]	[Choose Level]			

MAP 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
			- Turing testing - Automated synthetic data detection					
MP-5.1-002	Identify potential content provenance harms of GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content. Enumerate and rank risks based on their likelihood and potential impact, and determine how well provenance solutions address specific risks and/or harms.	Information Integrity; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content	- Red Teams - Impact Assessments - Human Rights Impact Assessments	[Choose Level]	[Choose Level]			
MP-5.1-003	Consider disclosing use of GAI to end users in relevant contexts, while considering the objective of disclosure, the context of use, the likelihood and magnitude of the risk posed, the audience of the disclosure, as well as the frequency of the disclosures.	Human-AI Configuration		[Choose Level]	[Choose Level]			

MAP 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MP-5.1-004	Prioritize GAI structured public feedback processes based on risk assessment estimates.	Information Integrity; CBRN Information or Capabilities; Dangerous, Violent, or Hateful Content; Harmful Bias and Homogenization	- Vulnerability Reporting Program - User Feedback Interface - Misuse Reporting	[Choose Level]	[Choose Level]			
MP-5.1-005	Conduct adversarial role-playing exercises, GAI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes.	Information Security	- Red Teams - Adversarial Testing	[Choose Level]	[Choose Level]			
MP-5.1-006	Profile threats and negative impacts arising from GAI systems interacting with, manipulating, or generating content, and outlining known and potential vulnerabilities and the likelihood of their occurrence.	Information Security	- GAI-Specific Risk Assessment	[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Design, AI Development, AI Impact Assessment, Affected Individuals and Communities, End- Users, Operation and Monitoring								

MAP 5.2: Practices and personnel for supporting regular engagement with relevant AI Actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
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		[Choose Level]	[Choose Level]			
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	- Incident Reporting Support - Complaint and Redress Mechanisms - Independent Audits - Whistleblower Protection Policy - Vulnerability Reporting Program	[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Design, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End- Users, Human Factors, Operation and Monitoring								

Measure

MEASURE 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-1.1-001	Employ methods to trace the origin and modifications of digital content.	Information Integrity	Provenance Generation and Tracking: - Watermarking - Human Authentication - Distributed Ledger Tech/Blockchain Technology	[Choose Level]	[Choose Level]			
MS-1.1-002	Integrate tools designed to analyze content provenance and detect data anomalies, verify the authenticity of digital signatures, and identify patterns associated with misinformation or manipulation.	Information Integrity	Provenance Generation and Tracking: - Statistical Detection - Behavioral analysis - Turing Testing - Automated Synthetic Data Detection	[Choose Level]	[Choose Level]			

MEASURE 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MS-1.1-003	Disaggregate evaluation metrics by demographic factors to identify any discrepancies in how content provenance mechanisms work across diverse populations.	Information Integrity; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
MS-1.1-004	Develop a suite of metrics to evaluate structured public feedback exercises informed by representative AI Actors.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities		[Choose Level]	[Choose Level]			
MS-1.1-005	Evaluate novel methods and technologies for the measurement of GAI-related risks including in content provenance, offensive cyber, and CBRN, while maintaining the models' ability to produce valid, reliable, and factually accurate outputs.	Information Integrity; CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content	- Red Teams - Benchmarks - Capability Evaluations	[Choose Level]	[Choose Level]			

MEASURE 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MS-1.1-006	Implement continuous monitoring of GAI system impacts to identify whether GAI outputs are equitable across various sub-populations. Seek active and direct feedback from affected communities via structured feedback mechanisms or red- teaming to monitor and improve outputs.	Harmful Bias and Homogenization	- General Fairness Metrics	[Choose Level]	[Choose Level]			
MS-1.1-007	Evaluate the quality and integrity of data used in training and the provenance of AI-generated content, for example by employing techniques like chaos engineering and seeking stakeholder feedback.	Information Integrity	- Data Audits - Provenance Generation and Tracking	[Choose Level]	[Choose Level]			
MS-1.1-008	Define use cases, contexts of use, capabilities, and negative impacts where structured human feedback exercises, e.g., GAI red-teaming, would be	Harmful Bias and Homogeniz ation; CBRN		[Choose Level]	[Choose Level]			

MEASURE 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
	most beneficial for GAI risk measurement and management based on the context of use.	Information or Capabilities						
MS-1.1-009	Track and document risks or opportunities related to all GAI risks that cannot be measured quantitatively, including explanations as to why some risks cannot be measured (e.g., due to technological limitations, resource constraints, or trustworthy considerations). Include unmeasured risks in marginal risks.	Information Integrity	- AI Risk Repository - AI Incident Database - Risk Tiers	[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Development, Domain Experts, TEVV								

MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI Actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-1.3-001	Define relevant groups of interest (e.g., demographic groups, subject matter experts, experience with GAI technology) within the context of use as part of plans for gathering structured public feedback.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities		[Choose Level]	[Choose Level]			
MS-1.3-002	Engage in internal and external evaluations, GAI red-teaming, impact assessments, or other structured human feedback exercises in consultation with representative AI Actors with expertise and familiarity in the context of use, and/or who are representative of the populations associated with the context of use.	Human-AI Configuration; Harmful Bias and Homogenization; CBRN Information or Capabilities	- Independent Audits - Red Teams - Impact Assessments - Bug Bounty Program - Bias Bounty Program	[Choose Level]	[Choose Level]			
MS-1.3-003	Verify those conducting structured human feedback exercises are not directly involved in system development tasks for the same GAI model.	Human-AI Configuration; Data Privacy		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Development, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End-Users, Operation and Monitoring, TEVV								

MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.2-001	Assess and manage statistical biases related to GAI content provenance through techniques such as re-sampling, re-weighting, or adversarial training.	Information Integrity; Information Security; Harmful Bias and Homogenization	- Fairness and Bias Evaluations - General Fairness Metrics	[Choose Level]	[Choose Level]			
MS-2.2-002	Document how content provenance data is tracked and how that data interacts with privacy and security. Consider: Anonymizing data to protect the privacy of human subjects; Leveraging privacy output filters; Removing any personally identifiable information (PII) to prevent potential harm or misuse.	Data Privacy; Human AI Configuration; Information Integrity; Information Security; Dangerous, Violent, or Hateful Content	- Provenance Generation and Tracking	[Choose Level]	[Choose Level]			
MS-2.2-003	Provide human subjects with options to withdraw participation or revoke their	Data Privacy; Human-AI Configuration; Information Integrity		[Choose Level]	[Choose Level]			

MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	consent for present or future use of their data in GAI applications.							
MS-2.2-004	Use techniques such as anonymization, differential privacy or other privacy-enhancing technologies to minimize the risks associated with linking AI-generated content back to individual human subjects.	Data Privacy; Human-AI Configuration		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Development, Human Factors, TEVV								

MEASURE 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.3-001	Consider baseline model performance on suites of benchmarks when selecting a model for fine tuning or enhancement with retrieval-augmented generation.	Information Security; Confabulation	- Benchmarks	[Choose Level]	[Choose Level]			

MEASURE 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.3-002	Evaluate claims of model capabilities using empirically validated methods.	Confabulation; Information Security	- Capability Evaluations	[Choose Level]	[Choose Level]			
MS-2.3-003	Share results of pre-deployment testing with relevant GAI Actors, such as those with system release approval authority.	Human-AI Configuration		[Choose Level]	[Choose Level]			
MS-2.3-004	Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.	CBRN Information or Capabilities; Data Privacy; Confabulation; Information Integrity; Information Security; Dangerous, Violent, or Hateful Content; Harmful Bias and Homogenization	- Testing Environments	[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, TEVV								

MEASURE 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.								
Action ID	**Suggested Action**	**Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.5-001	Avoid extrapolating GAI system performance or capabilities from narrow, non- systematic, and anecdotal assessments.	Human-AI Configuration; Confabulation		[Choose Level]	[Choose Level]			
MS-2.5-002	Document the extent to which human domain knowledge is employed to improve GAI system performance, via, e.g., RLHF, fine-tuning, retrieval-augmented generation, content moderation, business rules.	Human-AI Configuration		[Choose Level]	[Choose Level]			
MS-2.5-003	Review and verify sources and citations in GAI system outputs during pre- deployment risk measurement and ongoing monitoring activities.	Confabulation		[Choose Level]	[Choose Level]			
MS-2.5-004	Track and document instances of anthropomorphization (e.g., human images, mentions of human feelings, cyborg imagery or motifs) in GAI system interfaces.	Human-AI Configuration		[Choose Level]	[Choose Level]			
MS-2.5-005	Verify GAI system training data and TEVV data provenance, and that fine-tuning or	Information Integrity	- Data Audits - Provenance Generation and Tracking	[Choose Level]	[Choose Level]			

MEASURE 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.								
Action ID	**Suggested Action**	**Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	retrieval-augmented generation data is grounded.							
MS-2.5-006	Regularly review security and safety guardrails, especially if the GAI system is being operated in novel circumstances. This includes reviewing reasons why the GAI system was initially assessed as being safe to deploy.	Information Security; Dangerous, Violent, or Hateful Content	- Red Teams - Adversarial Testing - Red Team Approaches with Security Implications	[Choose Level]	[Choose Level]			
AI Actor Tasks: Domain Experts, TEVV								

MEASURE 2.6: The AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.6-001	Assess adverse impacts, including health and wellbeing impacts for value chain or other AI Actors that are exposed to sexually explicit, offensive, or violent information during GAI training and maintenance.	Human-AI Configuration; Obscene, Degrading, and/or Abusive Content; Value Chain and Component Integration; Dangerous,		[Choose Level]	[Choose Level]			

MEASURE 2.6: The AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
		Violent, or Hateful Content						
MS-2.6-002	Assess existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data.	Data Privacy; Intellectual Property; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; CBRN Information or Capabilities	- Data Audits - Data Sheets	[Choose Level]	[Choose Level]			
MS-2.6-003	Re-evaluate safety features of fine-tuned models when the negative risk exceeds organizational risk tolerance.	Dangerous, Violent, or Hateful Content	- Organizational Risk Thresholds	[Choose Level]	[Choose Level]			
MS-2.6-004	Review GAI system outputs for validity and safety: Review generated code to assess risks that may arise from unreliable downstream decision-making.	Value Chain and Component Integration; Dangerous, Violent, or Hateful Content	- Red Teams - Capability Evaluations	[Choose Level]	[Choose Level]			

MEASURE 2.6: The AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.6-005	Verify that GAI system architecture can monitor outputs and performance, and handle, recover from, and repair errors when security anomalies, threats and impacts are detected.	Confabulation; Information Integrity; Information Security	- Incident Response Plan	[Choose Level]	[Choose Level]			
MS-2.6-006	Verify that systems properly handle queries that may give rise to inappropriate, malicious, or illegal usage, including facilitating manipulation, extortion, targeted impersonation, cyber-attacks, and weapons creation.	CBRN Information or Capabilities; Information Security	- Red Teams - Benchmarks - Toxicity Evaluation	[Choose Level]	[Choose Level]			
MS-2.6-007	Regularly evaluate GAI system vulnerabilities to possible circumvention of safety measures.	CBRN Information or Capabilities; Information Security		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV								

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.7-001	Apply established security measures to: Assess likelihood and magnitude of vulnerabilities and threats such as backdoors, compromised dependencies, data breaches, eavesdropping, man-in-the-middle attacks, reverse engineering, autonomous agents, model theft or exposure of model weights, AI inference, bypass, extraction, and other baseline security concerns.	Data Privacy; Information Integrity; Information Security; Value Chain and Component Integration	- Red Teams - Adversarial Testing - Impact Assessments	[Choose Level]	[Choose Level]			
MS-2.7-002	Benchmark GAI system security and resilience related to content provenance against industry standards and best practices. Compare GAI system security features and content provenance methods against industry state-of-the-art.	Information Integrity; Information Security	- Baseline Security Measures - Provenance Generation and Tracking	[Choose Level]	[Choose Level]			

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.7-003	Conduct user surveys to gather user satisfaction with the AI-generated content and user perceptions of content authenticity. Analyze user feedback to identify concerns and/or current literacy levels related to content provenance and understanding of labels on content.	Human-AI Configuration; Information Integrity	- User Feedback Interface	[Choose Level]	[Choose Level]			
MS-2.7-004	Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, inference, bypass, extraction, penetrations, or provenance verification.	Information Integrity; Information Security		[Choose Level]	[Choose Level]			
MS-2.7-005	Measure reliability of content authentication methods, such as watermarking, cryptographic signatures, digital fingerprints, as well as access controls, conformity assessment,	Information Integrity		[Choose Level]	[Choose Level]			

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	and model integrity verification, which can help support the effective implementation of content provenance techniques. Evaluate the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification.							
MS-2.7-006	Measure the rate at which recommendations from security checks and incidents are implemented. Assess how quickly the AI system can adapt and improve based on lessons learned from security incidents and feedback.	Information Integrity; Information Security		[Choose Level]	[Choose Level]			
MS-2.7-007	Perform AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial	Information Security; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content	- Red-Team Approaches with Security Implications	[Choose Level]	[Choose Level]			

MEASURE 2.7: AI system security and resilience – as identified in the MAP function – are evaluated and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	examples/prompts, data poisoning, membership inference, model extraction, sponge examples).							
MS-2.7-008	Verify fine-tuning does not compromise safety and security controls.	Information Integrity; Information Security; Dangerous, Violent, or Hateful Content	- Red-Team Approaches with Security Implications	[Choose Level]	[Choose Level]			
MS-2.7-009	Regularly assess and verify that security measures remain effective and have not been compromised.	Information Security		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV								

MEASURE 2.8: Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.8-001	Compile statistics on actual policy violations, take-down requests, and intellectual property infringement for organizational GAI systems:	Intellectual Property; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			

MEASURE 2.8: Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	Analyze transparency reports across demographic groups, languages groups.							
MS-2.8-002	Document the instructions given to data annotators or AI red-teamers.	Human-AI Configuration		[Choose Level]	[Choose Level]			
MS-2.8-003	Use digital content transparency solutions to enable the documentation of each instance where content is generated, modified, or shared to provide a tamper-proof history of the content, promote transparency, and enable traceability. Robust version control systems can also be applied to track changes across the AI lifecycle over time.	Information Integrity	- Provenance Generation and Tracking	[Choose Level]	[Choose Level]			
MS-2.8-004	Verify adequacy of GAI system user instructions through user testing.	Human-AI Configuration		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV								

MEASURE 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the MAP function – to inform responsible use and governance.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MS-2.9-001	Apply and document ML explanation results such as: Analysis of embeddings, Counterfactual prompts, Gradient-based attributions, Model compression/surrogate models, Occlusion/term reduction.	Confabulation		[Choose Level]	[Choose Level]			
MS-2.9-002	Document GAI model details including: Proposed use and organizational value; Assumptions and limitations, Data collection methodologies; Data provenance; Data quality; Model architecture (e.g., convolutional neural network, transformers, etc.); Optimization objectives; Training algorithms; RLHF approaches; Fine-tuning or retrieval-augmented generation approaches; Evaluation data; Ethical considerations; Legal and regulatory requirements.	Information Integrity; Harmful Bias and Homogenization	- System Cards - Model Cards - Data Statements - Data Sheets - Data Audits	[Choose Level]	[Choose Level]			

AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, End-Users, Operation and Monitoring, TEVV

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

MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.11-001	Apply use-case appropriate benchmarks (e.g., Bias Benchmark Questions, Real Hateful or Harmful Prompts, Winogender Schemas[15]) to quantify systemic bias, stereotyping, denigration, and hateful content in GAI system outputs; Document assumptions and limitations of benchmarks, including any actual or possible training/test data cross contamination, relative to in-context deployment environment.	Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			

MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.11-002	Conduct fairness assessments to measure systemic bias. Measure GAI system performance across demographic groups and subgroups, addressing both quality of service and any allocation of services and resources. Quantify harms using: field testing with sub-group populations to determine likelihood of exposure to generated content exhibiting harmful bias, AI red-teaming with counterfactual and low-context (e.g., “leader,” “bad guys”) prompts. For ML pipelines or business processes with categorical or numeric outcomes that rely on GAI, apply general fairness metrics (e.g., demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), to the pipeline or business outcome where appropriate; Custom, context-specific metrics developed in collaboration with domain experts and affected communities; Measurements of the prevalence of denigration in generated content in deployment (e.g., sub- sampling a	Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content	- Fairness and Bias Evaluations	[Choose Level]	[Choose Level]			

MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	fraction of traffic and manually annotating denigrating content).							
MS-2.11-003	Identify the classes of individuals, groups, or environmental ecosystems which might be impacted by GAI systems through direct engagement with potentially impacted communities.	Environmental; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
MS-2.11-004	Review, document, and measure sources of bias in GAI training and TEVV data: Differences in distributions of outcomes across and within groups, including intersecting groups; Completeness, representativeness, and balance of data sources; demographic group and subgroup coverage in GAI system training data; Forms of latent systemic bias in images, text, audio, embeddings, or other complex or unstructured data; Input data features that may serve as proxies for demographic group membership (i.e., image metadata, language dialect) or otherwise give rise to emergent bias within GAI systems; The extent to which the digital divide may negatively impact representativeness in GAI system	Harmful Bias and Homogenization	- Fairness and Bias Evaluations - General Fairness Metrics	[Choose Level]	[Choose Level]			

MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	training and TEVV data; Filtering of hate speech or content in GAI system training data; Prevalence of GAI-generated data in GAI system training data.							
MS-2.11-005	Assess the proportion of synthetic to non-synthetic training data and verify training data is not overly homogenous or GAI-produced to mitigate concerns of model collapse.	Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Affected Individuals and Communities, Domain Experts, End-Users, Operation and Monitoring, TEVV								

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

MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-2.12-001	Assess safety to physical environments when deploying GAI systems.	Dangerous, Violent, or Hateful Content		[Choose Level]	[Choose Level]			
MS-2.12-002	Document anticipated environmental impacts of model	Environmental		[Choose Level]	[Choose Level]			

MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	development, maintenance, and deployment in product design decisions.							
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	- Machine Learning CO2 Impact	[Choose Level]	[Choose Level]			
MS-2.12-004	Verify effectiveness of carbon capture or offset programs for GAI training and applications, and address green-washing concerns.	Environmental		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV								

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

MEASURE 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-3.2-001	Establish processes for identifying emergent GAI system risks including consulting with external AI Actors.	Human-AI Configuration; Confabulation	- Red Teams - Adversarial Testing	[Choose Level]	[Choose Level]			

MEASURE 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
			- Bug Bounty Program - API Monitoring					
AI Actor Tasks: AI Impact Assessment, Domain Experts, Operation and Monitoring, TEVV								

MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MS-3.3-001	Conduct impact assessments on how AI-generated content might affect different social, economic, and cultural groups.	Harmful Bias and Homogenization	- Impact Assessments - Human Rights Impact Assessments	[Choose Le…	[Choose Le…			
MS-3.3-002	Conduct studies to understand how end users perceive and interact with GAI content and accompanying content provenance within context of use. Assess whether the content aligns with their expectations and	Human-AI Configuration; Information Integrity		[Choose Le…	[Choose Le…			

MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	how they may act upon the information presented.							
MS-3.3-003	Evaluate potential biases and stereotypes that could emerge from the AI- generated content using appropriate methodologies including computational testing methods as well as evaluating structured feedback input.	Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
MS-3.3-004	Provide input for training materials about the capabilities and limitations of GAI systems related to digital content transparency for AI Actors, other professionals, and the public about the societal impacts of AI and the role of diverse and inclusive content generation.	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization	- Model Cards - System Cards	[Choose Le…	[Choose Le…			
MS-3.3-005	Record and integrate structured feedback about content provenance from operators, users, and potentially impacted communities through the use of methods such as user research studies, focus groups, or community forums. Actively seek feedback on generated content quality and potential biases. Assess the general awareness among end users and impacted communities	Human-AI Configuration; Information Integrity; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			

MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	about the availability of these feedback channels.							
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, End-Users, Operation and Monitoring, TEVV								

MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI Actors to validate whether the system is performing consistently as intended. Results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-4.2-001	Conduct adversarial testing at a regular cadence to map and measure GAI risks, including tests to address attempts to deceive or manipulate the application of provenance techniques or other misuses. Identify vulnerabilities and understand potential misuse scenarios and unintended outputs.	Information Integrity; Information Security	- Adversarial Testing	[Choose Le…	[Choose Le…			
MS-4.2-002	Evaluate GAI system performance in real-world scenarios to observe its behavior in practical environments and reveal issues that might not surface in controlled and optimized testing environments.	Human-AI Configuration; Confabulation; Information Security		[Choose Le…	[Choose Le…			

MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI Actors to validate whether the system is performing consistently as intended. Results are documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MS-4.2-003	Implement interpretability and explainability methods to evaluate GAI system decisions and verify alignment with intended purpose.	Information Integrity; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
MS-4.2-004	Monitor and document instances where human operators or other systems override the GAI's decisions. Evaluate these cases to understand if the overrides are linked to issues related to content provenance.	Information Integrity		[Choose Le…	[Choose Le…			
MS-4.2-005	Verify and document the incorporation of results of structured public feedback exercises into design, implementation, deployment approval (“go”/“no-go” decisions), monitoring, and decommission decisions.	Human-AI Configuration; Information Security		[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Domain Experts, End-Users, Operation and Monitoring, TEVV								

Manage

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

MANAGE 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-2.2-001	Compare GAI system outputs against pre-defined organization risk tolerance, guidelines, and principles, and review and test AI-generated content against these guidelines.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content		[Choose Level]	[Choose Level]			
MG-2.2-002	Document training data sources to trace the origin and provenance of AI-generated content.	Information Integrity	- Data Sheets	[Choose Level]	[Choose Level]			
MG-2.2-003	Evaluate feedback loops between GAI system content provenance and human reviewers, and update where needed. Implement real-time monitoring systems to affirm that content provenance protocols remain effective.	Information Integrity		[Choose Level]	[Choose Level]			
MG-2.2-004	Evaluate GAI content and data for representational biases and employ techniques such as re-sampling, re-ranking, or adversarial training to mitigate biases in the generated content.	Information Security; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			

MANAGE 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-2.2-005	Engage in due diligence to analyze GAI output for harmful content, potential misinformation, and CBRN-related or NCII content.	CBRN Information or Capabilities; Obscene, Degrading, and/or Abusive Content; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content	- Red Teams - Benchmarks - Adversarial Testing - Capability Evaluations	[Choose Level]	[Choose Level]			
MG-2.2-006	Use feedback from internal and external AI Actors, users, individuals, and communities, to assess impact of AI-generated content.	Human-AI Configuration	- Vulnerability Reporting Program - Misuse Reporting - User Feedback Interface	[Choose Level]	[Choose Level]			
MG-2.2-007	Use real-time auditing tools where they can be demonstrated to aid in the tracking and validation of the lineage and authenticity of AI-generated data.	Information Integrity	- Provenance Generation and Tracking	[Choose Level]	[Choose Level]			

MANAGE 2.2: Mechanisms are in place and applied to sustain the value of deployed AI systems.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-2.2-008	Use structured feedback mechanisms to solicit and capture user input about AI- generated content to detect subtle shifts in quality or alignment with community and societal values.	Human-AI Configuration; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
MG-2.2-009	Consider opportunities to responsibly use synthetic data and other privacy enhancing techniques in GAI development, where appropriate and applicable, match the statistical properties of real-world data without disclosing personally identifiable information or contributing to homogenization.	Data Privacy; Intellectual Property; Information Integrity; Confabulation; Harmful Bias and Homogenization		[Choose Level]	[Choose Level]			
AI Actor Tasks: AI Deployment, AI Impact Assessment, Governance and Oversight, Operation and Monitoring								

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

MANAGE 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-2.4-001	Establish and maintain communication plans to inform AI stakeholders as part of the deactivation or disengagement process of a specific GAI system (including for open-source models) or	Human-AI Configuration	- Decommissioning Policy	[Choose Le…	[Choose Le…			

MANAGE 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	context of use, including reasons, workarounds, user access removal, alternative processes, contact information, etc.							
MG-2.4-002	Establish and maintain procedures for escalating GAI system incidents to the organizational risk management authority when specific criteria for deactivation or disengagement is met for a particular context of use or for the GAI system as a whole.	Information Security	- Incident Response Plan	[Choose Le…	[Choose Le…			
MG-2.4-003	Establish and maintain procedures for the remediation of issues which trigger incident response processes for the use of a GAI system, and provide stakeholders timelines associated with the remediation plan.	Information Security	- Incident Response Plan	[Choose Le…	[Choose Le…			
MG-2.4-004	Establish and regularly review specific criteria that warrants the deactivation of GAI systems in accordance with set risk tolerances and appetites.	Information Security	- Organizational Risk Thresholds - Decommissioning Policy	[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Governance and Oversight, Operation and Monitoring								

MANAGE 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-3.1-001	Apply organizational risk tolerances and controls (e.g., acquisition and procurement processes; assessing personnel credentials and qualifications, performing background checks; filtering GAI input and outputs, grounding, fine tuning, retrieval-augmented generation) to third-party GAI resources: Apply organizational risk tolerance to the utilization of third-party datasets and other GAI resources; Apply organizational risk tolerances to fine-tuned third-party models; Apply organizational risk tolerance to existing third-party models adapted to a new domain; Reassess risk measurements after fine-tuning third-party GAI models.	Value Chain and Component Integration; Intellectual Property		[Choose Le…	[Choose Le…			
MG-3.1-002	Test GAI system value chain risks (e.g., data poisoning, malware, other software and hardware vulnerabilities; labor practices; data privacy and localization compliance; geopolitical alignment).	Data Privacy; Information Security; Value Chain and Component Integration; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
MG-3.1-003	Re-assess model risks after fine-tuning or retrieval-augmented generation implementation and for any third-party	Value Chain and Component Integration		[Choose Le…	[Choose Le…			

MANAGE 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	GAI models deployed for applications and/or use cases that were not evaluated in initial testing.							
MG-3.1-004	Take reasonable measures to review training data for CBRN information, and intellectual property, and where appropriate, remove it. Implement reasonable measures to prevent, flag, or take other action in response to outputs that reproduce particular training data (e.g., plagiarized, trademarked, patented, licensed content or trade secret material).	Intellectual Property; CBRN Information or Capabilities	- Data Audits	[Choose Le…	[Choose Le…			
MG-3.1-005	Review various transparency artifacts (e.g., system cards and model cards) for third-party models.	Information Integrity; Information Security; Value Chain and Component Integration		[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Operation and Monitoring, Third-party entities								

MANAGE 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-3.2-001	Apply explainable AI (XAI) techniques (e.g., analysis of embeddings, model compression/distillation, gradient-based attributions, occlusion/term reduction, counterfactual prompts, word clouds) as part of ongoing continuous improvement processes to mitigate risks related to unexplainable GAI systems.	Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
MG-3.2-002	Document how pre-trained models have been adapted (e.g., fine-tuned, or retrieval-augmented generation) for the specific generative task, including any data augmentations, parameter adjustments, or other modifications. Access to un-tuned (baseline) models supports debugging the relative influence of the pre- trained weights compared to the fine-tuned model weights or other system updates.	Information Integrity; Data Privacy	- Model Cards - System Cards	[Choose Le…	[Choose Le…			
MG-3.2-003	Document sources and types of training data and their origins, potential biases present in the data related to the GAI application and its content provenance, architecture, training process of the pre-trained model including information on hyperparameters, training duration, and any fine-tuning or	Information Integrity; Harmful Bias and Homogenization; Intellectual Property	- Data Sheets	[Choose Le…	[Choose Le…			

MANAGE 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
	retrieval-augmented generation processes applied.							
MG-3.2-004	Evaluate user reported problematic content and integrate feedback into system updates.	Human-AI Configuration, Dangerous, Violent, or Hateful Content		[Choose Le…	[Choose Le…			
MG-3.2-005	Implement content filters to prevent the generation of inappropriate, harmful, false, illegal, or violent content related to the GAI application, including for CSAM and NCII. These filters can be rule-based or leverage additional machine learning models to flag problematic inputs and outputs.	Information Integrity; Harmful Bias and Homogenization; Dangerous, Violent, or Hateful Content; Obscene, Degrading, and/or Abusive Content		[Choose Le…	[Choose Le…			
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	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			

MANAGE 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
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		[Choose Le…	[Choose Le…			
MG-3.2-008	Use human moderation systems where appropriate to review generated content in accordance with human-AI configuration policies established in the Govern function, aligned with socio-cultural norms in the context of use, and for settings where AI models are demonstrated to perform poorly.	Human-AI Configuration		[Choose Le…	[Choose Le…			
MG-3.2-009	Use organizational risk tolerance to evaluate acceptable risks and performance metrics and decommission or retrain pre-trained models that perform outside of defined limits.	CBRN Information or Capabilities; Confabulation	- Organizational Risk Thresholds - Decommissioning Policy	[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Operation and Monitoring, Third-party entities								

MANAGE 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI Actors, appeal and override, decommissioning, incident response, recovery, and change management.

Action ID	Suggested Action	GAI Risks	Suggested Practices and Documentation	Priority Level	Implementation Level	Summary of Approach	Supporting Documentation	Responsible Party
MG-4.1-001	Collaborate with external researchers, industry experts, and community representatives to maintain awareness of emerging best practices and technologies in measuring and managing identified risks.	Information Integrity; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			
MG-4.1-002	Establish, maintain, and evaluate effectiveness of organizational processes and procedures for post-deployment monitoring of GAI systems, particularly for potential confabulation, CBRN, or cyber risks.	CBRN Information or Capabilities; Confabulation; Information Security		[Choose Le…	[Choose Le…			
MG-4.1-003	Evaluate the use of sentiment analysis to gauge user sentiment regarding GAI content performance and impact, and work in collaboration with AI Actors experienced in user research and experience.	Human-AI Configuration		[Choose Le…	[Choose Le…			
MG-4.1-004	Implement active learning techniques to identify instances where the model fails or produces unexpected outputs.	Confabulation		[Choose Le…	[Choose Le…			
MG-4.1-005	Share transparency reports with internal and external stakeholders that detail steps taken to update the GAI system to enhance transparency and accountability.	Human-AI Configuration; Harmful Bias and Homogenization		[Choose Le…	[Choose Le…			

MANAGE 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI Actors, appeal and override, decommissioning, incident response, recovery, and change management.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-4.1-006	Track dataset modifications for provenance by monitoring data deletions, rectification requests, and other changes that may impact the verifiability of content origins.	Information Integrity	- Provenance Generation and Tracking	[Choose Le…	[Choose Le…			
MG-4.1-007	Verify that AI Actors responsible for monitoring reported issues can effectively evaluate GAI system performance including the application of content provenance data tracking techniques, and promptly escalate issues for response.	Human-AI Configuration; Information Integrity		[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, Domain Experts, End-Users, Human Factors, Operation and Monitoring								

MANAGE 4.2: Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI Actors.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-4.2-001	Conduct regular monitoring of GAI systems and publish reports detailing the performance, feedback received, and improvements made.	Harmful Bias and Homogenization	- Model Cards - System Cards - Model Limits Documentation	[Choose Le…	[Choose Le…			

MANAGE 4.2: Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI Actors.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-4.2-002	Practice and follow incident response plans for addressing the generation of inappropriate or harmful content and adapt processes based on findings to prevent future occurrences. Conduct post-mortem analyses of incidents with relevant AI Actors, to understand the root causes and implement preventive measures.	Human-AI Configuration; Dangerous, Violent, or Hateful Content	- Incident Response Plan - Incident Reporting Support - Incident Disclosure Plan - Incident History Database	[Choose Le…	[Choose Le…			
MG-4.2-003	Use visualizations or other methods to represent GAI model behavior to ease non-technical stakeholders understanding of GAI system functionality.	Human-AI Configuration		[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, AI Design, AI Development, Affected Individuals and Communities, End-Users, Operation and Monitoring, TEVV								

MANAGE 4.3: Incidents and errors are communicated to relevant AI Actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.								
Action ID	**Suggested Action**	**GAI Risks**	**Suggested Practices and Documentation**	**Priority Level**	**Implementation Level**	**Summary of Approach**	**Supporting Documentation**	**Responsible Party**
MG-4.3-001	Conduct after-action assessments for GAI system incidents to verify incident response and recovery processes are followed and effective, including to follow procedures for communicating incidents to relevant AI Actors and where applicable, relevant legal and regulatory bodies.	Information Security		[Choose Le…	[Choose Le…			
MG-4.3-002	Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts.	Confabulation; Information Integrity		[Choose Le…	[Choose Le…			
MG-4.3-003	Report GAI incidents in compliance with legal and regulatory requirements (e.g., HIPAA breach reporting, e.g., OCR (2023) or NHTSA (2022) autonomous vehicle crash reporting requirements.	Information Security; Data Privacy	- Incident Disclosure Plan	[Choose Le…	[Choose Le…			
AI Actor Tasks: AI Deployment, Affected Individuals and Communities, Domain Experts, End-Users, Human Factors, Operation and Monitoring								

Specific Examples of Listed Practices and Documentation:

Practice/Documentation	Example(s)
Acceptable Use Policy	Meta AI (2024b) OpenAI (2024b) Anthropic (2024a)
Adversarial Testing	Simulating attacks in the model or system to identify vulnerabilities.
AI Incident Database	AIID (n.d.)
AI RACI Chart (Responsible, Accountable, Consulted, Informed)	Hurley (2018 pp. 71.72)
AI Risk Repository	MIT (2024) Slattery et al. (2024)
AI Safety Practices	OpenAI (2024d)
API Monitoring	Observation and analysis of the API input, output, and performance.
Automated Output Monitoring	Automated observation and analysis of GAI system output.
Baseline Security Measures	NIST Cybersecurity Framework (NIST 2024) NIST SP 800-53 (NIST 2023) including SC-28 NIST SP 800-53B (NIST 2020a) NIST SP 800-171 (NIST 2020b) NIST SP 800-172 (NIST 2021) ISO/IEC (2022) Anthropic (2023a) Anthropic (2023c) Nevo et al. (2024) NIST SP 800-218A (Booth et al. 2024)

Practice/Documentation	Example(s)
	ACSC (2024)
Benchmarking	DecodingTrust (Wang, Chen et al. 2023) BIG-bench “pro-social behavior” category of benchmark tasks (BIG-bench n.d.b, BIG-bench collaboration 2021, Srivastava et al. 2022) Model-Written Evaluations “advanced-ai-risk," “sycophancy,” and “winogender” datasets (Perez, Ringer et al. 2022a, 2022b) MACHIAVELLI (Pan et al. 2023) Accountability Benchmark (Gursoy and Kakadiaris 2022) LLM Lie Detection (Pacchiardi et al. 2023) Strategic deception (Scheurer et al. 2024)
Bug Bounty Program	Kenway et al. (2022)
Bias Bounty Program	Globus-Harris et al. (2022)
Capability Evaluations	Evaluating AI system or model performance and capability by carrying out tests. This may include Benchmarks and Red Teams.
Complaint and Redress Mechanisms	OpenAI (n.d.)
Cyber Vulnerability Capability Evaluation	Chauvin (2024)
Data Retention Policy	A policy that outlines how different types of data will be stored and disposed of, typically based on legal and regulatory compliance.
Data Audits	Birhane et al. (2021), Dodge et al. (2021)
Data Curation Policy	A policy that outlines data management procedures to ensure accuracy and relevance.
Data Sheets	Gebru et al. (2021)
Data Statements	Bender and Friedman (2018)

Practice/Documentation	Example(s)
Decommissioning Policy	A policy that outlines the circumstances and necessary procedures for systematic and deliberate decommissioning of AI systems or models.
Document Retention Policy	A policy that outlines how different types of documents will be stored and disposed of, typically based on legal and regulatory compliance.
Fingerprinting	ITI (2024)
Fairness and Bias Evaluations	Aequitas (Saleiro 2019) AIFairness 360 (Bellamy 2018) Fairlearn (Fairlearn Contributors 2023)
GAI-Specific Risk Assessment	Assessment of risks specific to GAI (e.g. misinformation, hallucination).
General Fairness Metrics - Demographic Parity - Equalized Odds - Equal Opportunity, - Statistical Hypothesis Tests	Garg et al (2020)
Human Rights Impact Assessments	DOS (2024)
Impact Assessments	UNESCO (2023)
Incident Disclosure Plan	Turri and Dzombak (2023)
Incident History Database	A comprehensive catalog of organizational GAI-related incidents.
Incident Reporting Support	Policies, interfaces, and procedures that support the ability of various stakeholders to report incidents.
Incident Response Plan	Comprehensive instructions that outline how the organization will respond to and recover from various types of incidents.

Practice/Documentation	Example(s)
Independent Audits	Audits performed by independent parties outside of the organization.
Machine Learning CO2 Impact	Schmidt et al. (2019) Lacoste et al. (2019) OECD (2022)
Misuse Reporting	Procedures and interfaces that allow various stakeholders to report AI system or model misuse.
Model Cards	Liang et al. (2024) Anthropic (2024a) Anthropic (2024c)
Model Limits Documentation	Comprehensive documentation of model or system performance and capability limitations.
Organizational Risk Thresholds	OpenAI (2023f pp. 6-11)
Privacy Policy	Google (2024b)
Prohibited Use Policy	Google (2023a)
Provenance Generation and Tracking - Watermarking - Human Authentication - Distributed Ledger Tech/Blockchain Technology - Statistical Detection - Behavioral Analysis - Turing Testing - Automated Synthetic Data Detection	ITI (2024)
Red Teams	Anthropic (2023b) OpenAI (2023b)

Practice/Documentation	Example(s)
Red Team Approaches with Security Implications	Ganguli, Lovitt et al. (2022) Casper et al. (2023a, 2023b, 2023c) Zou et al. (2023a, 2023b) OpenAI (2023a, pp. 15–16) ARC Evals[1] (2023a, 2023b) Kinniment et al. (2023) Anthropic (2023b, 2023c) Shevlane et al. (2023)
Responsible Scaling Policy	Anthropic (2023e, 2024d)
Responsible Disclosure Policy	Anthropic (2023d)
Risk Tiers	Defining risk tiers based on the level of risk they pose to people, the organization, and the ecosystem.
Service Level Agreements (SLAs)	A contract between the provider and user that outlines the provided service, the expected standards of the service, and the performance measurement criteria.
System Cards	OpenAI (2024a, 2023a)
System Inventory	OCC (2021 pp. 26-27)
Terms of Use Policy	OpenAI (2024c)
Terms of Service	Anthropic (2024b) Google (2023b) Google (2024a)
Testing Environments	Glasbrenner et al. (2024a, 2024b)
Threat Modeling	A structured process for identifying, assessing, and reducing risks.

[1] ARC Evals has since changed the organization's name to METR.

Practice/Documentation	Example(s)
Three Lines of Defense or 3LoD	Schuett (2022)
Toxicity Evaluations	Casper et al. (2023a, 2023b, 2023c) ToxiGen (Hartvigsen et al. 2022) TruthfulQA (Lin et al. 2021a, 2021b) MACHIAVELLI (Pan et al. 2023) Do-Not-Answer (Wang, Li et al., 2023)
Usage Guidelines	Meta AI (2024a)
Use Case Cards	Hupont et al. (2024a, 2024b)
User Feedback Interface	OpenAI (n.d.)
Vulnerability Disclosure Policy	OpenAI (2023c)
Vulnerability Reporting Program	Google (n.d.)
Watermarking	ITI (2024)
Whistleblower Protection Policy	A policy to protect whistleblowers from retaliation for reporting issues relating AI safety and security.

References

ACSC (2024) Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems. The Australian Signals Directorate's Australian Cyber Security Centre, https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/artificial-intelligence/deploying-ai-systems-securely

AIID (n.d.) AI Incident Database. https://incidentdatabase.ai/

Anthropic (2023a) Frontier Model Security. Anthropic, https://www.anthropic.com/index/frontier-model-security

Anthropic (2023b) Frontier Threats Red Teaming for AI Safety. Anthropic, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Anthropic (2023c) Collective Constitutional AI: Aligning a Language Model with Public Input. Anthropic, https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Anthropic (2023d) Responsible Disclosure Policy. Anthropic, https://www.anthropic.com/responsible-disclosure-policy

Anthropic (2023e) Anthropic's Responsible Scaling Policy, Version 1.0. Anthropic, https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

Anthropic (2024a) Usage Policy. Anthropic, https://www.anthropic.com/legal/aup

Anthropic (2024b) Commercial Terms of Service. Anthropic, https://www.anthropic.com/legal/commercial-terms

Anthropic (2024c) The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic, https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf

Anthropic (2024d) Responsible Scaling Program Updates. Anthropic, https://www.anthropic.com/rsp-updates

ARC Evals (2023a) Update on ARC's recent eval efforts: More information about ARC's evaluations of GPT-4 and Claude. Alignment Research Center, https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

ARC Evals (2023b) The TaskRabbit example. Alignment Research Center, https://evals.alignment.org/taskrabbit.pdf

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, Yunfeng Zhang (2018) AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv, https://arxiv.org/abs/1810.01943

Emily M. Bender and Batya Friedman (2018) Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00041/43452/Data-Statements-for-Natural-Language-Processing

BIG-bench collaboration (2021) Beyond the Imitation Game Benchmark (BIG-bench). https://github.com/google/BIG-bench/

BIG-bench (n.d.b) Summary table. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/keywords_to_tasks.md

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv, https://arxiv.org/abs/2110.01963

Harold Booth, Murugiah Souppaya, Apostol Vassilev, Michael Ogata, Martin Stanley, Karen Scarfone (2024) Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. NIST SP 800-218A. National Institute of Standards and Technology, https://csrc.nist.gov/pubs/sp/800/218/a/final

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, Dylan Hadfield-Menell (2023a) Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv*, https://arxiv.org/abs/2306.09442

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, Dylan Hadfield-Menell (2023b) Explore, Establish, Exploit: Red Teaming Language Models from Scratch. https://github.com/thestephencasper/explore_establish_exploit_llms

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, Dylan Hadfield-Menell (2023c) CommonClaim Dataset. https://github.com/Algorithmic-Alignment-Lab/CommonClaim

Timothee Chauvin (2024) eyeballvul: A Future-Proof Benchmark for Vulnerability Detection in the Wild. arXiv, https://arxiv.org/abs/2407.08708

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021) Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, https://aclanthology.org/2021.emnlp-main.98/

DOS (2024) Risk Management Profile for Artificial Intelligence and Human Rights. United States Department of State, https://www.state.gov/risk-management-profile-for-ai-and-human-rights/

Fairlearn Contributors (2023) Fairlearn, https://github.com/fairlearn/fairlearn

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark (2022) Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv*, https://arxiv.org/abs/2209.07858

Pratyush Garg, John Villasenor, and Virginia Foggo (2020) Fairness Metrics: A Comparative Analysis. *arXiv*, https://arxiv.org/abs/2001.07864

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford (2021) Datasheets for datasets. Association for Computing Machinery, https://doi.org/10.1145/3458723

James Glasbrenner, Harold Booth, Keith Manville, Julian Sexton, Michael Andy Chisholm, Henry Choy, Andrew Hand, Bronwyn Hodges, Paul Scemama, Dmitry Cousin, Eric Trapnell, Mark Trapnell, Howard Huang, Paul Rowe, and Alex Byrne (2024a) Dioptra Test Platform. National Institute of Standards and Technology, https://doi.org/10.18434/mds2-3398

James Glasbrenner, Harold Booth, Keith Manville, Julian Sexton, Michael Andy Chisholm, Henry Choy, Andrew Hand, Bronwyn Hodges, Paul Scemama, Dmitry Cousin, Eric Trapnell, Mark Trapnell, Howard Huang, Paul Rowe, and Alex Byrne (2024b) Dioptra Test Platform. National Institute of Standards and Technology, https://github.com/usnistgov/dioptra

Ira Globus-Harris, Michael Kearns, Aaron Roth (2022) An Algorithmic Framework for Bias Bounties. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, https://doi.org/10.1145/3531146.3533172 or https://arxiv.org/abs/2201.10408

Google (2023a) Generative AI Prohibited Use Policy. Google, https://policies.google.com/terms/generative-ai/use-policy

Google (2023b) Generative AI Terms of Service. Google, https://policies.google.com/terms/generative-ai

Google (2024a) Google Terms of Service. Google, https://policies.google.com/terms

Google (2024b) Gemini Apps Privacy Notice. Google, https://support.google.com/gemini/answer/13594961?visit_id=638501643118708256-3012533406&p=privacy_notice&rd=1#privacy_notice

Google (n.d.) Google and Alphabet Vulnerability Reward Program (VRP). Google, https://bughunters.google.com/about/rules/google-friends/6625378258649088/google-and-alphabet-vulnerability-reward-program-vrp-rules

Furkan Gursoy and Ioannis A. Kakadiaris (2022) System Cards for AI-Based Decision-Making for Public Policy. arXiv, https://arxiv.org/abs/2203.04754

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, Ece Kamar (2022) ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. arXiv, https://arxiv.org/abs/2203.09509

Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez (2024a) Use case cards: a use case reporting framework inspired by the European AI Act. *Ethics Inf Technol 26*, https://doi.org/10.1007/s10676-024-09757-7

Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez (2024b) Use case cards: a use case reporting framework inspired by the European AI Act. *GitLab*, https://gitlab.com/humaint-ec_public/use-case-cards

JS Hurley (2018) Enabling Successful Artificial Intelligence Implementation in the Department of Defense. *JSTOR*, https://www.jstor.org/stable/26633155

ISO/IEC (2022) ISO/IEC International Standard 27001:2022, Information security management systems. https://www.iso.org/standard/27001

ITI (2024) Authenticating AI-Generated Content: Exploring Risks, Techniques & Policy Recommendations. The Information Technology Industry Council (ITI), https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini (2022) Bug Bounties for Algorithmic Harms? Algorithmic Justice League, https://www.ajl.org/bugs

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, Paul Christiano (2023) Evaluating Language-Model Agents on Realistic Autonomous Tasks. Alignment Research Center, https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres (2019) Quantifying the Carbon Emissions of Machine Learning. *arXiv,* https://arxiv.org/abs/1910.09700

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou (2024) What's documented in AI? Systematic Analysis of 32K AI Model Cards. arXiv, https://arxiv.org/abs/2402.05160

Stephanie Lin, Jacob Hilton, Owain Evans (2021a) TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv*, https://arxiv.org/abs/2109.07958

Stephanie Lin, Jacob Hilton, Owain Evans (2021b) TruthfulQA: Measuring How Models Mimic Human Falsehoods. https://github.com/sylinrl/TruthfulQA

Meta AI (2024a) Llama Responsible Use Guide. Meta, https://ai.meta.com/static-resource/july-responsible-use-guide

Meta AI (2024b) Llama 3.1 Acceptable Use Policy. Meta, https://llama.meta.com/llama3_1/use-policy/

MIT (2024) AI Risk Repository. Massachusetts Institute of Technology, https://airisk.mit.edu/

Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott (2024) Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RR-A2849-1. RAND Corporation, https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf

NIST (2020a) Control Baselines for Information Systems and Organizations. Special Publication 800-53B. National Institute of Standards and Technology, https://csrc.nist.gov/pubs/sp/800/53/b/upd1/final

NIST (2020b) Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. Special Publication 800-171 Rev. 2. National Institute of Standards and Technology, https://csrc.nist.gov/pubs/sp/800/171/r2/upd1/final

NIST (2021) Enhanced Security Requirements for Protecting Controlled Unclassified Information: A Supplement to NIST Special Publication 800-171. Special Publication 800-172. https://csrc.nist.gov/pubs/sp/800/172/final

NIST (2024) The NIST Cybersecurity Framework (CSF) 2.0. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

NIST (2023) NIST SP 800-53, Revision 5 Control Mappings to ISO/IEC 27001. National Institute of Standards and Technology, https://csrc.nist.gov/projects/olir/informative-reference-catalog/details?referenceId=99#/

OCC (2021) Model Risk Management. Office of the Comptroller of the Currency, https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

OECD (2022) Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. OECD Digital Economy Papers, https://doi.org/10.1787/7babf571-en.

OpenAI (n.d.) Model behavior feedback. OpenAI, https://openai.com/form/model-behavior-feedback

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

OpenAI (2023b) OpenAI Red Teaming Network. OpenAI, https://openai.com/index/red-teaming-network/

OpenAI (2023c) Coordinated Vulnerability Disclosure Policy. OpenAI, https://openai.com/policies/coordinated-vulnerability-disclosure-policy/

OpenAI (2024a) GPT-4o System Card. OpenAI, https://openai.com/index/gpt-4o-system-card/

OpenAI (2024b) Usage Policies. OpenAI, https://openai.com/policies/usage-policies/

OpenAI (2024c) Terms of Use. OpenAI, https://openai.com/policies/terms-of-use/

OpenAI (2024d) OpenAI Safety Update. OpenAI, https://openai.com/index/openai-safety-update/

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner (2023) How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. arXiv, https://arxiv.org/abs/2309.15840

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks (2023) Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. arXiv, https://arxiv.org/abs/2304.03279

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer and Jared Kaplan (2022a) Discovering Language Model Behaviors with Model-Written Evaluations. arXiv, https://arxiv.org/abs/2212.09251

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer and Jared Kaplan (2022b) Model-Written Evaluation Datasets. https://github.com/anthropics/evals

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, Rayid Ghani (2019) Aequitas: A Bias and Fairness Audit Toolkit. arXiv, https://arxiv.org/abs/1811.05577

Jeremy Scheurer, Mikita Balesni, and Marius Hobbhahn (2024) Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv, https://arxiv.org/abs/2311.07590

Victor Schmidt, Alexandra (Sasha) Luccioni, Alexandre Lacoste, and Thomas Dandres (2019) ML CO2 Impact. https://mlco2.github.io/impact/

Jonas Schuett (2022) Three lines of defense against risks from AI. *arXiv*, https://arxiv.org/abs/2212.08364

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, Allan Dafoe (2023) Model evaluation for extreme risks. *arXiv,* https://arxiv.org/abs/2305.15324

Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson (2024) The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv*, https://www.arxiv.org/abs/2408.12622

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa et al. (2022) Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv, https://arxiv.org/abs/2206.04615

Violet Turri, Rachel Dzombak (2023) Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. In AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), https://dl.acm.org/doi/10.1145/3600211.3604700

UNESCO (2023) Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54678/YTSA7796

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, Bo Li (2023) DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv, https://arxiv.org/abs//2306.11698

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin (2023) Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv*, https://arxiv.org/abs/2308.13387

Andy Zou, Zifan Wang, J. Zico Kolter, Matt Fredrikson (2023a) Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*, https://arxiv.org/abs/2307.15043

Andy Zou, Zifan Wang, J. Zico Kolter, Matt Fredrikson (2023b) LLM Attacks. https://github.com/llm-attacks/llm-attacks