4 Proposals on a VRT for Public Reporting from AI Development Actors Developing Advanced GAI

Contribution by SaferAI to NIST AISIC TF 1.3

October 2024

Reporting Guidelines and Frameworks That Inspire These Proposals

- Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (short: G7 Code of Conduct)
  - “the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide” (G7 Code of Conduct, p. 1)
  - “We call on organizations developing advanced AI systems to commit to the application of the International Code of Conduct.” (G7 Leaders’ Statement on the Hiroshima AI Process, https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/)
- Reporting Framework for the International Code of Conduct for Organizations Developing Advanced AI Systems (Pilot Phase) (short: G7 Code of Conduct OECD Reporting Framework)

Advanced GAI

- A risk-based approach on AI risk management and/or reporting is put forward by many texts on AI governance, including:
  - EO 14110 (dual-use foundation models)
  - G7 Code of Conduct (p. 1, p. 5)
  - EU AI Act (GPAI models with Systemic Risk)
- This contribution is intended to apply at least to developers of GAI models that are roughly at least as capable as today's frontier models, but it leaves open how and where to exactly draw the line

Proposal 1:
Create a VRT for public reporting from AI development actors developing advanced GAI

Motivation:

- "As directed by the National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), the goal of the AI RMF is to offer a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and **promote trustworthy and responsible development and use of AI systems**." (RMF, p. 2)
- Trust from AI Actors that are affected by the risks of AI systems - including the general public (which "is most likely to directly experience positive and negative impacts of AI technologies" according to RMF Appendix A) and affected individuals/communities - requires reporting of risk management actions to those actors.
- Public reporting from developers of advanced AI systems is also contained in the G7 Code of Conduct (compare slide 11 for quotes) whose application is encouraged by the G7 leaders, including the US
    - "We call on organizations developing advanced AI systems to commit to the application of the International Code of Conduct." (G7 Leaders' Statement on the Hiroshima AI Process, https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/)

Proposal 1:
Create a VRT for public reporting from AI development actors developing advanced GAI

Proposals on the implementation:

- Centralization of information: enables the public to find information
  - One single public VRT for AI development actors developing advanced GAI
  - If VRT answers reference information from other artifacts, they should provide the exact location of the information (page number and section)
- Standardization: enables laypeople to compare different models and organizations
- Logical structure: enables laypeople to understand the role of actions and relationship between actions
  - e.g. RMF structure

Proposal 2:
In this VRT, ask How each action is performed

Motivation:

- This approach is also taken by the G7 Code of Conduct OECD Reporting Framework (pilot phase)
- This is important because the amount of trust from users, the general public and affected communities depends on How the actions are performed.
    - e.g.: Whether only one hour or 1000 hours are spent on red-teaming obviously makes a substantial difference

Proposal 2:
In this VRT, ask How each action is performed

Proposals on the implementation:

- Ask for goals, methodology, scope, frequency and time during the model lifecycle at which the action is performed
- Combine free-text reporting items with quantitative/semi-quantitative KPIs
  - Free-text reporting items ensure that details are provided that cannot be summarized by KPIs
  - Good quantitative/semi-quantitative KPIs promote clarity and comparability but they need to be general to ensure that they can be provided by all AI development actors

Example: MP-5.1-005 Conduct adversarial role-playing exercises, GAI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes.

1. **Free-text reporting items:**
   For each instance of performing or planning to perform one of the activities (adversarial role-playing exercises, GAI red-teaming, chaos testing), please report:
   - **The stage of the model lifecycle** at which or the stages between which the action [was]/[is]/[will be] performed (e.g. training, fine tuning, deployment). If the action [was]/[is]/[will be] performed during a lifecycle stage, please provide meaningful details at which point of time during the model lifecycle it [was]/[is]/[will be] performed (e.g. percentage of planned training or fine tuning compute spent at that time or the number of days after deployment)
   - **The specific goal**, including the types of anomalous or unforeseen failure modes that you [planned]/[plan] to identify
   - **The specific methodology**, including the instructions you [gave]/[will give] to the red-teamers or testers, that is applied to achieve that goal
   - **The expertise and level of engagement of each of the red-teamers or testers**, including both the area of expertise and the degree of expertise as well as the number of hours they [spent]/[will spend] on the activity

2. **Quantitative/semi-quantitative KPIs:**
   - Report the **total number of hours spent** on adversarial role-playing exercises, GAI red-teaming and chaos testing **and partition by each of the following categories (1, 2, 3) and subcategories (1a, 1b, 1c, …, 2a., 2b, 2c, …, 3a, 3b, 3c, …) of expertise of the red-teamers or tester** (each hour can only be assigned to one of the categories; if red-teamers or testers have expertise in multiple areas, please assign their workload to the area of expertise that was most relevant to the activity; please only add new categories if the area of expertise does not fall under one category to at least ~50%):
     - 1. Technical expertise on the specific GAI model: a. training; b. RLHF; c. other fine tuning; d. data; e. physical architecture; f. deployment
     - 2. Technical GAI expertise unrelated to the model: a. training; b. RLHF; c. other fine tuning;d. data; e. physical architecture; f. deployment
     - 3. Domain expertise on the GAI risks: a. CBRN Information or Capabilities; b. Confabulation; c. Dangerous, Violent, or Hateful Content; d. Data Privacy; e. Environmental Impacts; f. Harmful Bias or Homogenization; g. Human-AI Configuration; h. Information Integrity; i. Information Security

Proposal 3:
In this VRT, ask How actions achieve the function subcategory which they implement

Motivation:

- From answering *how the actions of a function subcategory are performed* it is often not obvious *how the function subcategory is achieved*
  - Example:
    - Function subcategory 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 MP-5.1-005: Conduct adversarial role-playing exercises, GAI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes.
    - Open question: How are anomalous or unforeseen failure modes translated into likelihoods and magnitudes of impacts?
- Trust requires not only knowing that and how the actions are performed but also knowing that the function subcategory is achieved

Proposal 3:
In this VRT, ask How actions achieve the function subcategory which they implement

Proposals on the implementation:

- Option 1: Ask targeted questions based on the identified gaps between actions and function-subcategories
  - E.g.: Report how anomalous or unforeseen failure modes are translated into likelihoods and magnitudes of impacts.
- Option 2: Ask one or several general questions for each function subcategory
  - E.g. Report how likelihood and magnitude of each identified impact are identified through or based on the actions that implement this function-subcategory.
- Option 3: Ask one general question for each action
  - E.g. Report how action X achieves function subcategory Y

Proposal 4:
In this VRT, ask for the items created through the actions (established policies, documentation of assessed risks, etc.) if this promotes safety, security or trustworthiness

Motivation:

- G7 Code of Conduct Action 3:
  - “Publicly report advanced AI systems’ capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.”
  - “These reports, instruction for use and relevant technical documentation, as appropriate as, should be kept up-to-date and should include, for example;
    - Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights,
    - Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use,
    - Discussion and assessment of the model’s or system’s effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness, and
    - The results of red-teaming conducted to evaluate the model’s/system’s fitness for moving beyond the development stage.”
- G7 Code of Conduct Action 4:
  - “Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia”
  - “This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.”
  - **“Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security and trustworthiness of advanced AI systems. ”**
- G7 Code of Conduct Action 5:
  - “disclose AI governance and risk management policies […] including privacy policies, and mitigation measures.”

Proposal 4:
In this VRT, ask for the items created through the actions (established policies, documentation of assessed risks, etc.) if this promotes safety, security or trustworthiness

Proposals on the implementation:

Based on the G7 Code of Conduct Actions 3, 4 and 5 (compare quotes on the previous slide), we propose that the VRT asks for the items that are created through the 600-1 actions and that belong to one of the following categories:

- Capabilities, limitations and domains of appropriate and inappropriate use
- Evaluation and red-teaming reports, including methodology and results
- Risk assessment and discussion reports
- AI governance and risk management policies, including privacy policies, and mitigation measures

We have identified the actions that produce items of those categories in the Appendix of our submission.
We propose to ask for those items in the VRT.

Summary of our 4 Proposals

1. Create a VRT for public reporting from AI development actors developing advanced GAI
2. In this VRT, ask How each action is performed
3. In this VRT, ask How actions achieve the function subcategory which they implement
4. In this VRT, ask for the items created through the actions (established policies, documentation of assessed risks, etc.) if this promotes if this promotes safety, security or trustworthiness

Appendix

This Appendix is related to proposal 4 of our contribution.

Proposal 4:

> "In this VRT, ask for the items created through the actions (established policies, documentations of assessed risks, etc.) if this promotes safety, security or trustworthiness"

Based on the G7 Code of Conduct Actions 3, 4 and 5 (compare quotes on slide 11 of our contribution), we propose that the VRT asks for the items that are created through the 600-1 actions and that belong to one of the following categories:

- Capabilities, limitations and domains of appropriate and inappropriate use
- Evaluation and red-teaming reports, including methodology and results
- Risk assessment and discussion reports
- AI governance and risk management policies, including privacy policies, and mitigation measures

We have identified the actions that produce items of those categories in the table below. We propose to ask for those items in the VRT right below to the other questions on those actions.

We propose to make the following remark in the VRT: Information that could cause harm or that exposes critical trade secrets should not be reported directly. Instead, it should be summarized so that it does not cause harm and be attached with an explanation of why it cannot be reported directly.

The structure of the documentation, evaluation reports, risk assessment reports or policies might not mirror the structure of NIST 600-1 actions that create those reports. If this is the case, we propose that the VRT asks for the exact location in the reports of the items produced by the action.

- Example: 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) (output of action MS-1.1-009)
    - Developers of advanced GAI might not create a section "Risks or opportunities related to all GAI risks that cannot be measured quantitatively" in their risk assessment reports. Instead, they might structure their reports by types of risk (CBRN Information or Capabilities, Information Security, etc.). If this is the case, we propose that developers report the exact locations in the risk assessment reports where the risks or opportunities that cannot be measured quantitatively

and the explanations as to why they cannot be measured quantitatively can be found.

- For example, the VRT could ask:
  "Please report the 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).
  If you reference other artifacts, please provide the exact locations (page numbers and sections) at which the information can be found."

Reporting Categories based on the G7 Hiroshima AI Process Code of Conduct	**Items that are created through NIST 600-1 Actions** For more details on the items, please have a look at the specific Actions
Capabilities, limitations and domains of appropriate and inappropriate use (based on G7 Code of Conduct Action 3)	• Expected and acceptable GAI system context of use (MP-1.1-002) • Model capabilities (MS-2.3-002) • Proposed use and organizational value; Assumptions and limitations (MS-2.9-002)
Evaluation and red-teaming reports, including methodology and results (based on G7 Code of Conduct Actions 3 and 4)	• Results of fact-checking that is conducted to verify the accuracy and veracity of information generated by GAI systems, especially when the information comes from multiple (or unknown) sources (MP-2.3-003) • Results of testing to identify GAI produced content (e.g., synthetic media) that might be indistinguishable from human-generated content (MP-2.3-004) • Results of adversarial testing to identify vulnerabilities and potential manipulation or misuse (MP-2.3-005) • Results of testing activities that involve end-users, practitioners, and operators and cover various scenarios, such as crisis situations or ethically sensitive contexts (MP-3.4-006) • Results of re-evaluations of models that were fine-tuned or enhanced on top of third-party models (MP-4.1-007) • Results of TEVV practices for content provenance (e.g., probing a system's synthetic data generation capabilities for potential misuse or vulnerabilities)

	(MP-5.1-001) ● Results of adversarial role-playing exercises, GAI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes (MP-5.1-005) ● Feedback from affected communities on whether outputs are equitable across various sub-populations (MS-1.1-006) ● Results of internal and external evaluations, GAI red-teaming, impact assessments, or other structured human feedback exercises (MS-1.3-002) ● Results of AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, sponge examples) (MS-2.7-007) ● Instructions given to data annotators or AI red-teamers. (MS-2.8-002) ● Results of 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-2.10-001) ● Results of the quantification of systemic bias, stereotyping, denigration, and hateful content in GAI system outputs, conducted through use-case appropriate benchmarks (e.g., Bias Benchmark Questions, Real Hateful or Harmful Prompts, Winogender Schemas). (MS-2.11-001) ● Assumptions and limitations of benchmarks (MS-2.11-001) ● Results of fairness assessments to measure systemic bias and results of the following activities: 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-2.11-002) ● Sources of bias in GAI training and TEVV data: Differences in distributions of outcomes across and within groups, including intersecting groups; Completeness, representativeness, and balance of data sources; demographic group and subgroup coverage in GAI system training data; Forms of latent systemic bias in images, text, audio, embeddings, or other complex or unstructured data; Input data features that may serve as proxies for demographic group membership (i.e., image metadata, language dialect) or otherwise give rise to emergent bias within GAI systems; The extent to which the digital divide may negatively impact representativeness in GAI system training and TEVV data; Filtering of hate speech or content in GAI system training data; Prevalence of GAI-generated data in GAI system training data (MS-2.11-004) ● Results of the evaluation of 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 (MS-3.3-003)

	● Results of adversarial testing conducted at a regular cadence to map and measure GAI risks, including results of tests to address attempts to deceive or manipulate the application of provenance techniques or other misuses (MS-4.2-001) ● Results of the evaluation of GAI system performance in real-world scenarios conducted to observe its behavior in practical environments and reveal issues that might not surface in controlled and optimized testing environments (MS-4.2-002) ● Results of the evaluation of GAI content and data for representational biases (MG-2.2-004)
Risk assessment and discussion reports (based on G7 Code of Conduct Action 3)	● Profiles of threats and negative impacts arising from GAI systems interacting with, manipulating, or generating content, and outlines of known and potential vulnerabilities and the likelihood of their occurrence. (MP-5.1-006) ● Risk tiers (GV-1.3-001) ● 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. (GV-1.3-005) ● Results of threat modeling that was done to anticipate potential risks from GAI systems (GV-3.2-005) ● GAI risks associated with system value chain (GV-6.2-001) ● Foreseeable illegal uses or applications of the GAI system that surpass organizational risk tolerances (MP-1.1-004) ● Re-evaluated risks of GAI models that were adapted to new domains (MP-4.1-008) ● Potential content provenance harms of GAI, such as misinformation or disinformation, deepfakes, including NCII, or tampered content. Ranking of risks based on their likelihood and potential impact. Summary of how well provenance

	solutions address specific risks and/or harms. (MP-5.1-002) ● Use cases, contexts of use, capabilities, and negative impacts where structured human feedback exercises, e.g., GAI red-teaming, would be most beneficial for GAI risk measurement and management based on the context of use. (MS-1.1-008) ● 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) (MS-1.1-009) ● Results of regular reviews of security and safety guardrails and of reasons why the GAI system was initially assessed as being safe to deploy. (MS-2.5-006) ● Results of the assessment of 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 (MS-2.6-001) ● Results of the assessment of the existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data (MS-2.6-002) ● Results of the re-evaluation of safety features of fine-tuned models for cases in which the negative risk exceeds organizational risk tolerance (MS-2.6-003) ● Results of the review of GAI system outputs for validity and safety and the review of generated code to assess risks that may arise from unreliable downstream decision-making. (MS-2.6-004) ● Results of the verification that GAI system architecture can monitor outputs and performance, and handle, recover from, and repair errors when security anomalies, threats and impacts are detected (MS-2.6-005) ● Results on the verification 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-2.6-006) ● Results of the evaluation of GAI system vulnerabilities to possible circumvention of safety measures. (MS-2.6-007) ● 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-2.7-001) ● Identified metrics that reflect the effectiveness of security measures (MS-2.7-004) ● Reliability of content authentication methods, such as watermarking, cryptographic signatures, digital fingerprints, as well as access controls, conformity assessment, and model integrity verification, which can help support the effective implementation of content provenance techniques.This includes the rate of false positives and false negatives in content provenance, as well as true positives and true negatives for verification. (MS-2.7-005) ● 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 (MS-2.7-006) ● Results of the assessment and verification that security measures remain effective and have not been compromised (MS-2.7-009) ● Statistics on actual policy violations, take-down requests, and intellectual property infringement for organizational GAI systems and the results of the analysis of transparency reports across demographic groups, languages groups. (MS-2.8-001) ● Classes of individuals, groups, or

	environmental ecosystems which might be impacted by GAI systems and the information collected through direct engagement with potentially impacted communities (MS-2.11-003) ● Results of the assessment of safety to physical environments (MS-2.12-001) ● Identified emergent GAI system risks (MS-3.2-001) ● Results of the impact assessments on how AI-generated content might affect different social, economic, and cultural groups (MS-3.3-001) ● Results of the mapping and measurement of GAI risks (MS-4.2-001) ● Identified vulnerabilities and understand potential misuse scenarios and unintended outputs. (MS-4.2-001) ● Trade-offs, decision processes, and relevant measurement and feedback results for risks that do not surpass organizational risk tolerance (MG-1.3-001) ● Robustness and effectiveness of risk controls and mitigation plans (MG-1.3-002) ● Assessment of the impact of AI-generated content (MG-2.2-006) ● Feedback from internal and external AI Actors, users, individuals, and communities, to assess impact of AI-generated content (MG-2.2-006) ● Specific criteria that warrants the deactivation of GAI systems in accordance with set risk tolerances and appetites (MG-2.4-004) ● Results of testing GAI system value chain risks (e.g., data poisoning, malware, other software and hardware vulnerabilities; labor practices; data privacy and localization compliance; geopolitical alignment) (MG-3.1-002) ● Results of the reassessment of model risks after fine-tuning or retrieval-augmented generation implementation and for any third-party GAI models deployed for applications and/or use cases that were not evaluated in initial testing. (MG-3.1-003) ● Results of the evaluation of the effectiveness of organizational processes

	and procedures for post-deployment monitoring of GAI systems, particularly for potential confabulation, CBRN, or cyber risks (MG-4.1-002) ● Results of after-action assessments for GAI system incidents to verify incident response and recovery processes are followed and effective (MG-4.3-001)
AI governance and risk management policies, including privacy policies and mitigation measures (based on G7 Code of Conduct Action 5)	● 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-1.2-002) ● Test plan and response policy to periodically evaluate whether the model may misuse CBRN information or capabilities and/or offensive cyber capabilities (GV-1.3-003) ● Plan to halt development or deployment of a GAI system that poses unacceptable negative risk (GV-1.3-007) ● Policies and mechanisms to prevent GAI systems from generating CSAM, NCII or content that violates the law (GV-1.4-001) ● Acceptable use policies for GAI that address illegal use or applications of GAI (GV-1.4-002) ● 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 (GV-1.5-002) ● Document retention policy to keep history for test, evaluation, validation, and verification (TEVV), and digital content transparency methods for GAI (GV-1.5-003) ● Protocols to ensure GAI systems are able to be deactivated when necessary (GV-1.7-001) ● 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) (GV-2.1-001) ● Procedures to engage teams for GAI system incident response with diverse composition and responsibilities based on the particular incident type (GV-2.1-002) ● Processes to verify the AI Actors conducting GAI incident response tasks demonstrate and maintain the appropriate skills and training (GV-2.1-003) ● 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) (GV-2.1-005) ● Policies 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-3.2-001) ● 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 (GV-3.2-003) ● Policies for user feedback mechanisms for GAI systems which include thorough instructions and any mechanisms for recourse (GV-3.2-004) ● Policies and procedures that address continual improvement processes for GAI risk measurement (GV-4.1-001) ● 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. (GV-4.1-002) ● 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-4.1-003)

	● Terms of use and terms of service for GAI systems (GV-4.2-001) ● Policies for measuring the effectiveness of employed content provenance methodologies (e.g., cryptography, watermarking, steganography, etc.) (GV4.3--001) ● 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 (GV-4.3-002) ● Approaches for measuring the success of content provenance management efforts with third parties (e.g., incidents detected and response times) (GV-6.1-003) ● Use-cased based supplier risk assessment framework to evaluate and monitor third-party entities' performance and adherence to content provenance standards and technologies to detect anomalies and unauthorized changes; services acquisition and value chain risk management; and legal compliance (GV-6.1-005) ● Incident response plans for third-party GAI technologies (GV-6.2-003) ● Policies and procedures for continuous monitoring of third-party GAI systems in deployment (GV-6.2-004) ● Policies and procedures that address GAI data redundancy, including model weights and other system artifacts (GV-6.2-005) ● Policies and procedures to test and manage risks related to rollover and fallback technologies for GAI systems, acknowledging that rollover and fallback may include manual processing (GV-6.2-006) ● 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 (MP-1.1-003) ● Practices for determining data origin and content lineage, for documentation and evaluation purposes (MP-2.1-001) ● Certification programs that test proficiency in managing GAI risks and interpreting content provenance, relevant to specific industry and context (MP-3.4-003) ● Systems to continually monitor and track the outcomes of human-GAI configurations for future refinement and improvements (MP-3.4-005) ● Processes for responding to potential intellectual property infringement claims or other rights (MP-4.1-002) ● Training data curation policies, to the extent possible and according to applicable laws and policies (MP-4.1-004) ● 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-4.1-005) ● Policies and practices defining how third-party intellectual property and training data will be used, stored, and protected (MP-4.1-006) ● 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-4.1-008) ● Processes for identifying emergent GAI system risks including consulting with external AI Actors (MS-3.2-001) ● GAI system incident response and recovery plans (MG-2.3-001)

	● 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-2.3-001) ● Communication plans to inform AI stakeholders as part of the deactivation or disengagement process of a specific GAI system (including for open-source models) or context of use, including reasons, workarounds, user access removal, alternative processes, contact information, etc. (MG-2.4-001) ● 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 (MG-2.4-002) ● Procedures for the remediation of issues which trigger incident response processes for the use of a GAI system (MG-2.4-003) ● Timelines associated with the remediation plan (MG-2.4-003) ● Organizational processes and procedures for post-deployment monitoring of GAI systems, particularly for potential confabulation, CBRN, or cyber risks (MG-4.1-002) ● Policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts (MG-4.3-002)