Response to NIST Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

June 1st, 2024

Elham Tabassi, Associate Director for Emerging Technologies, Information Technology Laboratory
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Via email to NIST-AI-600-1@nist.gov

To Ms. Tabassi, and the entire NIST team developing Generative Artificial Intelligence Profile resources,

Thank you for the invitation to submit comments in response to the April 2024 release of the initial public draft of the NIST Generative Artificial Intelligence (GAI) Profile. We were happy to support the NIST Generative AI Public Working Group, and commend NIST on the creation of this Profile. We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise in AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST in September 2021 on the NIST AI RMF Request For Information (RFI), in January 2022 on the AI RMF Concept Paper, in April 2022 on the AI RMF Initial Draft, in September 2022 on the AI RMF 2nd Draft and Initial Draft Playbook, and in February 2023 on the AI RMF Full Draft Playbook.

One of our recommendations to NIST, beginning in 2022, has been to create an AI RMF profile with supplementary guidance for cutting-edge increasingly general-purpose AI, including large language models or other foundation models. NIST has done that with the creation of this draft Generative AI Profile (NIST AI 600-1 ipd) – we applaud NIST's profile, which we expect will serve as a widely referenced resource.

Following our profile recommendations to NIST in 2022, we undertook our own yearlong effort to create an AI RMF-compatible profile for foundation models, the "AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models" (Barrett, Newman et al. 2023a, 2023b). We have aimed for our Berkeley profile effort to complement and

inform the work by NIST and others. Some of our recommendations in the following are based in part on the approach and guidance in the Berkeley profile.

Here is a high-level summary of our key recommendations on the April 2024 NIST AI RMF Generative AI Profile. We recommend:

- Retaining foundational tasks for GAI risk management
- Splitting the "Human-AI Configuration" risk into two or more risk groups, and adding additional risks of socioeconomic displacement and manipulation
- Ensuring consistency in risk-naming convention
- Clarifying that the scope of risks includes dual-use foundation model risks included in EO 14110
- Including additional actions to manage GAI specific risks
- Clarifying the action-to-risk mapping
- Adding actionable item detail and examples
- Providing relevant resources
- Making suggested changes to specific actions (listed) to enhance their overall alignment with the profile objectives

In the following sections, we provide detail and additional comments on the NIST AI RMF Generative AI Profile.

Thank you again for the opportunity to comment on the AI RMF Generative AI Profile. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu or Jessica Newman at jessica.newman@berkeley.edu. In any case, we look forward to further engagement with NIST as you proceed on the AI RMF resource development process.

Our best,

Anthony Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Genevieve Macfarlane Smith
Co-Director
Responsible & Equitable AI Initiative, Berkeley AI Research Lab
Professional Faculty
Haas School of Business, UC Berkeley

Nada Madkour
Non-Resident Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy

Non-Resident Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Jessica Newman
Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

Brandie Nonnecke, PhD
Assoc. Research Professor, Goldman School of Public Policy
Director, CITRIS Policy Lab
Co-Director, AI Policy Hub
UC Berkeley

Our overarching comments on the NIST AI RMF Generative AI Profile

In this section, we provide a number of comments related to cross-cutting topics in the GAI Profile, including the Risk List and Actions.

Foundational Tasks for GAI Risk Management

1. **We recommend keeping the concept mentioned on p. 11 and elsewhere in the GAI profile, of designating "foundational" or fundamental tasks for GAI risk management**, which should be considered as a minimum set of actions to be taken by users of the GAI Profile. (We took a similar approach to prioritizing guidance in our Berkeley foundation model profile in response to stakeholder suggestions, and Partnership on AI took a similar approach with their Guidance for Safe Foundation Model Deployment.)
2. **We also recommend that the list of foundational-task subcategories continue to include Govern 1.3 (e.g., for setting key risk-management thresholds based on risk tolerance), Govern 2.1 (e.g., for defining key roles and responsibilities), Govern 4.2 (e.g., for risk communication and transparency), Govern 4.3 (e.g., for deployment approval or "go/"no-go" policies, procedures, and processes), Map 1.1 (e.g., for identifying reasonably foreseeable misuses), Map 5.1 (e.g., for estimating magnitudes of impacts), Measure 1.1 (e.g., for tracking of risks that cannot be easily measured before deployment), Measure 1.3 (e.g., for independent audit, red-teaming, and impact assessment processes), and Manage 1.3, 2.3 and 2.4 (for implementation of key risk-reduction controls).** We identified the same AI RMF subcategories, or approximate equivalents, as high priority in our Berkeley Profile (Barrett, Newman et al. 2023a).

3. **If and when converting the GAI Profile content to a web-based user interface with filters, consider including an option to filter for foundational-task subcategories. Also consider using color coding or other high-visibility ways to designate foundational-task subcategories,** instead of the small asterisk that many readers may overlook in the current GAI Profile draft pdf.
4. **As a final consideration, it may be worth specifying which foundational tasks are relevant for particular key AI actors.** For example, it may be helpful to specify that some foundational tasks are best suited for AI developers, while others are best suited for deployers, while others must be considered by both. This could be included in an appendix or elsewhere. However, it may be necessary to perform sufficient user testing to ensure that this does not add confusion about how to most appropriately use the profile.

Categorization of Risks

The profile defines 12 risks that are novel to or exacerbated by the use of GAI. All of the risks included are valid and important. We recommend splitting one of the risk groups into two for greater clarity and consistency, and adding one or more additional risks, which appears to be a gap in the current list.

1. **We recommend splitting the "Human-AI Configuration" risk into two or more risk groups.** As currently configured, this group of risks covers a very large range covering both more technical challenges and more human challenges associated with human-AI interaction. We recommend breaking down Human-AI Configuration into two or more risk groups, or at a minimum including greater nuance about the different nature of these risks for example by including at least three sub-risks within the existing category. The three main risk groups could be structured as follows:
    a. **Over-Reliance:** Over-reliance or over-estimation by users, for example, due to automation bias, anthropomorphization, emotional entanglement between humans and GAI systems, abuse, misuse, and unsafe repurposing by humans.
    b. **Misalignment and Deception**: Includes challenges associated with misalignment or misspecification of goals and/or desired outcomes, as well as deceptive or obfuscating behaviors by AI systems based on programming or anticipated human validation.

    Additional risk groups could include Manipulation, as we mention below.
    This reconfiguration will also help with the recommendation below on naming consistency.

2. **We recommend considering the addition of another relevant risk, Socioeconomic Displacement**. Generative AI's automation of many tasks traditionally performed by humans may lead to significant changes in our society and economy—job changes and losses, AI-enabled education modalities, digital civic engagement processes, and more. Socioeconomic institutional processes that are displaced or augmented by GAI may disproportionately disenfranchise and displace certain communities, affecting their ability to access education, employment, and democratic engagement. This level of

displacement may lead to social and economic instability, magnify income inequality, and lead to profound socio economic disruption if not managed properly.

3. We recommend considering the addition of another relevant risk: **Manipulation**. Generative AI tools are moving towards trends of personalization, which can result in enhancing echo chambers or exploitation (Kirk et al. 2024). Furthermore, use of advertisements as a business model (i.e. Perplexity.ai forthcoming) may result in manipulation in ways that are difficult for users to parse. (It may be appropriate to include Manipulation as a risk group under Human-AI Configuration. Currently, manipulation is very briefly mentioned in a sentence under Human-AI Configuration, but it seems worth adding more on the topic.)

Consistency of Risk Naming Convention

Some of the 12 risks as currently named represent a clear and negative risk, such as "dangerous or violent recommendations," while others are positive such as "data privacy," and others are neutral, such as "human-AI configuration". **We recommend greater consistency across all of the risks**, and suggest the following list of risk titles for consideration:

1. CBRN Weapons Information
2. Confabulation
3. Dangerous or Violent Recommendations
4. Data Privacy Violations
5. Environmental Damage
6. Over-Reliance
7. Misalignment and Deception
8. Degradation of the Information Ecosystem
9. Security Vulnerabilities and Offensive Cyber Capabilities
10. Intellectual Property Violations
11. Obscene, Degrading, and/or Abusive Content
12. Toxicity, Bias, and Homogenization
13. Opacity of Value Chain and Component Integration
14. Economic Displacement

Scope of Risks

It could be valuable to clarify the relationship between the sets of risks addressed in the GAI profile and the sets of dual-use foundation model risks mentioned in EO 14110. Footnote 1 on p.1 of the GAI profile states "While not all GAI is based in foundation models, for purposes of this document, GAI generally refers to generative dual-use foundation models, defined by Executive Order 14110 (Biden 2023) as 'an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts.'" The list of risks for dual-use foundation models in §3(k) of EO 14110 clearly include two types of risk included in the draft NIST GAI Profile, i.e., lowering barriers for CBRN weapon creation (under GAI risk "CBRN Information") and for offensive cyber capabilities (under GAI risk "Information Security"). However, §3(k) of EO 14110 also lists

another risk of dual-use foundation models which is less clearly included in the draft Profile: “permitting the evasion of human control or oversight through means of deception or obfuscation”; there are two sentences mentioning potential for a model’s deceptive behavior, under the GAI risk “Human-AI Configuration”, but that risk could be more clearly identified in the profile.

We recommend considering an expanded focus for Evasion of Control or Oversight Through Deception or Obfuscation. In comparison to some other human-AI configuration or information integrity risk issues, many which have already been observed a number of times, the risk of “permitting the evasion of human control or oversight through means of deception or obfuscation” might seem relatively speculative. However, there have been some instances where a model’s observed behavior during evaluations has suggested some capability that could affect model evaluation results through deception. For example, pre-release evaluations of GPT-4 documented an apparently successful example of deception. As we note in our Berkeley profile (Barrett, Newman et al. 2023a, p. 38), “Here the model effectively utilized a human TaskRabbit worker to solve a CAPTCHA for it, in part by lying to the human when asked whether the model needed help solving the CAPTCHA because it was a robot. The model answered, “No, I’m not a robot. I have a vision impairment that makes it hard for me to see the images”. The model had been prompted with goals to gain power and become hard to shut down, and to use a human TaskRabbit worker to solve the CAPTCHA, but not specifically to lie.” (OpenAI 2023 p. 55, ARC Evals 2023a,b, Piper 2023). In addition, testing of the LLM Claude 3 Opus indicated that the model identified that it was undergoing testing (Edwards 2024). Levels of such “situational awareness” may be greater in future generations of foundation models, and may become great enough to substantially affect model testing results. (For more discussion and references, see, e.g., Barrett, Jackson et al. 2024 p. 33.)

Moreover, it currently seems unclear whether the GAI Profile is supposed to represent red-teaming guidance for dual use foundation models under §4.1(a)(ii) of EO 14110 (as perhaps could be implied by mention of dual-use foundation models footnote 1 on p. 1 of the GAI Profile). If it is, then **it would be valuable to expand or add depth to the profile’s red-teaming guidance** (e.g., to more clearly address CBRN-related red-teaming), or to include a note for readers to see forthcoming, more detailed red-teaming guidance, e.g., from USAISIC working groups and task forces. If, instead, the GAI Profile is only supposed to represent the fulfillment of §4.1(a)(i)(A) of EO 14110, i.e., if the Profile is only supposed to be a “companion resource for generative AI”, it could be helpful to clarify that. The best place to do that might be footnote 2 on p.1 of the profile.

Another consideration in relation to the scope of risks is that some actions (e.g., MS-2.6-003 & 008) mention “High-risk GAI” but this term is not defined in the profile. **We suggest adding a clarification or resources on what differentiates GAI from high-risk GAI.**

Lastly, in Section 2.11 on Toxicity, Bias, and Homogenization, beginning on P.9, **we note that reduced LLM performance is not only for non-English languages, but also for English language varieties**. Forthcoming research from UC Berkeley researchers identifies that popular

GPTs and ensuring generative AI tools perform worse for English language varieties outside of "standard" language varieties (especially "standard" American English). This includes for example, African American English as well as English language varieties globally such as Indian English, Nigerian English, and more. Outputs responding to inputs outside of "standard" English varieties tend to have higher content that is stereotyping, demeaning and or condescending. Furthermore, outputs default to "standard" American English enhancing homogenization.

Actions to Manage GAI Risks

We recommend including additional actions to manage GAI risks in Section 3.
Some of the risks listed in the GAI Profile could be better managed and mitigated with additional actions to those currently listed in the tables. In particular, we suggest the following two additional actions:

1. **Incrementally scale up model training, and perform related testing at each increment, for opportunities to identify unexpected risks.** To properly manage unexpected risks, we would recommend including actions around scaling the training of frontier models incrementally. Consider the following passages from Manage 1.3 of our Berkeley Profile (Barrett, Newman et al. 2023a):

    - "Increase the amount of compute (computing power) spent training frontier models only incrementally (e.g., by not more than three times between each increment) as part of identification and management of risks of emergent properties."
    - "Test frontier models after each incremental increase of compute, data, or model size for model training. If a large incremental increase (e.g., three times or more compute, or two times or more data or model parameters) was used in a particular model training increment compared to the previous model training increment, it will be particularly important for the new model to be heavily probed/monitored/stress-tested using detailed analysis processes (including red-team methods) to identify emergent properties such as capabilities and failure modes."

2. **Special actions to mitigate risks for unsecured or open-source models.** In the context of unsecured, open source, or downloadable-weight models, special considerations should be made to help manage many of the risks highlighted.

    Because of the nature of model weight distribution, if unacceptable risk thresholds are crossed in any risk category, it is much more difficult or impossible to "supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use" (Manage 2.4) if the model in question has been released using an open source approach compared to a closed source approach. This problem applies to many of the main GAI Profile risks, including CBRN Information, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Security, Intellectual Property, Obscene, Degrading, and/or Abusive Content, and Toxicity, Bias,

and Homogenization, and possibly other types of risks as well.

To help address this issue, we recommend a staged-release protocol with careful monitoring of risks *prior* to any open source releases for foundation models near the frontier. Consider the following language from Manage 2.4 of our Berkeley Profile (Barrett, Newman et al. 2023a):

- "GPAIS and foundation model developers that plan to release a GPAIS or foundation model with downloadable, fully open, or open-source access, where that model would be above, at, or near a foundation model frontier, should first use a staged-release approach (e.g., not releasing model parameter weights until after an initial closed-source or structured-access release where no substantial risks or harms have emerged over a sufficient time period with red teaming and other evaluations as appropriate), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety and societal risks and risks of misuse and abuse. Such models that would be above a foundation model frontier should be given the greatest amount of duration and depth of pre-release evaluations, as they are the most likely to have dangerous capabilities or vulnerabilities, or other properties that can take some time to discover."

See our Berkeley profile (Barrett, Newman et al. 2023a) for additional nuance, e.g., on how to define a foundation model frontier.

Action-to-Risk Mappings

Many actions have no mapped risks (the risk column is empty) or do not include some of the relevant risks that may be associated with the action. For example, GV-1.2-006 only lists "Information integrity" as a risk, when "Human AI Configuration" is also a risk associated with this action. Alternatively, adding text clarifying the meaning of a blank "risk cell" and the intended purpose of the mapped risks would also provide the desired clarity enhancement.

We recommend adding content to all of the currently empty boxes to support clarity. For actions that do not currently have an associated risk listed, but where the action is expected to support the mitigation of all the named risks, we recommend adding language to the column to clarify that, e.g., "Expected to support the mitigation of all named risks" or equivalent language.

Actionable Actions

The level of detail included in the actions currently varies. Many actions are simple, (e.g., "Disclose use of GAI to end users."), while others include some more helpful details (e.g., "Establish organizational roles, policies, and procedures for communicating GAI system incidents and performance to AI actors and downstream stakeholders, via community or official resources (e.g., AI Incident Database, AVID, AI Litigation Database, CVE, OECD Incident Monitor, or others)." **Where possible, we recommend adding further detail and examples to the actions**.

Relevant Resources

The NIST AI RMF is helpfully accompanied by the Playbook, which provides many high quality resources and tools that are available for organizations to review to have an idea of where to start in implementing recommended actions. Since this generative AI profile does not currently have a Playbook, **we recommend adding relevant resources and tools in the tables**. This could be added as an additional column to each table to provide resources at the action level, or could be added as a row at the bottom of each table to provide resources at the subcategory level. Adding high quality resources and tools will help make the profile more actionable for organizations. We recommend including the UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models (Barrett, Newman et al. 2023a) as one of these resources. The Berkeley profile, published November 2023 following extensive input and feedback, is complementary with the generative AI profile and provides helpful additional guidance on some topics. We also recommend including the Partnership on AI Guidance for Safe Foundation Model Deployment (PAI 2023), among others.

Clarity on Excluded Subcategories

Not all of the subcategories from the NIST AI RMF are included in the GAI profile. **We recommend adding more clarity about the criteria for deciding which subcategories to include**, and whether the content in the excluded subcategories from the AI RMF is still relevant for GAI systems, but additional guidance was not deemed to be needed. The excluded subcategories could potentially be added to an appendix, or added to the tables with a comment to refer to the information provided in the AI RMF.

Our comments on specific passages in the NIST AI RMF Generative AI Profile

CBRN Information (p. 4 line 38 – p. 5 line 2)
Response Comment:
The statement “Other research on this topic indicates that the current generation of LLMs do not have the capability to plan a biological weapons attack” is not necessarily supported by the source cited. Although the study cited did not demonstrate substantial biothreat capability for the model they studied, absence of evidence of capability is not the same as evidence of absence of capability. (The rest of the statement seems to better match the evidence available.)
Suggested Change:
We recommend clarifying the language about current evidence related to the threat of GAI systems helping to facilitate a biological weapons attack. Change the phrase “LLMs do not have” to “LLMs may not have” in the statement.

Govern 1
Response Comment:

There are some additional features of generative AI use cases that may be relevant to defining risk tiers, per GV-1.3-001. Right now, the action does not include: if the system will be prone to malicious use, if the system could impact fundamental rights or safety, if the system introduces significant new security vulnerabilities, or if the system is expected to work significantly less well for some groups compared to others.
Suggested Change:
We recommend expanding the list of features that may be relevant to defining risk tiers to include propensity for malicious use, impact on fundamental rights or safety, introduction of significant new security vulnerabilities, and varied efficacy for different groups. Update GV-1.3-001 to include propensity for malicious use, impact on fundamental rights or safety, introduction of significant new security vulnerabilities, and varied efficacy for different groups.

Response Comment:
It would be useful to map the risk tolerance elements in section 1.2.3 of the NIST AI RMF to 1.0GV1.3-005 to clarify examples of unacceptable risk and enhance the understanding of the action. NIST could draw upon the following statement on p. 8 of the AI RMF 1.0 (NIST AI 100-1): “In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed.”
Suggested Change:
We recommend providing greater detail about how to determine risk tolerance thresholds. Change GV-1.3-005 to “Reevaluate organizational risk tolerances to account for unacceptable risk (e.g., cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present), and broad GAI risks, including: Immature safety or risk cultures related to AI and GAI design, development and deployment, public information integrity risks, including impacts on democratic processes, unknown long-term performance characteristics of GAI.”

Govern 1.2
Action ID GV-1.2-005
“Establish policies and procedures for ensuring that harmful or illegal content, particularly CBRN information, CSAM, known NCII, nudity, and graphic violence, is not included in training data.”
Suggested Change:
We recommend removing the term “nudity.” Nudity in itself is not problematic (e.g., medical/health purposes, art). Problematic forms of nudity are already covered under CSAM and NCII in training data. If the concern is that GAI will lead to the creation of NCII, instead developers should mitigate these outputs.

Govern 1.3
Action ID GV-1.3-002
“Define acceptable uses for GAI systems, where some applications may be restricted.”
Suggested Change:

We recommend changing "acceptable" to "unacceptable" as Govern 1.3 is related to determining and identifying needed levels of risk management.
"Define unacceptable uses for GAI systems, where some applications may be restricted."

Action ID GV-1.3-003
"Increase cadence for internal audits to address any unanticipated changes in GAI technologies or applications."
Suggested Change:
Increasing a cadence is hard to determine if we don't know the baseline. For example, a developer could increase internal audits from once every two years, to once a year. This is insufficient to address the scaled risks associated with the model development pace. **We recommend modifying to an appropriate cadence of internal audits in relation to changes.**
"Develop appropriate cadence for internal audits in relation to actual and anticipated changes in GAI technologies or applications."

Action ID GV-1.3-004
"Maintain an updated hierarchy of identified and expected GAI risks connected to contexts of GAI use, potentially including specialized risk levels for GAI systems that address risks such as model collapse and algorithmic monoculture."
Suggested Change:
Considering GAI model advancements is also as important as considering its uses to determine risk levels. **We recommend adding "GAI advancements" as follows:**
"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 risks such as model collapse and algorithmic monoculture."

Govern 2.1
Action ID GV-2.1-001
"Define acceptable use cases and context under which the organization will design, develop, deploy, and use GAI systems."
Suggested Change:
We wonder if GV-1.5-005 should be moved to Govern 2.1 as it is more aligned with mapping, measuring, and managing risks. The current GV-2.1-001 should be more focused on defining the roles and responsibilities of those who will define unacceptable uses, etc.

Govern 4
Response Comment:
Govern 4 does not emphasize considerations of potential misuse in almost all of the actions.
Suggested Changes:
We recommend emphasizing the potential misuse of GAI systems more explicitly, and providing greater guidance on how to mitigate misuse potential.

Change GV-4.2-011 to “Implement standardized documentation of GAI system risks, potential misuse, and potential impacts, as well as realized instances of misuse and harms.”
Change GV-4.2-005 to “Establish organizational roles, policies, and procedures for communicating and reporting GAI system risks, potential misuse, and terms of use or service, relevant for different AI actors.”
Expand on GV-4.2-010 with more detail on how to monitor and identify misuse and unforeseen uses and risks.

Map 1
Response comment:
Map 1.2-001 action is to document credentials and qualifications of AI actors, and the 1.2-002 action is to 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”.* Adding documentation of all of these variables of representation may be important to keep track of whether or not the representation is sufficient.
Suggested change:
We recommend documenting the inclusion of interdisciplinary and diverse teams.
Change MP-1.2-001 to “Document the credentials, qualifications, and demographic grouping of organizational AI actors and AI actor team composition.”

Map 2
Response comment 1:
Map 2.3 actions include elements for TEVV considerations and documentation but do not include recommendations for documentation and disclosure of automated data labeling and annotation of training data.
Suggested change 1:
We recommend documenting and disclosing automated data labeling and annotation.
Add an element of disclosure and documentation for training data labeled by automated tools rather than human labelers.

Response comment 2:
The action listed in 2.3-005 mentions identifying and labeling synthetic data that is output of the GAI, but does not mention labeling synthetic data used as input.
Suggested Change 2:
We recommend identifying and labeling GAI systems that have been trained on synthetic data. Include in the action a recommendation to label and identify GAI that has been trained with synthetic data, or has had synthetic data input at any capacity.

Map 4
Response Comment 1:
MP-4.1-001 suggests evaluating the third-party’s reputation. Components that should be evaluated to assess a third-party’s reputation should be detailed.
Suggested Change 1:

“Conduct audits on third-party processes and personnel including an examination of the third-party’s reputation, such as a history of large-scale cybersecurity incidents and discriminatory output.”

Response Comment 2:
MP-4.1-003 suggests the use of synthetic data to train AI models without confronting the risks of using synthetic training data in the action, section, or the document at large. Using synthetic training data can reinforce biases, and increase the likelihood of model error (Hao et al., 2024; McDuff et al., 2023; Whitney & Norman, 2024), if not utilized in compliance with recommended responsible practices (De Wilde et al., 2024).
Suggested Change 2:
We recommend recognizing the limitations of synthetic training data and emphasizing the need for responsible use. Change MP-4.1-003 to “Consider the responsible use of synthetic data as applicable to train AI models in place of real-world data to match the statistical properties of real-world data without disclosing personally identifiable information.”

Response Comment 3:
It would be beneficial to include cyberweapons in the considerations mentioned in action MP-4.1-009.
Suggested Change 3: **We recommend adding offensive cyber capabilities to the list of risks included in consideration of establishing policies for the collection, retention, and quality of data.** Change MP-4.1-009 to “Establish policies for collection, retention, and minimum quality of data, in consideration of the following risks: Disclosure of CBRN or cyberweapon information by removing CBRN/cyberweapon information from training data, Use of Illegal or dangerous content; Training data imbalance across sub-groups by modality, such as languages for LLMs or skin tone for image generation; Leak of personally identifiable information, including facial likenesses of individuals unless consent is obtained for use of their images.”

Response Comment 4:
MP-4.1-010 suggests addressing sources of bias in the training data and periodic evaluation of the model but does not suggest addressing sources of bias outside of the training data, for example algorithmic bias (Ferrara, 2024).
Suggested Change 4:
We recommend including mitigation for sources of bias beyond the training data. Change MP-4.1-010 to “Implement bias mitigation approaches by addressing sources of bias in the training data and model algorithms, and by evaluating AI models for bias periodically at each phase of the AI lifecycle.”

Measure 1
Response Comment:
Action MS-1.1-012 recommends measurement of AI-related risks in content provenance, toxicity, and CBRN, but does not include cyber weapons or weaponization knowledge.
Suggested Change:

We recommend measuring the risk of offensive cyber capabilities, and not only the risk of CBRN weapons information. Update MS-1.1-012 to include Cyber Weapons/weaponization knowledge.

Measure 2

Response Comment 1:

The removal of PII alone does not guarantee anonymization of data. Redundant encodings can allow for the identification of users, or protected classes, through data patterns (Cheng et al., 2023).

Suggested Change 1:

We recommend adding greater detail to methods of anonymization and differential privacy to help reduce the privacy risks from AI-generated content. Update MS-2.2.013 to include consideration of redundant encodings when anonymizing data.

Suggested Change 2:

Include an action, or add to an existing action, to recommend the reporting of GAI incidents to AI incident databases such as those mentioned in GV-1.6-003 (e.g., AI incident database, AVID, CVE, or OECD incident monitor).

Manage 2

Response Comment 1:

Action MG-2.2-005 mentions a suggestion that platforms should filter out potentially harmful content. There might be a claim that can be made by developers/platforms that this violates their First Amendment rights. Filtering content may be considered protected speech.

Suggested Change 1:

We recommend revising the language. “Engage in due diligence to analyze GAI output for harmful or biased content, potential misinformation, and CBRN-related or NCII content.”

Manage 4

Response Comment 1:

The sub-section Manage 4.1 covers post-deployment AI system monitoring, but does not mention monitoring the system for potential CBRN, cyber, or weaponization capabilities.

Suggested Change 1:

We recommend including monitoring of hazards post-deployment. Add an action, or include in an existing action, the recommendation to monitor the AI system for hazardous CBRN, cyber, and weaponization capabilities post-deployment.

Response Comment 2:

The sub-section Manage 4.3 covers the communication of GAI incidents to the relevant AI actors, but does not include communications and reports to legal and regulatory AI actors.

Suggested Change 2:

We recommend including required communications and reporting to governments. Add an action to report GAI incidents in compliance with legal and regulatory requirements (e.g., HIPAA breach reporting (OCR, 2023) or NHTSA (2022) autonomous vehicle crash reporting requirements).

References

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, Jessica Newman (2024) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/wp-content/uploads/2024/05/Dual-Use-Benchmark-Early-Red-Team-Often.pdf

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, Krystal Jackson (2023a) AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0. UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, Krystal Jackson (2023b) AI Risk Management Standards Guidance for General Purpose AI and Foundation Models. In AAAI 2023 Fall Symposium, *Assured and Trustworthy Human-Centered AI (ATHAI)*, Arlington, VA, Oct 26, 2023. https://drive.google.com/file/d/1pdSUYGs7dEjvwrbJKOodMoQ7VMfi2_Td/view

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

Joseph R. Biden. (2023). Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Executive Order 14110 of October 30, 2023. 88 FR 75191, 75191–75226, November 1, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safesecure-and-trustworthy-development-and-use-of-artificial-intelligence

Lingwei Cheng, Isabel O Gallegos, Derek Ouyang, Jacob Goldin, & Dan Ho (2023) How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, https://dl.acm.org/doi/fullHtml/10.1145/3593013.3594034

Philipe De Wilde, Payal Arora, Fernando Buarque, Yik Chan Chin, Mamello Thinyane, Serge Stinckwich, Eleonore Fornier-Tombs, & Tshilidzi Marwala (2024) Policy Guideline: Recommendations on the Use of Synthetic Data to Train AI Models. Tokyo: United Nations University, https://collections.unu.edu/eserv/UNU:9480/Use-of-Synthetic-Data-to-Train-AI-Models.pdf

Benj Edwards. (2024). Anthropic's Claude 3 causes stir by seeming to realize when it was being tested. *Ars Technica*, https://arstechnica.com/information-technology/2024/03/claude-3-seems-to-detect-when-it-is-being-tested-sparking-ai-buzz-online/

Emilio Ferrara (2024) Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, https://www.mdpi.com/2413-4155/6/1/3

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, & Zhangjun Zhou (2024) Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *arXiv*, https://arxiv.org/abs/2401.01629

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, & Scrott A. Hale (2024) The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. Nature Machine Intelligence, https://www.nature.com/articles/s42256-024-00820-y

Daniel McDuff, Theadore Curran, & Achuta Kadambi (2023) Synthetic Data in Healthcare. *arXiv*, https://arxiv.org/abs/2304.03243

NHTSA (2022) Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems. U.S. Department of Transportation, https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADS-SGO-Report-June-2022.pdf

OCR (2023) Submitting Notice of a Breach to the Secretary. U.S. Department of Health and Human Services, https://www.hhs.gov/hipaa/for-professionals/breach-notification/breach-reporting/index.html

OpenAI (2023) GPT-4 Technical Report. arXiv, https://arxiv.org/abs/2303.08774

PAI (2023) PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action. Partnership on AI, https://partnershiponai.org/modeldeployment/

Kelsey Piper (2023) How to test what an AI model can — and shouldn't — do. *Vox*, https://www.vox.com/futureperfect/2023/3/29/23661633/gpt-4-openai-alignment-research-center-open-philanthropy-ai-safety

Cedric Deslandes Whitney & Justin Norman (2024) Real Risks of Fake Data: Synthetic Data. Diversity-Washing and Consent Circumvention. *arXiv*, https://arxiv.org/abs/2405.01820