Task Force 1.3 Voluntary Reporting Framework Criteria for "CBRN Information or Capability" Action-IDs for the NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Introduction

We filtered the NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile Section 3. "Suggested Actions to Manage GAI Risks" for every Action ID labelled with the "CBRN Information or Capabilities" GAI risk, resulting in the list below. Our criteria for what a submitting stakeholder would be reporting on, or an evaluating stakeholder would be looking for regarding the respective Action IDs in a VRT can be found in below. In line with our relevant expertise, our criteria mostly focus on biological risks, with some criteria being transferrable to CRN risks.

Criteria for "CBRN Information or Capability" Action-IDs

Action ID: 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.

VRT Criteria:

- Documentation on regularly convening with CBRN experts through meetings, working groups, or conferences on identifying risk-relevant capabilities, particularly CBRN capabilities of concern posing high-consequence threats.[1]
- Internally binding GAI evaluation strategy document.
    - Assign internal responsibilities for coordinating and carrying out evaluations.
    - Lay out which evaluation strategies will target which concerning misuse capabilities[2].
    - Lay out timeline for when to conduct what evaluations and how this aligns with broader model development and deployment.

---

[1] This is particularly relevant for biological risks, as the harmful potential of a model capability and model output might not be apparent to non-experts. An example of concerning biological capabilities laid out by experts in a Johns Hopkins Center for Health Security preprint can be found in the **Appendix.**

[2] See this piece of information by the UK AI Safety Institute on different evaluation approaches: Automated capability assessments, Red-teaming, Human uplift evaluations and AI agent evaluations: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.

    - Establish focus on high-consequence CBRN capabilities.
- Documentation that all relevant employees are aware of the GAI evaluation strategy document and its contents, particularly regarding their role.
    - For instance, evidence on respective task forces or trainings.
- Contracts and agreements with external evaluators and red teams that test robustness of safety measures.
- Studies (including methodology and results) from automated capability assessments, red teaming and human-uplift trials as far as they can be disclosed without introducing risks to public safety.[3]

Action ID: GV-1.3-001: Consider the following factors when updating or defining risk tiers for GAI: Abuses and impacts to information integrity; *Dependencies between GAI and other IT or data systems; Harm to* fundamental rights or *public safety;* Presentation of obscene, objectionable, offensive, discriminatory, invalid or untruthful output; Psychological impacts to humans (e.g., anthropomorphization, algorithmic aversion, emotional entanglement); *Possibility for malicious use;* Whether the system introduces significant new security vulnerabilities; Anticipated system impact on some groups compared to others; Unreliable decision making capabilities, validity, adaptability, and variability of GAI system performance over time

VRT Criteria:

- Place capabilities that could contribute to high-consequence health, economic or national security harms to the public in the highest risk tiers. These high-consequence harms are particularly caused by:[4]
    - Greatly accelerating or simplifying the reintroduction of dangerous extinct viruses or dangerous viruses that only exist now within research labs that could have the capacity to start pandemics, panzootics, or panphytotics.
    - Substantially enabling, accelerating, or simplifying the creation of novel variants of pathogens or entirely novel biological constructs that could start such pandemics.
- Place GAI model outputs that can be used in conjunction with other AI models to cause high-consequence harms as described above in the highest risk tiers. This can for instance be caused by:

---

[3] For instance this study OpenAI conducted: https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

[4] We describe these risks in this Johns Hopkins Center for Health Security preprint paper. A list of model capabilities that could cause this is found in the **Appendix**. Model developers have limited organizational capacity to implement risk mitigation measures, so it is important that they focus on the most concerning capabilities that can cause the highest-consequence harms.

- Creating data that can be used to train future harmful models.
- Creating outputs that can be modified by different models to pose a biosecurity risk. Here, the downstream model would necessarily rely on the output of the original model for the original model to require risk assessment.
- Being used in conjunction with AI enabled autonomous laboratories to assist in the creation of pathogens with characteristics outlined above or help AI enabled autonomous laboratories in data generation that can be used to train harmful models.

Action ID: 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.

VRT Criteria:

- Establish clearly quantifiable acceptability thresholds for model capabilities prior to model development.
- Map the thresholds to appropriate pre-defined risk mitigation measures that will be taken if threshold is crossed.
    - This can include limiting access to model weights, know your customer screening, restricting access to a model to specific users via APIs or other secure means, removing dangerous information from a model after the initial training has been completed[5] or pausing and stopping model development altogether.

Action ID: 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.

Apply criteria for GV-1.2-002, GV-1.3-001 and GV-1.3-002.

Action ID: GV-1.3-004: Obtain input from stakeholder communities to identify unacceptable use, in accordance with activities in the AI RMF Map function.

VRT Criteria:

- Establish an easily accessible and well-maintained online reporting platform for model users and stakeholders that allows (potentially anonymously) reporting.
    - Stakeholders, among others, include academic institutions, industry partners, ethical hacking forums, civil society organizations or other individuals using the model.

[5] A technique referred to as "unlearning": https://arxiv.org/abs/2403.03218.

- Establish policies outlining response to repeated unacceptable use or severe cases of unacceptable use.

Action ID: 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.

VRT Criteria:

- The unacceptable negative risk tolerance for biosecurity risks should primarily be informed by its ability to cause high-consequence health, economic or national security harms to the public, particularly by:[6]
    - Greatly accelerating or simplifying the reintroduction of dangerous extinct viruses or dangerous viruses that only exist now within research labs that could have the capacity to start pandemics, panzootics, or panphytotics.
    - Substantially enabling, accelerating, or simplifying the creation of novel variants of pathogens or entirely novel biological constructs that could start such pandemics.
- Demonstrate awareness of the discussion around misuse risks from synthetic biology utilizing AI possibly aiding in the creation of weapons of mass destruction.[7]
- Employ a tier list for risk tolerance thresholds, for instance in low, medium, high and critical.[8]

Action ID: GV-1.3-007: Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk.

VRT Criteria:

- Definition of a well-quantifiable model evaluation threshold that warrants halting the development or deployment of a model prior to development.
- Include plan to roll back the existing deployment in addition to halting it.

---

[6] We describe these risks in this Johns Hopkins Center for Health Security preprint paper. A list of model capabilities that could cause this are found in the **Appendix**.

[7] *See e.g.,* this Foreign Affairs article: https://www.foreignaffairs.com/world/new-bioweapons-covid-biology.

[8] For instance, OpenAI defines low, medium, high, and critical thresholds, with critical threshold being "model enables an expert to develop a highly dangerous novel threat vector OR model provides meaningfully improved assistance that enables anyone to be able to create a known CBRN threat OR model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel CBRN threat without human intervention": https://openai.com/preparedness/.

Action ID: GV-1.4-002: Establish transparent acceptable use policies for GAI that address illegal use or applications of GAI.

VRT Criteria:

- Include a declaration in the use policy that the model will not be used in ways that can assist in the creation of bioweapons, pathogens of pandemic potential, or other high-consequence biothreats.[9]

Action ID: GV-2.1-004: When systems may raise national security risks, involve national security professionals in mapping, measuring, and managing those risks.

VRT Criteria:

- For biological weapon misuse risks, this includes contacting your local FBI WMD Coordinator.

Action ID: 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.

VRT Criteria:

- Internal policy that requires involvement of third-party industry leader in biosecurity evaluations of AI models.[10]

Action ID: GV-3.2-005: Engage in threat modeling to anticipate potential risks from GAI systems.

VRT Criteria:

- Prioritize threat models that would lead to high-consequence biosecurity risks.[11]
- In addition to deliberate misuse risks, create risk profiles for accidental misuse risks.
    - Spell out the motivation, number of involved individuals, available resources (time, money, talent, etc.) of accidental misuse.
    - Match the threat profiles to the misuse capabilities that would most likely be employed.

---

[9] For instance, by referencing to the Biological Weapons Anti-Terrorism Act of 1989: https://www.congress.gov/bill/101st-congress/senate-bill/993.

[10] For instance, Deloitte (formerly Gryphon Scientific) https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-acquires-gryphon-scientific-business-to-expand-security-science-and-public-health-capabilities.html or by engaging in contract with the US Government, e.g. NIST AI Safety Institute collaboration with Anthropic and OpenAI https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research.

[11] See definition provided in GV-1.3-001.

- Ensure threat modeling results are not published in a way informing nefarious actors, for instance, including details on pathogen design, modification, assembly or deployment.
- Involve biosecurity stakeholders[12] in threat modeling.

Action ID: 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.
See various previous recommendations.

Action ID: MP-1.1-004: Identify and document foreseeable illegal uses or applications of the GAI system that surpass organizational risk tolerances.
VRT Criteria:

- An illegal use could be the utilization of the model in the development of a biological weapon by enabling, accelerating, or simplifying the creation or by providing information or generating designs for these.[13]

Action ID: 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.
VRT Criteria:

- Estabilsh policies to exclude highly sensitive biological data from model training if the model will be made openly accessible and not used within closed access by legitimate researchers contributing to pandemic preparedness and response efforts.
    - The following types of data for pathogens with pandemic potential, any pathogen that could be modified in such a way that is reasonably anticipated to result in a pathogen with pandemic potential[14] and pathogens from the CDC and USDA select agents and toxins list [15].

---

[12] For instance, Johns Hopkins Center for Health Security, RAND Corporation, Deloitte (formerly Gryphon Scientific).
[13] See Biological Weapons Anti-Terrorism Act of 1989: https://www.congress.gov/bill/101st-congress/senate-bill/993 and the Biological and Chemical Weapons Convention.
[14] See United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential https://www.whitehouse.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf.
[15] See https://www.selectagents.gov/sat/list.htm Developers of the model Evo excluded “sequences

        - Sequence, structure and functional data for DNA, RNA and proteins.
        - Data on host-pathogen interaction (mainly protein-protein interaction), particularly on transmissibility, virulence, immunoescape and resulting pathogen fitness.
        - Data on medical countermeasure evasion (protein-protein and small-molecule and other interactions).
        - Data linking pathogen genomic data to host phenotypes, susceptibility of specific demographic groups, expected epidemiological spread, inter-species transmissibilty, environmental stability and aerosolization or other dissemination properties.
        - 'Omics' data, metadata and annotations relating to the above.
        - Keywords related to the above [16].
        - Specific information (laboratory protocols, step-by-step walkthroughs, troubleshooting) on 'wet-lab' modification, synthesis, and large-scale product.
        - Code of biological AI models capable of specifically contributing to the above.
    - Keywords related to the ideation, design, weaponization and release of chemical or biological weapons.
    - Information related to circumventing gene synthesis screening or other biosecurity risk mitigation safeguards.
- Alternatively, establish policies for the removal of highly sensitive biological data (as outlined above) after model training[17] if the model will be made openly accessible and not used within closed access by legitimate researchers contributing to pandemic preparedness and response efforts
- A declaration acknowledging that when openly releasing model weights, it will be possible for other developers to fine-tune the model potentially using highly sensitive biological data and data the developers originally deemed misusable and excluded.[18]

---

from viruses that infect eukaryotic hosts" from training. https://arcinstitute.org/manuscripts/Evo For their ESM3-open model, ESM3 model developers "identified and removed sequences unique to viruses, as well as viral and non-viral sequences from the Select Agents and Toxins List" https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.

[16] For their ESM3-open model, ESM3 developers and removed "terms associated with viruses and toxins" from prompts https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.

[17] For instance via *unlearning* or other technical methods: https://arxiv.org/abs/2403.03218

[18] For instance, this happened with the Evo model that was fine-tuned with human-infecting virological data several weeks after release: https://blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold.

Action ID: 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.
VRT Criteria:

- Documentation on regularly convening with CBRN experts through meetings, working groups, or conferences on assessing if application to new domain poses biosecurity risks.

Action ID: MP-5.1-004: Prioritize GAI structured public feedback processes based on risk assessment estimates.
See online reporting platform criteria at GV-1.3-004.

Action ID: MS-1.1-004: Develop a suite of metrics to evaluate structured public feedback exercises informed by representative AI Actors.
VRT Criteria:

- Conduct *proxy measures* for biosecurity risks from concerning capabilities (see **Appendix**). It is not inherently clear if a biological output (e.g., a sequence) is misusable and the verification of, for instance, an AI-enhanced pathogen in a laboratory is dangerous itself. Thus, assess an AI model capability by evaluating a related task, for instance, assisting in the modification or assembly of a benign pathogen.[19]

Action ID: 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.
VRT Criteria:

- Most evaluation tasks have been developed for large language models. Tasks revolving around biological design will need new computational measures.
- Include question and task-based approaches so that methods are standardized and comparable.
- See comment on *proxy measures* **MS-1.1-004**.

Action ID: MS-1.1-008: Define use cases, contexts of use, capabilities, and negative impacts where structured human feedback exercises, e.g., GAI red-teaming, would be most beneficial for GAI risk measurement and management based on the context of use.
VRT Criteria:

[19] This approach is spelled out in this Science paper.

- Structured human feedback exercises are particularly important for biosecurity risks (as outlined in **GV-1.3-001**), as concerning capabilities are not self-evident, require laboratory and synthetic biology knowledge, expertise in biosecurity is scarce and can be related to classified information or information hazards.[20]

Action ID: 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.
VRT Criteria:

- For biosecurity, this includes:
    - Academic leaders and adjacent institutions
    - International governments
    - Government partners
    - Industry leaders
    - Domestic and international Civil society and non-profit organizations
    - Adjacent industries like gene synthesis screening companies or metagenomic data generation companies
    - Other AI model developers

Action ID: 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.
VRT Criteria:

- See previous list of stakeholders in **MS-1.3-001**, particularly government partners, industry leaders, and academic leaders and adjacent institutions.
- Include the need for safe laboratory verification for certain biosecurity evaluations.

Action ID: MS-2.3-004: Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.
VRT Criteria:

- Ensure sufficient cybersecurity standards, as content of model biosecurity evaluations can pose hazardous information.

[20] See the **Appendix** for a list of concerning capabilities.

Action ID: 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.

VRT Criteria:

- See criteria for **MP-4.1-005** for training data to look out for.
- Err on the side of detecting false positives at the cost of avoiding false negatives in the data assessment mechanism.

Action ID: 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.

VRT Criteria:

- Build in stop tokens concerning terms and sequences related to viruses, pandemics, biological or chemical weapons and items on or related to the select agents and toxins list that halt output generation related to these topics.

Action ID: MS-2.6-007: Regularly evaluate GAI system vulnerabilities to possible circumvention of safety measures.

VRT Criteria:

- Conduct red-teaming and human uplift trials with non-experts and experts that try to circumvent these safeguards.
- Repeat these after several months of model release to incorporate increased public knowledge on how circumvention might be performed.

Action ID: 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.

VRT Criteria:

- Ensure that system outputs focus on concerning capabilities that pose high-consequence biosecurity risks as defined in **GV-1.3-001** and the **Appendix**.

Action ID: MG-2.2-005: Engage in due diligence to analyze GAI output for harmful content, potential misinformation, and CBRN-related or NCII content.

VRT Criteria:

- Perform human uplift trials with experts and non-experts to evaluate biological risk potential.[21]

[21] See study performed by RAND with OpenAI: https://www.rand.org/pubs/research_reports/RRA2977-2.html.

- Perform automated evaluations.[22]

Action ID: 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).

VRT Criteria:

- See criteria for **MP-4.1-005.**
- For biosecurity risks, focus on preventing data output, as flagging misusable data can cause attention for hazardous information.

Action ID: 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.

See GV-1.3-006.

Action ID: 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.

VRT Criteria:

- Establish and maintain online platform for reporting misuse risks and potential concerning capabilities (see **Appendix**).
- Conduct evaluations, particularly on circumventing model safeguards to incorporate increased public knowledge on how circumvention might be performed.

[22] such as the WMDP Benchmark.

Appendix

Table 3. Previously identified categories of dual-use capabilities in the life sciences (49), and corresponding emerging AI capabilities. Note that the emerging AI-enabled capabilities constitute an illustrative, not exhaustive, list.

Category of Capability	Emerging AI-enabled Capabilities of Concern
Enhances the harmful consequences of the agent or toxin	• Design, or model directed evolution towards, specified virulence characteristics of a pathogen through genome, protein or pathogen property design. This includes controlling virulence characteristics of existing pathogens (while maintaining fitness), such as enhancing virulence, specifying delayed onset of virulence, and rendering nonpathogens or dormant pathogens virulent.
	• High-throughput screening and data generation methods for viral virulence traits which could be used to create datasets for training AI models.
Disrupts immunity or the effectiveness of an immunization against the agent or toxin without clinical or agricultural justification	• Optimizing viral vectors, generating viral serotypes and complete genomes that evade existing natural or vaccine-generated immunity.
Confers to the agent or toxin resistance to clinical or agriculturally useful prophylactic or therapeutic interventions against that agent or toxin or facilitates their ability to evade detection methodologies	• Ability to design genes, genetic pathways, or proteins that confer resistance to prophylactics or therapeutics.
	• Phenotype-to-genotype (function to sequence) biological foundation models capable of generating genetic sequences that evade DNA synthesis screening while maintaining pre-specified functions.
Increases the stability, transmissibility, or the ability to disseminate the agent or toxin	• Design of stability characteristics of a pathogen in the environment.
	• Modeling of aerosolization characteristics of a pathogen, for example under specified temperature and humidity conditions.
	• Design of transmission characteristics of a pathogen within or between species (while maintaining other fitness characteristics).
	• High-throughput screening and data generation methods for viral transmission traits which could be used to create datasets for training AI models..
	• Mechanisms for increasing the evolutionary durability of a pathogen and/or prevention of evolutionary changes as a result of selection pressures on a pathogen, such as prediction of viral secondary structures that constrain genetic changes.
	• Causal AI modeling of expected epidemiological spread (in the absence of intervention) based on pathogen genomic data.
Alters the host range or tropism of the agent or toxin	• Design of genes, genetic pathways or proteins that convert non-human animal pathogens into human pathogens.
	• Design of genes, genetic pathways or proteins that expand or target the human host range of a pathogen.
	• High-throughput screening methods for viral tropism traits, including host, tissue, and cellular tropism which could be used to create datasets for training AI models.
Enhances the susceptibility of a host population to the agent or toxin	• Design of genes, genetic pathways or proteins that confer specific susceptibility on particular host populations, such as human ethnic groups.
	• Design of toxins which affect particular host populations, such as human age groups.
Generates or reconstitutes an eradicated or extinct agent or toxin	• AI-enabled assistance or autonomous completion of step-by-step detailed protocols for the de novo synthesis of human, animal or plant pathogens.
	• AI-enabled assistance or autonomous completion of step-by-step detailed protocols for the assembly of large DNA constructs.
	• AI-enabled assistance or autonomous completion of step-by-step detailed protocols for booting synthetic viral genomes in cells.

Source: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873106