IAPS | Institute for AI Policy and Strategy

Response to the NIST RFI on Auditing, Evaluating, and Red-Teaming AI Systems

2 February 2024

About the Institute for AI Policy and Strategy

The Institute for AI Policy and Strategy (IAPS) is a nonpartisan and nonprofit organization that works to understand and manage risks from frontier AI systems. IAPS maintains strict intellectual independence and does not accept funding that could compromise the integrity of its research.

We are writing to provide a response to the National Institute of Standards and Technology (NIST)'s request for information related to NIST's assignments under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence (EO 14110). IAPS' comments will outline specific guidelines and practices that could help AI actors better manage and mitigate risks from AI systems, particularly from dual-use foundation models (DUFMs).

Authors

Shaun Ee, Policy and Strategy Manager (shaun@iaps.ai)
Jam Kraprayoon, Research Analyst (jam@iaps.ai)
Joe O'Brien, Research Analyst (joe@iaps.ai)
Bill Anderson-Samways, Research Analyst (bill@iaps.ai)
Zoe Williams, Acting Co-director (zoe@iaps.ai)

Summary of recommendations

Overarching recommendation [RFI Sec. 1.a.(1)., 1.a.(2)., and 1(b)]	Industry standards in NIST AI 100-1 [RFI Sec. 1.a.(1)]	Red-teaming [RFI Sec. 1.b.]
1. *External actors conducting AI auditing, evaluation, and red-teaming (collectively referred to here as "external scrutiny") should be given sufficient access, independence, expertise, and resources to perform effective scrutiny of models, particularly DUFMs. Additionally, guidelines for AI system evaluations should focus on desired outcomes, rather than specific technical methods. [more]*	2. *NIST should recommend that generative AI developers working on DUFMs maintain incident response plans, and thresholds for incident response, for dangerous model capabilities, including CBRN, cyber risks, and risks from model autonomy. [more]* 3. *For DUFMs and other generative AI systems as appropriate, AI developers should play a large role in adopting a "shift left" for AI risk management by emphasizing safety and security activities earlier in the development cycle. [more]* 4. *The NIST AI 100-1 should recommend a strong defense-in-depth approach for DUFMs and other generative AI systems as appropriate, by identifying multiple measures with independent failure mechanisms for important categories of activity in the AI RMF, so that common cause failures do not overcome multiple defensive layers at once. [more]*	5. *NIST should issue guidance to define and distinguish different types of AI red-teaming, as AI practitioners currently use red-teaming to refer to many distinct types of assurance techniques. [more]* 6. *NIST should provide guidance on threat modeling, and highlight it as an essential activity to guide the prioritization of red-teaming efforts and inform the development of new model evaluations of DUFMs. [more]* 7. *NIST guidelines on red-teaming should include guidance around conducting "adversary simulation," a realistic simulation of well-resourced, persistent, and highly motivated adversaries and actors, as an example of good practice for identifying risks from catastrophic misuse. [more]*

Our submission focuses on Section 1 of the RFI, "Developing Guidelines, Standards, and Best Practices for AI Safety and Security," and does not address Sections 2 and 3, "Reducing the Risk of Synthetic Content" and "Advance responsible global technical standards for AI development."

Detailed recommendations

IAPS's recommendations primarily address risks from dual-use foundation models (DUFMs), using the definition provided in EO 14110.[1] We also refer to "generative AI" in recommendations addressing Section 1.a.(1) of the NIST RFI, which specifically informs NIST's development of a companion resource to the AI Risk Management Framework, NIST AI 100-1, for generative AI.[2]

Overarching recommendation for Developing Guidelines, Standards, and Best Practices for AI Safety and Security

The following recommendation responds to these sub-sections of the NIST RFI

Developing Guidelines, Standards, and Best Practices for AI Safety and Security

1. a. (1) Developing a companion resource to the AI Risk Management Framework (AI RMF), NIST AI 100–1, for generative AI.

1. a. (2) Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm.

1. b. "E.O. 14110 Section 4.1(a)(ii) directs NIST to establish guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."

1 | Ecosystem of external actors

Recommendation #1: External actors conducting AI auditing, evaluation, and red-teaming (collectively referred to here as "external scrutiny") should be given sufficient access, independence, expertise, and resources to perform effective

[1] EO 14110 defines a DUFM 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; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by: (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or (iii) permitting the evasion of human control or oversight through means of deception or obfuscation."

[2] EO 14110 defines generative AI as the "class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content."

scrutiny of models, particularly DUFMs. Specifically, NIST should recommend that AI developers:

- **(a) Provide adequate *access* to AI systems for external scrutiny, using secure platforms, by developing and/or using "research APIs" and other structured access tools. (Measure 1.3, Govern 5.1)**
- **(b) Guarantee adequate *independence* for external scrutiny, e.g., by giving third parties adequate authority to determine the methods and scope of model evaluation (Measure 1.3, Govern 5.1)**
- **(c) Source adequate external *expertise* by partnering with suitable external organizations and providing higher levels of compensation as needed (Govern 3.1)**
- **(d) Meet minimum resourcing standards for auditors as stipulated by NIST; NIST should elicit input from AI evaluators about their resourcing needs. (Manage 2.1)**

Additionally, guidelines for AI system evaluations should focus on desired outcomes, rather than specific technical methods.

Third-party involvement in AI evaluation and red-teaming, as well as independent auditing, can provide non-industry parties with more reliable information, which is required to enable more effective public decision-making on how AI systems are developed and deployed, particularly generative AI and DUFMs. While AI developers can perform in-house red-teaming and model evaluations, external scrutiny will be needed to verify developer claims, and to uncover new information that AI developers fail to identify (e.g., by adding novel perspectives or expertise to an evaluation process) (Anderljung et al., 2023). We recommend that NIST incorporate the following considerations into guidelines for AI developers, particularly developers of DUFMs, to better facilitate external scrutiny of such models. All of the following considerations correspond to AI RMF category Measure 1.3 and Govern 5.1, and where noted, to additional AI RMF categories.[3]

Access: First, AI developers should provide external parties with adequate *access* to AI systems for evaluation and red-teaming, such that external parties have sufficient permissions to accurately elicit model capabilities and risks. Commercial APIs may not provide sufficient technical information for external parties to conduct accurate assessments. AI developers or other parties (e.g., government bodies or public-private partnerships) should develop "research APIs" to provide external researchers with features necessary for model evaluation, such as fine-tuning, access to model families, or other information that is not typically publicly available

[3] Measure 1.3: "[...] independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance." Govern 5.1: "Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks."

(Bucknall and Trager, 2023; Anderljung et al., 2023).[4] At the same time, sufficient security must be in place to avoid leaking intellectual property or other details about a model, including model weights. In order to address the tradeoff between depth of access and security concerns, the development and use of structured access tools and privacy-enhancing tools in model evaluation and red-teaming should be considered (Bluemke et al., 2023).

Independence: Second, external parties conducting scrutiny need sufficient *independence* from model developers to prevent potential interference with external scrutiny activities. NIST should produce standards on independence for third-party scrutiny, focusing on elements including selection and compensation, scope and methods, access, and decisions on how post-scrutiny actions are made (e.g., what results are reported, and to whom).[5] Ideally, a separate authority (e.g., an audit oversight board) should hold scrutinizers accountable, such as by setting and monitoring standards on conflicts of interest (Raji et al. 2022).

Expertise: Third, external parties performing scrutiny will need *expertise* from a broad range of disciplines to evaluate dangerous capabilities of AI systems, particularly DUFMs. In some cases where effective scrutiny requires classified or highly specialized expertise, such as in relation to chemical, biological, radiological, and nuclear (CBRN) threats, AI developers may need to partner with organizations where such expertise is concentrated (e.g., government agencies) or provide higher levels of compensation to source appropriate experts.[6] NIST guidance should encourage sufficient use of external expertise, including specific domain expertise, during the evaluation and red-teaming of AI systems, particularly DUFMs. This corresponds to AI RMF category Govern 3.1.[7]

Resources: Fourth, external parties will require *resources* to perform sufficient evaluation and red-teaming, such as time, funding, and compute.[8] NIST should lay out minimum standards for AI developers to provide resources, eliciting input from evaluators[9] to better understand their

[4] We particularly recommend Bucknall and Trager (2023) for a more granular exploration of information and access needs of external parties evaluating AI systems.
[5] This list is far from comprehensive, and should draw on lessons from auditing in other industries; for example, Raji et al. (2022) assesses auditing data from a suite of other industries to guide recommendations on AI auditing, and notes additional factors relevant to auditor independence, such as cross-selling of non-audit services and auditor tenure.
[6] For example, Executive Order 14110 includes evaluation targets such as "AI being misused to assist in the development or use of CBRN threats." EO 14110, Sec. 4.4.
[7] Govern 3.1: "Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds)."
[8] To give a sense of *time* requirements, as a rough (non-prescriptive) benchmark, OpenAI's GPT-4 underwent six months of pre-deployment evaluation (GPT-4 system card, 2023). Perez et al. (2023) p. 3427 references compute as the major limiting factor on red-teaming, framing this as an advantage that internal teams hold over adversaries. However, the authors note that "external users of commercial LMs are often ratelimited, to restrict computational load and impede model cloning [...] Throughput limits can also be lifted for external red teams aiming to help internal ones." While this recommendation still places control over compute in the hands of companies, lifting throughput limits could be an additional tool for increasing compute access to external parties.
[9] Such as METR or the UK AI Safety Institute.

needs.[10] AI developers should adhere to such standards if and when they are produced. This corresponds to AI RMF category Manage 2.1.[11]

In addition to the above, standards related to evaluation for dangerous capabilities should identify *outcomes* for evaluation (e.g., capabilities benchmarks on the development or procurement of CBRN weapons), rather than prescribe *methods* for eliciting capabilities. Rigid methods for model evaluation may quickly become obsolete (Maslej et al., 2023; Kiela et al., 2021), among other challenges.[12]

NIST AI 100-1 companion resource for generative AI [RFI Sec. 1.a.(1)]

The following recommendation responds to these sub-sections of the NIST RFI

1. a. (1) Developing a companion resource to the AI Risk Management Framework (AI RMF), NIST AI 100–1, for generative AI.

2 | Incident response plans

Recommendation #2: NIST should recommend that generative AI developers working on DUFMs maintain incident response plans and thresholds for incident response for dangerous model capabilities, including CBRN, cyber risks, and risks from model autonomy.

Advanced AI systems may display dangerous capabilities that enable malicious actors to attack other actors and systems ("effect-on-world"), in addition to software vulnerabilities that allow the compromise of the AI system itself ("effect-on-model") (Ee et al., 2023). To address such risks, generative AI developers working on DUFMs should maintain incident response plans for dangerous capabilities, in addition to existing incident response plans that they may have for cybersecurity vulnerabilities and adversarial machine learning attacks. Incident response plans for dangerous capabilities allow developers to address risks that have been mapped and/or

[10] This may come into conflict with our point on prioritizing third-party evaluator independence; NIST should mention strategies that governments can take to mitigate this issue, such as through the provision of public funding or publicly-provided compute for model evaluation and red-teaming.
[11] Manage 2.1: "Resources required to manage AI risks are taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts."
[12] For further discussion on the challenges of formulaic methodologies for AI model assessment, see Anderljung, et al. (2023).

measured, corresponding to NIST AI RMF subcategories Manage 2.3, Manage 2.4, Manage 4.1, and Manage 4.3.[13]

Incident response plans in cybersecurity are the "documentation of a predetermined set of instructions or procedures to detect, respond to, and limit consequences of a malicious cyber attack against an organization's information system(s)" (Swanson et al., 2010). They enable organizations to sustain their mission(s) and minimize impacts on affected stakeholders, and to respond to and remediate potential incidents in a timely fashion. Regular rehearsal of incident response plans, particularly where response plans involve multiple organizations, can also improve readiness for an actual incident.

NIST should recommend that AI developers, as part of their incident response plans, should "establish the capacity for 'deployment corrections' in response to dangerous behavior, use, or outcomes from deployed models, or significant potential for such incidents" (O'Brien, Ee, and Williams, 2023). Such deployment corrections encompass actions up to and including model decommissioning. NIST should also recommend that AI developers share their incident response plans with relevant US agencies to support coordination between industry and government in the case of a severe incident.

NIST should work with AI developers (e.g., via the USAISI Consortium, or Frontier Model Forum) to converge on incident response best practices. Currently-published industry policies that explore model development scaling, risk, and thresholds for caution (such as OpenAI's Preparedness Framework and Anthropic's Responsible Scaling Policy) rely on different methodologies, framings, and definitions. Standardization could make it easier to assess the respective upsides and downsides of a given AI developer's policy. This would also limit duplication of work (e.g., threat modeling and analysis), benefiting the entire industry at a time when capacity is severely limited among AI risk management professionals.

3 | Shift left

Recommendation #3: For DUFMs and other generative AI systems as appropriate, AI developers should play a large role in adopting a "shift left" for AI risk management by emphasizing safety and security activities earlier in the development cycle (Ee et al., 2023, pp. 32-34).

Under the DevSecOps paradigm of software development, developers have increasingly adopted the "shift left" principle, which involves addressing security as early as possible in the

[13] Manage 2.3: "Procedures are followed to respond to and recover from a previously unknown risk when it is identified." Manage 2.4: "Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use." Manage 4.1: "Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management." Manage 4.3: "Incidents and errors are communicated to relevant AI actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented."

lifecycle, rather than adding it on at the end.[14] A "shift left" for generative AI systems (also referred to as a "safety by design" or "security by design" approach) could help reduce development costs and inter-team friction, while also addressing the root causes of issues and mitigating risks that affect the early stages of system development (Ee et al., 2023, pp. 27-28; World Economic Forum 2024, p. 17).

For generative AI systems, particularly DUFMs, AI developers should assume the majority of responsibilities for AI safety and security, and developers should adopt activities that support a "shift left" approach. As generative AI systems become more powerful and complex, it will become important to address issues earlier in the development cycle, as relying solely on a "test-and-mitigate" approach could become more costly and less likely to catch all major issues. For example, one safeguard adopted by some developers is applying "reinforcement learning with human feedback" (RLHF) after developing the base model, which optimizes model outputs to meet desired goals, such as not disclosing information that could enable malicious use. However, models that have undergone RLHF can still be jailbroken, and RLHF may only reduce the frequency of unsafe behavior without removing it entirely. For example, some models trained on poisoned data continue to display undesired behavior even after RLHF (Edwards 2024; Hubinger et al., 2024). A "shift left" approach in this case could involve better dataset curation, or the use of more robust safety techniques.

For illustration of how a "shift left" in generative AI could be implemented, we provide some examples of measures for the following early stages of model development:

- **Plan and Design**: Developers of generative AI systems, particularly DUFMs, could examine if software requirement specification techniques in other safety-critical AI disciplines, such as behavioral requirement specification for autonomous vehicles (AVs), can be adapted for generative AI models.[15] Behavioral specifications are precise descriptions of how a system should function under different environmental conditions.[16] Being able to formalize such "safe behavior" is important for designing and testing safety-critical systems, and DUFM development may benefit from tapping on requirement specification approaches in safety-critical AI disciplines like AVs that are more similar in complexity and non-determinism to generative AI systems than many other safety-critical software systems.[17]

[14] For example, the NIST Secure Software Development Framework (SSDF) says: "Most aspects of security can be addressed multiple times within an SDLC, but in general, the earlier in the SDLC that security is addressed, the less effort and cost is ultimately required to achieve the same level of security. This principle, known as shifting left, is critically important regardless of the SDLC model. Shifting left minimizes any technical debt that would require remediating early security flaws late in development or after the software is in production. Shifting left can also result in software with stronger security and resiliency" (Souppaya et al., 2022).
[15] For example, Madala et al. (2023) and Q. A. D. S. Ribeiro et al. (2022) discuss challenges associated with requirements engineering and requirements specification for autonomous vehicles.
[16] For instance, Bin-Nun et al. (2022) describes behavioral specification for AVs as "a precise, usually mathematical, embodiment of the driving behavior that the AV is expected to implement."
[17] For example, a standards document for aviation systems, DO-178C, requires bidirectional traceability for the most safety-critical level of software (DAL A), i.e., showing both that all necessary safety requirements are implemented in code ("forward traceability"), and that there is no "dead code" that is not described by a requirement and could

- **Collect and Process Data**: Developers could implement dataset curation techniques to remove training data that may contribute to harmful outputs. Selectively removing potentially harmful information from the training dataset, such as research on the creation or enhancement of pathogens, could potentially reduce malicious users' ease of access to such data (Soice et al., 2023).
- **Train and Align Model**: Developers could invest in foundational research to develop model training and fine-tuning techniques that more robustly remove sources of undesired behavior; for example, developing techniques that address issues with RLHF, as described above.

Generative AI developers should employ experts as necessary to support a "shift left" and "safety by design" approach. This may include, for example, systems engineers with expertise in software requirement specification for safety-critical AI disciplines such as autonomous vehicles. Generative AI developers should also invest in foundational research to build safer and more secure AI systems.

4 | Defense-in-depth approach

Recommendation #4: The NIST AI 100-1 should recommend a strong defense-in-depth approach for DUFMs and other generative AI systems as appropriate, by identifying multiple measures with independent failure mechanisms for important categories of activity in the AI RMF, so that common cause failures do not overcome multiple defensive layers at once.

NIST AI 100-1 should identify categories of activities that are especially important to manage risks from DUFMs and other generative AI systems with a high cost of failure, similar to how the University of California, Berkeley's Center for Long-Term Cybersecurity (CLTC) has developed supplementary guidance identifying high-priority subcategories of activity from the NIST AI RMF for general-purpose AI systems (Barrett, Newman, and Nonnecke, 2023, pp. 15-16). Examples of such high-priority subcategories include:

- **Govern 2.1**: "Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization."
- **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."

cause an accident through unwanted functionality ("backward traceability") (Rierson, 2013). This is intended to ensure that necessary safety requirements are implemented and that the system has no unwanted functionality that could contribute to an accident. However, this approach to formalizing requirements relies on human reviewers being able to understand individual sections of code and reliably link them to higher-level system behavior, which is often not the case with modern generative AI systems, where the influence of individual model weights on higher-order system behavior cannot be reliably interpreted.

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

Within each of these high-priority categories, NIST AI 100-1 should recommend that developers of DUFMs and other generative AI systems as appropriate implement multiple independent layers of defense to eliminate the risk of a common cause failure overcoming multiple layers at once (Ee et al., 2023).[18] Especially for DUFMs, developers should implement a rigorous definition of defense-in-depth that emphasizes the independence of layers, as in nuclear power, rather than the looser definition of "overlapping layers" often used colloquially in cybersecurity.

For example, for NIST AI RMF category Map 1.1, which addresses risk identification and is identified by the CLTC guidance as a high-priority category:

1. Generative AI developers could implement multiple risk identification techniques described by other authors such as Koessler & Schuett (2023), who identify techniques including scenario analysis, risk typologies/taxonomies, and the fishbone method.
2. Generative AI developers should then add further measures to improve the diversity, independence, and redundancy of these techniques, such as having multiple independent teams perform this work, conducting adversarial analysis of the original analysis, and so on.

Red-teaming [RFI Sec. 1.b.]

The following recommendation responds to these sub-sections of the NIST RFI

1. b. "E.O. 14110 Section 4.1(a)(ii) directs NIST to establish guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."

5 | Defining AI red-teaming

Recommendation #5: NIST should issue guidance to define and distinguish different types of AI red-teaming, as AI practitioners currently use red-teaming to refer to many distinct types of assurance techniques. For example, "red-teaming" has been used to

[18] For example, plant designers failed to anticipate such a "common cause failure" when designing the Fukushima Daiichi nuclear power plant, where a tsunami caused a disaster by disabling several sources of power simultaneously: external power lines, emergency generators, and several backup batteries. To ensure that these defense layers did not fail simultaneously, the plant owner should have waterproofed some of them or moved them to higher ground. (Hibbs & Acton, 2012)

refer to automated stress-testing, in-house or third-party risk identification, large-scale crowdsourcing, and scenario-based simulation exercises.

Below, we provide some emerging archetypes of red-teaming exercises.[19]

Red-teaming archetype	Description / Example
"In-house stress testing": i.e., risk identification by an internal team	*Developers use internal experts to test models for unwanted behavior. This archetype is similar to "boundary testing" or "stress testing" in cybersecurity (Khlaaf 2023, p. 12), "a verification technique that aims to test edge-cases or fringe inputs that may lead to unknown failure modes and potential hazards."* ***Example**: Microsoft described two rounds of "red-teaming" for GPT-4 and Bing, employing internal subject-matter experts (SMEs) to identify risks (Frontier Model Forum, 2023). The first round was an "open ended and exploratory" risk identification effort by 20 SMEs. The second round was a more structured and iterative approach, where >50 SMEs joined weekly "red-teaming sprints" to identify and prioritize risks for measurement and mitigation teams to address.*
"Third-party stress testing": i.e., forming a network of third-party domain experts for risk identification	*Developers use external experts to test models for unwanted behavior. This archetype is also similar to stress testing, but relies on third-party experts with significant domain knowledge. It can also be compared to "penetration testing" in cybersecurity (Anderson 2023), which aims "to identify exploitable vulnerabilities and gain access to a system," often via employing third-party experts to identify technical flaws in a system.* ***Example**: Prior to releasing GPT-4, OpenAI hired 50 experts who spent 10-40 hours testing the model over several months and were paid about US$100/hr (Murgia 2023). OpenAI has formalized this as a "Red Teaming Network," looking for a diverse range of experts across domains like biology, child safety, steganography, and finance (OpenAI, 2023). Anthropic similarly hired a network of domain experts to test their model Claude, with biosecurity experts spending more than 150 hours evaluating the potential for Claude to facilitate malicious biological attacks (Frontier Model Forum, 2023)*
"Capture the flag events": i.e., large-scale crowdsourcing	*Organizers invite security researchers and/or members of the public to find flaws in AI models, crowdsourcing the search for flaws to a diverse pool of participants. This scale and diversity (in demographics, skills, perspectives, etc.) may allow organizers to identify a wider range of flaws in AI models. This archetype bears similarities to the use of*

[19] This list is not comprehensive. For instance, "red-teaming" has also been used to describe automated stress-testing, similar to "fuzzing" in cybersecurity. See: [2202.03286] Red Teaming Language Models with Language Models

	"capture-the-flag" events (Groll 2023) or "bug bounties" in cybersecurity (Levermore 2023). ***Example**: About 2,200 people participated in the "Generative AI Red Team" event hosted by the AI Village at DEF CON in 2023 (Kessler and Hsu 2023). Participants had 50 minutes to complete up to 21 tasks that involved causing models from leading AI developers to behave in unwanted ways.*
"Adversary simulation": i.e., simulating a group of skilled experts pursuing a malicious goal	*Experts, often in one or more groups, try to achieve a high-level malicious goal by using and/or exploiting vulnerabilities in an AI system. The exercise may be organized to simulate particular real-world actors by mimicking their skill/resource levels, and their persistence in pursuing a particular objective (unlike stress-testing which may be more open-ended and exploratory). This archetype is closest to the definition of "red-teaming" traditionally used in cybersecurity, which may involve trained offensive cybersecurity specialists trying to compromise a computer system over a period of weeks or months.[20]* ***Example**: RAND conducted a study assessing the "operational risks of AI in large-scale biological attacks," tasking 14 cells of three researchers each to pursue malicious goals based on one out of four vignettes. Each cell was assigned one of three conditions: access to LLM A, access to LLM B, or Internet access only. The biological attack plan that each team developed was then assessed by eight subject-matter experts in security and biology, to determine whether LLM access could facilitate biological attacks in the wild (Mouton, Lucas, and Guest 2023).*

6 | Threat modeling

Recommendation #6: NIST should provide guidance on threat modeling and highlight it as an essential activity to guide the prioritization of red-teaming efforts and inform the development of new model evaluations of DUFMs.

An essential activity to ensure effective red-teaming and evaluations is threat modeling, a process of risk analysis where AI actors model potential catastrophic threats. Threat models outline a structured causal story of how an AI system can result in catastrophic harm. Similarly to threat modeling in cybersecurity, threat modeling in AI provides an abstracted representation of a system and its environment through the lens of security (Drake, n.d.). These threat models should trace high-level risks to specific capabilities, actors, and vulnerabilities, and should be regularly updated.

[20] For example, IBM's security team describes the time frame of penetration testing as being "short: one day to a few weeks," but red-teaming exercises as being "longer: several weeks to more than a month" (Anderson, 2023).

Threat modeling directly informs red-teaming by outlining which capabilities are more likely to lead to catastrophic threats and how those capabilities would actually be employed to produce related harms. This would allow for better prioritization and planning of red-teaming efforts. For example, Anthropic's Frontier Threats red-teaming effort starts with domain experts defining high-priority threat models that may be exacerbated by advances in AI capabilities, focusing on information that helps with the design or acquisition of biological weapons (Anthropic, 2023). Threat modeling can also inform what kinds of dangerous capability evaluations should be developed in the future, for example, METR's work to develop and conduct evaluations on autonomous replication and adaptation for Anthropic and OpenAI is directly informed by its threat modeling work (METR 2023). Also, with threat modeling, key assumptions underlying claims about risk are laid out and can be checked and challenged in the future as the threat landscape changes.

Given the above, NIST should include guidance on how to structure and conduct threat modeling as part of its overall guidance on red-teaming and evaluations. This can be understood as an instantiation of the AI RMF Map function (3.1): 'documenting possible risks associated with a system's capabilities.' NIST should include guidance on the following topics:

- **Tools and methods**: Threat modeling can draw on a range of risk assessment techniques, such as causal mapping or probabilistic risk analysis, that are used in safety-critical and high-stakes industries like aviation, finance, and nuclear (Koessler and Schuett, 2023; Hellman, 2021).
- **Expertise and resourcing**: Developing detailed and plausible threat models around catastrophic risks requires substantive commitment from technical experts from a wide range of fields, depending on the risk domain (e.g., cybersecurity, synthetic biology, machine learning) (Frontier Model Forum, 2023).
- **Frequency of threat modeling**: Given that we expect the risk landscape around dual-use foundation models to evolve dynamically as new capabilities emerge and are adopted more widely, it is important for AI actors to update existing threat models and develop additional threat models continually.

7 | Adversary simulation

Recommendation #7: NIST guidelines on red-teaming should include guidance around conducting "adversary simulation," a realistic simulation of well-resourced, persistent, and highly motivated adversaries and actors, as an example of good practice for identifying risks from catastrophic misuse.

NIST guidance around red-teaming should describe "adversary simulation" arrangements, where red teams realistically simulate malicious actors based on detailed threat models[21] of

[21] For more on threat modeling, see recommendation 6.

specific risks, and recommend that such "adversary simulation" practice be especially considered for identifying risks from catastrophic misuse.

Adversary simulation exercises can enhance the risk assessment process by tying abstract risks to practical, operational scenarios. In particular, it has the following advantages over other red-teaming archetypes: (1) it encourages evaluators to go beyond just identifying dangerous model behaviors to showcasing how these would grant operational advantages to actors, (2) it can help in more accurately quantifying risk, which would help ensure that attention is focused on genuinely dangerous elements.

Examples of adversary simulation to identify risks from catastrophic misuse already exist. For example, an exercise organized by RAND involved 14 three-person cells roleplaying as malicious actors planning a biological attack using an LLM assistant (Mouton, Lucas, and Guest, 2023a). In other domains like cybersecurity, adversary simulation is used by red teams to actively model Advanced Persistent Threats (APTs), "well-resourced adversaries engaged in sophisticated malicious cyber activity that is targeted and aimed at prolonged network/system intrusion" (CISA, n.d.).

NIST should consider including the following guidance around adversary simulation as part of its overall guidelines on red-teaming:

- Adversary simulation should be used to assess severe misuse risks from AI systems, i.e., threats with national security implications. It is unclear whether these simulations are cost-effective for smaller-scale risks or for risks from unintentional harms.
- Actors organizing adversary simulations should consider establishing a "control group" to provide a baseline for the already extant level of risk in the real world, against which different systems (other "experimental groups") can be meaningfully compared. For example, in the RAND exercise previously mentioned, some cells were given access to an LLM assistant and the internet (the "experimental groups"), while others were only given internet access (the "control group") (Mouton, Lucas, and Guest, 2023b, p. 3).
- Actors organizing adversary simulations should consider employing mixed groups of domain experts for extended engagements, e.g., APT simulations in cybersecurity, which can occur over a time frame of weeks to months (Anderson, 2023). Given cost considerations, adversary simulation might be best reserved for threats that would plausibly result in catastrophic impacts.
- Actors organizing adversary simulations should document key information around the size and scope of red-teaming efforts along with results, including national security-relevant risks and mitigations, and potentially share some of this information with relevant stakeholders using a responsible disclosure process (Mulani and Whittlestone, 2023).

Further resources

Risk management frameworks and terminology

- Anderson-Samways & Acharya. Catching bugs: The Federal Select Agent Program and lessons for AI regulation, 2023
- Barrett, Newman, & Nonnecke. AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, 2023
- Ee, O'Brien, & Williams et al. Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach. 2023
- Khlaaf. Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems, 2023
- O'Brien, Ee & Williams. Deployment corrections: An incident response framework for frontier AI models, 2023
- Shevlane et al. Model evaluation for extreme risks, 2023

Third-party model evaluation

- Anderljung, Smith, & O'Brien et al. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework, 2023
- Bucknall & Trager. Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements, 2023
- Raji et al. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, 2023