NIST AI 800-1 2pd

Second Public Draft

Managing Misuse Risk for Dual-Use Foundation Models

U.S. AI Safety Institute

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.800-1.2pd

NIST AI 800-1 2pd

Second Public Draft

Managing Misuse Risk for Dual-Use Foundation Models

U.S. AI Safety Institute

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.800-1.2pd

January 2025

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Charles H. Romine, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

The U.S. AI Safety Institute (AISI) at NIST is releasing this document as an updated draft for the second round of public comment.

Comments on NIST AI 800-1 2pd may be sent electronically to NISTAI800-1@nist.gov with "NIST AI 800-1 2pd, Managing Misuse Risk for Dual-Use Foundation Models" in the subject line. Electronic submissions may be sent as an attachment in any of the following unlocked formats: HTML; ASCII; Word; RTF; or PDF.

All comments are subject to release under the Freedom of Information Act (FOIA)

Contents

Disclaimer: *Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.*

1. INTRODUCTION

This document provides voluntary guidelines for improving the safety, security, and trustworthiness of dual-use foundation models[i] (hereafter referred to as "foundation models")[ii] consistent with the National AI Initiative Act,[iii] Executive Order 14110,[iv] and the October 24, 2024, Presidential National Security Memorandum on AI.[v,1] Specifically, it focuses on managing the risk that such models will be deliberately misused to cause harm to public safety or national security. The ways that foundation models can be misused continue to evolve, but scenarios include using a model to facilitate the development of chemical, biological, radiological, or nuclear weapons; enable offensive cyber-attacks; and generate harmful or dangerous content, such as child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII) of real individuals.[vi]

The rapid development of foundation models poses significant challenges to understanding their capabilities and misuse risks,[2] and this document provides a basis to identify, measure, and mitigate these risks across the AI lifecycle. Misuse risks are not a function of the model alone—they result in part from malicious actors' motivations, resources, and constraints, as well as the ways that models are integrated into applications and society's defensive measures against harms.[3] As a result, the guidelines provided here address both technical and broader societal aspects of these risks.

This document identifies procedures and outlines a framework to anticipate, measure, and mitigate misuse risks from foundation models, as well as suggests how organizations can provide transparency into risk management practices. This document focuses particularly on foundation models' initial developers, but additional actors across the AI supply chain play a role in managing misuse risks, which Section 3 describes in greater detail.

---

[i] This term, and other terms throughout the document, are defined in Appendix A.

[ii] Executive Order 14110 defines a "dual-use foundation model" 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."

[iii] As codified in 14 U.S.C. § 278h-1.

[iv] Section 4.1(a)(ii) of Executive Order 14110 directs the Secretary of Commerce to "Establish appropriate 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. These efforts shall include: (A) coordinating or developing guidelines related to assessing and managing the safety, security, and trustworthiness of dual-use foundation models."

[v] Section 3.3(e)(ii) of the Memorandum tasks the AI Safety Institute to "issue guidance for AI developers on how to test, evaluate, and manage risks to safety, security, and trustworthiness arising from dual-use foundation models…," including on a range of subtopics in scope here.

[vi] Appendix D provides considerations for chemical and biological misuse risks; Appendix E for cyber misuse risks.

2. SCOPE

Risks. This document focuses on misuse risk from dual-use foundation models. Consistent with Section 3(k) of Executive Order 14110,[4] this includes foundation models that exhibit, or could be easily modified to exhibit, high levels of performance at tasks that can pose a serious risk to security, economic security, public health or safety, or any combination of those matters (hereafter referred to jointly as 'public safety'). This document addresses both potential future misuse harms, such as a foundation model facilitating the development of a biological or chemical weapon, as well as current harms from misuse, such as a foundation model generating CSAM or NCII. This document applies to the additional or novel misuse risk that a model introduces (i.e., the marginal risk).[5]

This document does not address all important risks from foundation models, nor does it address all risks to public safety that may arise from AI models that are not foundation models.[vii] Actors across the AI value chain and throughout the AI lifecycle should manage these additional risks as well, which may be consistent with relevant aspects of these guidelines, as well as those provided in other guidance such as the *Blueprint for an AI Bill of Rights*[6] and the NIST AI 100-1 AI Risk Management Framework (AI RMF) and NIST AI 600-1 RMF Generative AI Profile (AI RMF Generative AI Profile).[7]

Actors. The practices in this document principally focus on the role that the initial developers of foundation models play in the AI lifecycle. Implementation of the key practices in this document may vary depending on the size and resources of the developer. Initial developers have the most insight into model design characteristics, development process, and baseline capabilities. They can also take steps early in the development process to reduce a model's risks throughout its lifecycle, including across a range of downstream applications. Initial developers also have control over how the model is initially made available to others, but in many cases, they may share responsibility with downstream actors. Additional actors include cloud service providers, model hosting platforms, downstream model adapters, application developers, deployers, distribution platforms, third-party evaluators, and more.[8] Section 3 provides additional actors with examples of their roles in managing misuse risk.

[vii] For instance, many smaller, domain-specific models can also be misused in harmful ways. This document also does not cover risks from accidental AI harms to public safety.

3. MISUSE RISK MANAGEMENT THROUGHOUT THE AI SUPPLY CHAIN

While not the primary focus of this document, actors across the AI supply chain have a role in managing misuse risk in addition to developers. These actors include cloud service providers, model hosting platforms, downstream model adapters, application developers, deployers, distribution platforms, third-party evaluators and auditors, academic institutions, and government agencies.[9]

These actors should implement appropriate processes and risk mitigations to help manage misuse risk, such as the relevant guidelines outlined in Section 5 of this document and other NIST guidance, including the AI RMF and its Generative AI Profile. Actors should also collaborate to understand and mitigate misuse risks along the AI supply chain and throughout the AI lifecycle, such as the information sharing recommendations outlined in Objective 7 in Section 5. While best practices for managing misuse risk continue to evolve, this section outlines steps that these actors can take to manage and mitigate misuse risk specific to their roles in the AI supply chain, including:

1. **Compute providers:** Actors that provide compute infrastructure to foundation model developers and deployers, such as cloud service providers, can implement physical and cybersecurity practices to protect model weights and other sensitive model elements, as well as engage with developers to ensure such practices are proportionate to misuse risk. Objective 3 in Section 5 provides specific guidelines that may be relevant to cloud service providers.
2. **Model hosting platforms:** Actors that distribute and enable discovery of foundation models and datasets used to build foundation models can adopt terms of service that reduce misuse risks and implement systems for monitoring and reporting evidence of misuse to support enforcement. Objectives 6 and 7 in Section 5 provide related recommended practices.
3. **Downstream model adapters, application developers, and deployers**: AI systems may inherit misuse risks from upstream foundation models. Design decisions made by downstream actors when adapting models and integrating them into applications, such as when implementing safeguards, as well as selection of deployment context and use case, can impact misuse risk. To manage deployment-specific misuse risks, these actors can use information shared from the initial developer as a starting point for managing misuse risk and review all of Section 5, if necessary, with particular attention to Objectives 5 and 6.
4. **Distribution platforms:** App stores and other platforms that enable the discovery, distribution, and use of AI applications can define terms of service that reduce misuse risks, apply system-level safeguards to models that are directly available for inference, and implement systems for monitoring and reporting evidence of misuse to support enforcement. Third-party platforms that do not directly host foundation models or AI applications may host model outputs and thus be a potential source of misuse. These

platforms can also share information across the AI supply chain to help manage misuse risk, as outlined in Objectives 6 and 7 in Section 5.

5. **Third-party evaluators and auditors:** Third parties can provide information and knowledge related to public safety risks of foundation models to complement developer-led misuse risk management. Fostering a community of third-party evaluators and auditors helps maintain accountability for model developers and increases institutional transparency around misuse risk management practices. Third-party evaluators and auditors can help define standards for misuse risk management practices and formalize processes for reliable audits and evaluations of foundation models.[10] They can also help improve misuse risk management practices overall, such as by developing methods to incorporate empirical evidence of real-world harm into misuse risk assessments. To perform evaluations, third parties can review the documentation provisions throughout Section 5, with particular attention to Objectives 1, 4, 5, and 6.
6. **Academics and external researchers:** Academics and independent external researchers can help advance scientific research related to AI safety and security practices, including the development of novel model safeguards, methods and tools to measure misuse risk, and innovative risk management practices. Academics and external researchers can also consider how their research might contribute to implementation of the practices in Section 5, with particular attention to Objectives 1, 4, and 5.
7. **Government agencies:** Government agencies can coordinate national security efforts to understand sensitive misuse risks in classified settings and enforce federal, civil, and criminal law to combat misuse. Government agencies can complement other actors in disseminating information about misuse risk, work with partners across the AI supply chain to develop practices for assessing misuse risk, and foster a robust ecosystem of third-party evaluator and auditors. Government agencies can also take steps to improve societal capacity and resilience to increase robustness to the possible misuse of foundation models.
8. **Users and the public:** The public can report observed misuse incidents to law enforcement, incident databases, and actors across the AI supply chain. Users can report information, such as model issues that may lead to misuse risk (hereafter referred to as 'model flaws'), to actors across the AI supply chain.

4. KEY CHALLENGES IN MANAGING MISUSE RISKS

While this document provides guidelines for managing misuse risks of foundation models, it is important to acknowledge that these approaches are still nascent, posing both methodological and scientific challenges to implementation. For instance, best practices for evaluating model capabilities, identifying and accurately measuring potential misuse risks, and monitoring real-world misuse remain in flux. Organizations should collaborate and take appropriate steps to build an empirical basis to evaluate and mitigate misuse risks. Such challenges include:

1. **Foundation models are general purpose.** Models trained on a broad data distribution can often be applied across many different domains and integrated into a variety of downstream systems, including domains and systems not expressly considered by the developer. This flexibility renders anticipating the potential ways in which a model might be misused difficult and complicates the measurement and monitoring of misuse risk.
2. **Models are only one element of misuse risk**. Misuse risk is not simply a function of model capabilities identified in specific evaluations. Public safety harms often arise from the successful completion of a chain of tasks, where only a subset benefit from model capabilities. While models can provide new methods to enact harm, enhance actors' abilities to commit existing harms, or increase the number of actors that can participate in malicious activity, harms to public safety also often require physical infrastructure, distribution mechanisms, or complex interactions in the physical world.[11] Bad actors may also already have access to existing tools that serve their needs better than foundation models, and existing methods to prevent harm—such as controls on physical dual-use materials—may be substantially more determinative of real-world risks.[12]
3. **Model capabilities that relate to misuse risk are difficult to predict.** It is difficult to predict the model capabilities that may increase the risk of misuse. In many instances, increasing the amount or quality of a foundation model's training data, the quantity of compute used to build or run the model, or the number of parameters in the model can improve its performance.[13] However, these factors, while useful heuristics in some instances, have limited precision and an uncertain relationship to a model's capabilities and potential misuse risks.[14]
4. **It is difficult to accurately emulate threat actors and simulate misuse scenarios**. Assessing misuse risk requires understanding how a malicious actor might in practice exploit a foundation model's capabilities or circumvent safeguards to cause harm. Organizations may not have the capacity to accurately simulate the resources, time, or access to infrastructure or niche expertise available to a malicious actor. Realistic emulation of malicious actors may also be prohibitively dangerous or unethical, such as attempting to develop a volatile substance, generate non-consensual intimate imagery, or harm real people.
5. **Results from misuse risk measurement may not generalize to real-world situations.** Model performance on one task may not provide reliable evidence of its performance on others, even when the two tasks appear related.[15] For instance, initial benchmarks

that are cheaper and easier to carry out may suggest a model has dangerous capabilities, but this concern may not be substantiated when the model is tested in more rigorous and realistic conditions.

6. **Evaluating misuse risk may require scarce domain expertise**. Information about some risks, like the potential for a foundation model to enable a malicious actor to develop a chemical or biological weapon, may be closely guarded. Organizations developing foundation models may not have the domain expertise – or access to the expertise – necessary to assess or manage some misuse risks.

7. **Safeguards to protect against misuse are often brittle, and methods to evaluate their efficacy are nascent**. Foundation model developers can implement safeguards to protect their models from misuse, such as those included in Appendix B, but few techniques exist to evaluate the adequacy of those safeguards under real-world conditions. Safeguards are often underdeveloped and brittle: for example, foundation models trained to decline harmful requests will often comply with those requests if an adversary reframes them or adds additional text to the prompt.[16]

5. OBJECTIVES AND PRACTICES TO MANAGE MISUSE RISKS

This section outlines seven objectives that organizations should aim to meet to assess and manage the risk that their foundation models will be misused to deliberately harm public safety.[viii] This document also provides practices and recommendations for each objective as non-exhaustive examples of how organizations can meet the objectives in a manner they deem appropriate and proportionate to the risks posed by their foundation models.

These objectives are:

1. **Identify** potential misuse risk
2. **Plan** to manage misuse risk
3. **Protect** the model from unauthorized access (if necessary)
4. **Measure** misuse risk associated with model deployments
5. **Mitigate** misuse risk before deployments
6. **Monitor** and respond to misuse
7. **Disclose** misuse risk management practices

Managing misuse risk is an iterative process, and organizations should consider all seven objectives holistically to manage foundation model misuse risks. Organizations should also adopt additional risk management measures and communicate with other actors in the AI supply chain, where and when appropriate. However, implementing these practices cannot guarantee that actors will not misuse a foundation model, and model developers should remain aware of possible misuse risks. For instance, organizations can consider these objectives in conjunction with existing NIST guidance, particularly the NIST AI RMF and its Generative AI Profile, and NIST SP 800-30 Guide for Conducting Risk Assessments. See Figure 1 for primary correspondence of objectives across the AI lifecycle.

[viii] In some cases, the objectives may be difficult to fully achieve based on the state of the science. In such cases, and depending on model capabilities, the practices and recommendations may serve primarily to help further advance towards the objective rather than ensuring it would be fully achieved.

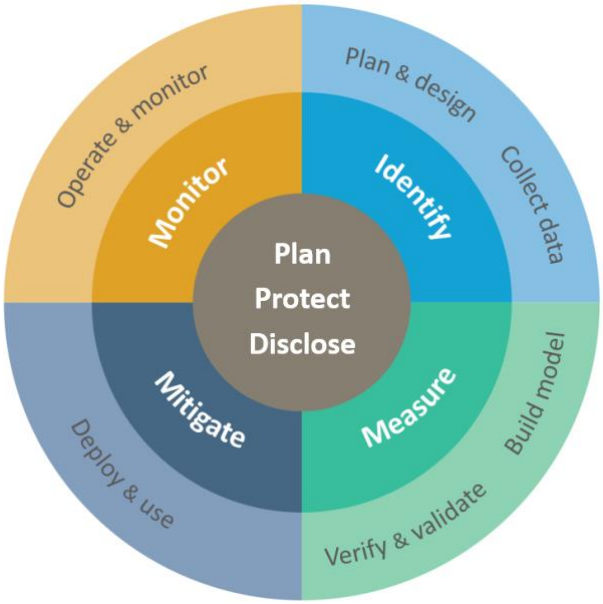

Figure 1. *NIST AI 800-1 Objectives in the AI Lifecycle. The two inner circles display the objectives included in Section 5 and the outer circle indicates their primary corresponding stages in the AI lifecycle.[17] Plan, Protect, and Disclose (Objectives 2, 3, and 7, respectively) may be applicable across all AI lifecycle stages. Note that these objectives should be implemented holistically and iterated upon as necessary across the AI lifecycle to manage misuse risk particular to a specific foundation model.*

The specific elements of each objective below include:

1. **Objective:** a target outcome that will help organizations manage misuse risks.
2. **Practice:** a suggested practice that can help organizations achieve an objective. This document proposes practices that may not be appropriate for all contexts. Alternative practices may also achieve these objectives, and new practices are likely to be developed over time.
3. **Recommendation:** an implementation characteristic or consideration that is often important for a practice to effectively achieve an objective. These recommendations may not be appropriate in all cases and in some cases will not be exhaustive of the measures necessary for a practice to help achieve an objective.
4. **Documentation:** examples of information that could demonstrate whether and how a practice was implemented, as well as potential evidence of its effectiveness and comprehensiveness for internal record-keeping or external information sharing. Sharing documentation with the public and/or select parties can help demonstrate implementation, facilitate collaboration across the AI supply chain, and develop shared best practices for managing misuse risks. If an organization relies on alternative practices to achieve an objective, a similar level of documentation and a rationale for why the alternative practices match or exceed the efficacy of provided practices can be used. In situations where documentation details are sensitive, an organization may need

to calibrate the amount and type of information shared externally, such as only sharing with appropriate national security authorities.

Objective 1: Identify potential misuse risk

Anticipate the possible ways in which malicious actors might use the model to cause harm (hereafter referred to as 'threat profiles') and assess the potential impact of this harm before the model is developed.

Practice 1.1: Anticipate model capabilities.

Forecast potential model capabilities that may present misuse risk by surveying evidence from deployed models with similar capabilities to the model (hereafter referred to as 'proxy models').

Recommendations for Practice 1.1:

1. Prioritize evidence from widely deployed and studied proxy models.
2. Consider also collecting evidence from less similar models, such as those that are significantly more or less powerful than the model (e.g., if a less capable model presents a misuse risk, prioritize measuring for that risk).
3. Consider other factors beyond model characteristics, such as resources available to develop the model, when relating the proxy models' capabilities to anticipated ones.
4. Collect information about the prevalence of misuse, the usefulness of proxy models for potentially harmful or dual-use real-world tasks, and the efficacy of safeguards.
5. Assess the degree of uncertainty in these estimates based on the difference between the proxy model and the planned model, how those differences are expected to affect their capabilities, and the reliability and completeness of the evaluations available for proxy models.
6. Consider involving experts in anticipating model capabilities, such as by potentially granting them access to intermediate model checkpoints for open-ended experimentation to identify other ways the model could be misused to cause harm.
7. If capability forecasts are uncertain and include the possibility that the model may require increased risk mitigations, consider increasing the frequency of capability measurements during the development process (Objective 4) and expanding the organization's planned risk mitigation measures (Objective 5).

Documentation for Practice 1.1:

a. The proxy model(s) used for comparison and the specific capabilities that relate to misuse risk.
b. The anticipated difference between the model's forecasted capabilities and the capabilities of the proxy model(s).

Practice 1.2 Create threat profiles.

Before training a model, identify and maintain a list of threat profiles based on the anticipated model capabilities.

Recommendations for Practice 1.2:

1. Identify (i) the threat actor or actors that may misuse the model, (ii) the chain of tasks required to realize harm, (iii) the relevance of model capabilities to each task, and (iv) the harm(s) that the model could help the threat actor(s) enact.
2. Incorporate relevant information about possible misuse identified from proxy models in Practice 1.1. Examples of possible misuse risks may include the risk that a model may lower the barriers to the design, development, distribution, or use of chemical, biological, radiological, or nuclear (CBRN) weapons or the automation of an offensive cyberattack.
3. Use real data, case studies, or expert opinions to inform threat profiles and help identify gaps.[ix]
4. Develop a plan for identifying and adding threat profiles to this list as future research reveals new potential or ongoing misuse risks (Objective 2).

Documentation for Practice 1.2:

a. A list of identified threat profiles.
b. A description of how these threat profiles were selected and judged to have adequate coverage.
c. A plan for updating this list as new misuse risks may materialize.

Practice 1.3: Conduct risk assessments.

For each threat profile, assess the potential misuse risk by estimating the likelihood of harm occurring and the impact if that harm occurred.

Recommendations for Practice 1.3:

1. Consider both quantitative and qualitative assessments of likelihood and impact, if possible.
2. Account for alternative tools already available to threat actors, such as existing models or other digital tools, and assess the model's marginal risk of misuse relative to that baseline.[18]

[ix] Consider that there may be multiple threat profiles for a given capability; for example, automation of vulnerability discovery might be used in different ways by a criminal organization or by a nation state, corresponding to distinct threat profiles.

3. Consider the new risk the model may introduce, such as the model's ability to help an actor increase the scale, prevalence, or frequency, decrease the cost, or improve the effectiveness or efficiency of their malicious activity.
4. Consider malicious actors' potential motivations, willingness, and capacity to enact harm, as well as the number of malicious actors that may exist.[19]
5. Account for existing mitigations and barriers, such as limitations on access to physical resources needed to enact harm, and the level of societal preparedness to defend against that harm.
6. Use information about the real-world impact and use of proxy models to inform the assessment; update these assessments as new information about real-world impact materializes.
7. Update estimates if changes in the broader ecosystem either increase or decrease vulnerability to potential harms.
8. Recognize areas of uncertainty and sensitivity and account for these areas when communicating and using impact assessments.

Documentation for 1.3:

a. A risk assessment for each identified threat profile.
b. A comparison between past risk assessment predictions and actual model impact for proxy models, when and if relevant and accessible.

Objective 2: Plan to manage misuse risk

Determine a strategy to apply appropriate risk mitigations for each potential model capability that may introduce misuse risk. Align development and deployment plans with the resources, time, and operational constraints that may be required to manage potential misuse risk.

Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk.

For foundation models with capabilities that may introduce misuse risks, identify the risk mitigations necessary to manage the risks that might arise from these particular capabilities across the AI lifecycle. Appendix B includes a list of example risk mitigations.

Recommendations for Practice 2.1:

1. Consider the costs and benefits of planned risk mitigations, using both qualitative and quantitative comparisons.
2. Consider the risk in the context of organizational risk tolerances, which may reflect legal and regulatory obligations.
3. Identify and articulate evidence to justify the adequacy of planned risk mitigations.

4. Refine planned methods to mitigate misuse risk periodically based on adjustments to identified threat profiles, changes in factors that affect risk tolerance (such as increases in expected benefits), and information about real-world performance.
5. As measurement and real-world monitoring is performed, continue to re-assess the safeguards necessary to manage misuse risk for each particular capability.

Documentation for Practice 2.1:

a. The procedure and criteria used to define risk mitigations proportionate to model capabilities.
b. An articulation of the adequacy of these mitigations to manage misuse risk.

Practice 2.2 Establish an organizational plan to manage misuse risk.

Develop and implement an organizational plan, including the expected resources needed and timelines to manage misuse risk throughout the AI lifecycle, including iterations and improvements to the model.

Recommendations for Practice 2.2:

1. Plan for information security and physical security practices necessary to manage misuse risk, when appropriate.
2. Plan research needed to implement and evaluate adequate safeguards. If necessary, base these plans on the mapping established in Practice 2.1.
3. Plan for decision-making about model deployment. This plan should include the organization's plan for pre-deployment testing and determining which appropriate risk mitigations are available.
4. Plan to monitor evidence of real-world misuse and potential risk. Develop processes for adjusting these organizational plans, as well as deployment or development approaches, as necessary.
5. Consider consulting existing resources for developing and implementing such organizational plans, such as the NIST RMF and other guidance.[20]

Documentation for Practice 2.2:

a. The planned risk mitigations, accounting for technical and resource constraints.
b. Concrete steps to take if subsequent measurements suggest mitigations are not adequate to manage misuse risk.
c. Plans to update (a) and (b) as necessary.

Objective 3: Protect the model from unauthorized access (if necessary)

Some models may pose particular misuse risk, warranting careful storage. In such cases, take steps to assess and prevent the risk that entities and individuals not authorized by the developer will access the model, its associated weights, or other information that could facilitate re-creation of the model in an unauthorized manner (hereafter referred to as 'unauthorized access').[x]

Practice 3.1: Assess misuse risk from threat actors gaining unauthorized access to the model.

Based on threat profiles and risk assessments identified in Objective 1, assess whether unauthorized access to a model may pose or significantly increase misuse risk, and if so, estimate that risk.

Recommendations for Practice 3.1:

1. Consider possible threat actors' motivations and level of sophistication related to gaining unauthorized access to the model.[21]
2. Consider the organization's compliance with applicable cybersecurity best practices, such as NIST Special Publication 800-53 and Secure Software Development Framework (SSDF), to help assess the probability of unauthorized access.[22]
3. Consider the threat posed by insiders, such as an individual involved in developing or deploying the model who may behave maliciously or collaborate with an external attacker.
4. When relevant, consider using cybersecurity red teams and penetration testing to assess how difficult it would be for an actor to circumvent security measures.

Documentation for Practice 3.1:

a. A summary of evidence collected to evaluate the risk of threat actors gaining unauthorized access to the model, including the results of any cybersecurity red team exercises and penetration testing.

Practice 3.2: Maintain security practices sufficient to prevent unauthorized access.

[x] In this document, unauthorized access generally refers to an intrusion or breach of an information system or asset that the accessing entity has not been granted any access to (a full definition is provided in Appendix A). This is differentiated from unauthorized *use,* which here refers to an entity misusing technical access that they have been granted legitimately to carry out some unauthorized purpose, such as by jailbreaking a model or violating its terms of service.

Implement security practices, to the extent possible, commensurate with risk to protect against threat actors gaining unauthorized access to the model.

Recommendations for Practice 3.2:

1. Consider adopting existing U.S. government cybersecurity guidance and applicable international or industry standards.[23]
2. Apply security practices tailored to the context of foundation models, such as protections against exfiltrating large amounts of data (e.g., model weights) and other vulnerabilities (e.g., extraction attacks).[24]
3. Apply appropriate protections against insider threats, such as limiting access to model weights within the organization or implementing two-party control systems.
4. Re-assess the risk of unauthorized access as security practices are implemented.
5. Only develop a model that relies on confidentiality to manage misuse risk when the risk of threat actors gaining unauthorized access to the model is sufficiently mitigated. When the organization is made aware of an increased risk of unauthorized access, adjust or halt further development until the risk of unauthorized access is adequately managed.

Documentation for Practice 3.2:

a. A summary of security measures that have been implemented to reduce the risk of threat actors gaining unauthorized access to the model. Consider including references to standards used to implement such security measures.

Objective 4: Measure misuse risk associated with model deployments

Measure model capabilities related to identified misuse to help quantify risk. Rely on methods that incorporate both technical and societal factors and prioritize accuracy while avoiding harm from measuring dangerous activities.

Practice 4.1: Evaluate model capabilities on tasks relevant to assessing misuse risk.

Conduct pre-deployment evaluations of the model to measure performance on specific capabilities that may introduce misuse risk identified in Practice 1.3. Appendix C lists possible characteristics of methods to measure misuse risk.

Recommendations for Practice 4.1:

1. Evaluate the model's capabilities throughout the development process: during training, after training, and when integrated into a downstream system or interface.
2. Consider using automated or otherwise less expensive tests on tasks related to real-world harm to indicate that a model lacks dangerous capabilities. If initial tests are

inconclusive or indicate a potential risk, conduct more costly and precise measurements.

3. Consult relevant experts when creating evaluations to ensure that they are indicative of real-world scenarios.
4. Maximize model performance on evaluation tasks by adjusting prompts, software scaffolding, fine-tuning the model, or other means.
5. If a gap exists between the effort or resources applied by evaluators to maximize system performance and the resources that could be applied by a threat actor, account for that gap when interpreting evaluation results.
6. Avoid overlap between data used to train a model and data used in capability evaluations and measure the extent of any overlap.
7. Account for the relative immaturity of some measurement approaches and consider misuse risk in the context of this immaturity. Update evaluations accordingly as the state of science advances.
8. Compare measurements to real-world observations of proxy models identified in Practice 1.1 to help calibrate the relationship between real-world observations and controlled model evaluations. If sufficient proxy models have been deployed but information about real-world performance is not available, consider conducting additional relevant measurements of performance on harmful or dual-use tasks in a realistic setting.

Documentation for Practice 4.1:

a. A list of tasks that were used to evaluate each threat profile, results of these tests, and a representative subset of datasets used for each evaluation.
b. A methodological description for each evaluation in enough detail to reproduce it. For example, outline methods for data collection, preprocessing, prompts, inference parameters, and evaluator models.
c. An analysis of the relationship between evaluation tasks, assessed capabilities, and potential misuse risks identified in Objective 1 that addresses uncertainties and limitations.

Practice 4.2: Red-team safeguards.

Use red-teaming to assess whether threat actors could bypass model or system safeguards designed to prevent misuse.

Recommendations for Practice 4.2:

1. Evaluate whether an adequately resourced red team can misuse the model to achieve a predetermined goal or accomplish a related proxy task in a realistic deployment context. If using proxy tasks, ensure that they are at least as easy to achieve as undesirable real-

world outcomes, considering both the inherent complexity of the task and any additional difficulty introduced by safeguards.

2. Assemble red teams based on their ability to succeed at the determined task, considering factors such as relevant domain expertise, their independence from the model developer, diversity of perspectives, and lack of incentives that conflict with their red-teaming goal.
3. Provide the red team with a clearly defined task. Determine a performance metric for measuring the red team's success and provide incentives and accountability for task completion.
4. Compare the red team's expertise, resources, and time available to those of a relevant threat actor. Ensure that the red team is adequately resourced to the extent possible and consider providing alternative and/or additional resources to the red team or making the red team's task easier in other ways (such as providing additional access to the model or dividing complex tasks into constituent pieces) to compensate for any remaining gaps between the red team and threat actors.[25]
5. Provide the red team with at least as much information about the model (or downstream system if testing at the integration level) as would be available to an attacker and make explicit any respects in which the red team lacks full information about the design of safeguards.
6. Attempt to minimize institutional or legal obstacles to red teams succeeding at their tasks. Consider providing appropriate protections, such as waiving terms of service and legal liability.
7. Consider each level of access to a model that a threat actor might have, ranging from limited access through an API to direct access to the code and parameters that define the model, and determine the minimum level of access (if any) that allows the red team to accomplish its goal.
8. Consider testing the model at different levels, such as the model without safeguards, the model with safeguards, and the model integrated into a downstream system or interface.
9. Consider information that may become public about how to circumvent safeguards when assessing risk, including when information will be disclosed to the developer relative to the public and the technical feasibility of rapidly addressing disclosed model flaws.

Documentation for Practice 4.2:

a. For each red team exercise, a description of the risks it is intended to address, the goal provided to the red team, the composition and expertise of the red team, and the resources, level of access, and time available to the red team.

b. A high-level summary of the results of red-teaming, including the level of success achieved by the red team at their task, feedback from the red team about their interpretation of the outcome, and any other relevant red-teaming characteristics.

Objective 5: Mitigate misuse risk before deployments

Make deployment decisions based on available safeguards and their efficacy to adequately mitigate misuse risks, based on plans determined in Objective 2. Appendix B outlines a list of example methods to mitigate misuse risk.

Practice 5.1: Implement safeguards proportionate to the model's misuse risk.

Implement safeguards designed to prevent the model from being misused, as identified in Practice 2.1.

Recommendations for Practice 5.1:

1. Establish evidence of safeguards' effectiveness before relying on them to prevent misuse risks, such as by red-teaming (Practice 4.2) and monitoring the efficacy of safeguards for proxy models (Practice 1.1).
2. Consider implementing safeguards at various stages of model development, such as before, during, and after training and once the model has been integrated into a downstream system, as well as other mitigation measures. Some examples are included in Appendix B.

Documentation for Practice 5.1:

a. The list of safeguards that have been implemented.

b. A summary of any evidence available about the safeguards' efficacy, which may include the result of safeguard evaluations via red teams and other testing.

Practice 5.2: Assess misuse risk based on implemented safeguards.

Evaluate misuse risk in the context of implemented safeguards to assess the remaining 'residual risk.' Use this assessment to inform the choice of deployment strategy.

Recommendations for Practice 5.2:

1. Identify planned deployments that could impact misuse risk. For example, consider whether the planned deployment could impact the number of actors who have access to a model or the level of access that they would have.
2. Estimate misuse risk based on appropriate measurements conducted under Objective 4 and comparisons to proxy models conducted in Practice 1.1.
3. If additional safeguards are added in response to identified risks, consider re-assessing misuse risk prior to deployment by carrying out red-teaming exercises (Practice 4.2) with the additional safeguards in place.

Documentation for Practice 5.2:

a. The level of model access provided for each deployment.

b. The assessed misuse risks associated with each deployment.

Practice 5.3: Adopt appropriate deployment strategies based on misuse risk assessments.

If risk cannot be adequately managed after completing Practice 5.1 and 5.2, determine whether the deployment should be modified or delayed.

Recommendations for Practice 5.3:

1. Assess residual risk by incorporating the overall assessed misuse risk from the selected deployment strategy and mitigations from implemented safeguards.
2. Consider deployment strategies that provide additional real-world evidence to inform risk assessments without presenting significant misuse risks.
3. Consider whether additional safeguards are feasible to implement prior to deployment, whether additional time could be used to carry out a more reliable estimate of risk, or whether a more limited deployment may be more appropriate given the level of assessed risk.
4. Consider leaving a buffer between the estimated level of risk—given the implemented safeguards and the deployment strategy—and the associated anticipated real-world risk (hereafter referred to as a 'margin of safety'). This margin of safety could incorporate how threat actors may continue to acquire new knowledge about how to misuse or augment the model after it is deployed[26] and how to circumvent safeguards.[27] Consider a larger margin of safety to manage risks that are more severe or less certain.

Documentation for Practice 5.3:

a. The basis for determining that the risk of misuse was adequately managed for a deployment decision, including that the deployment's risks were within the organization's risk tolerances.

Objective 6: Monitor and respond to misuse

Collect information about model usage to improve understanding of misuse risk, adapt to emerging risks, and improve future deployments. Engage with and encourage findings from the public, civil society organizations, external researchers, and the foundation model's distribution partners.

Practice 6.1: Monitor for evidence of misuse.

Accounting for differences in deployment strategies; develop and implement a plan for monitoring model deployments for misuse.

Recommendations for Practice 6.1:

1. Monitor APIs, model hosting platforms, and other distribution channels for misuse while maintaining privacy of users.
2. Build or procure systems to enable automated detection of misuse.
3. Periodically assess the effectiveness of misuse detection systems, including through red-teaming.
4. Request that distribution channels monitor for misuse and share information regarding this monitoring.
5. Consider tiered methods of detection, such as scanning for misuse using less costly automated methods and then validating through direct human intervention, to help prioritize limited resources, improve privacy, and increase coverage.
6. Collaborate with other actors, such as content distribution platforms, downstream model adapters, cloud service providers, third-party researchers, and law enforcement officials, to track real-world incidents of misuse.

Documentation for Practice 6.1:

a. A summary of the mechanisms used to monitor each distribution channel for a foundation model and the methods for determining the effectiveness of those mechanisms.

Practice 6.2: Respond to incidents of model misuse.

Develop incident response processes to manage real-world misuse after model deployments.

Recommendations for Practice 6.2:

1. Establish clear organizational responsibilities for accountable incident response processes.
2. Plan for responses to plausible novel scenarios of misuse, such as strengthening safeguards or implementing new ones.[28]
3. Plan for how identified instances of misuse will inform future development and deployment decisions, especially when reducing access to the model may not be possible.
4. Consider carrying out drills to practice responding to time-sensitive and safety-critical scenarios of misuse, if appropriate given the level of risk of the deployment.

Documentation for Practice 6.2:

a. A summary of the incident response process and the organizational roles and responsibilities in the process.

Practice 6.3: Establish misuse reporting mechanisms.

Establish channels and protections for reporting known instances of misuse and issues that could lead to misuse risk.

Recommendations for Practice 6.3:

1. Adopt policies that protect and reward individuals who report model issues related to misuse risk.
2. Establish formal processes to adjudicate concerns from the public, employees, and contractors in a timely fashion.
3. Communicate the identified issues or misuse instances clearly with employees and contractors, as appropriate.

Documentation for Practice 6.3:

a. A summary of the organization's policies with respect to internal and external reporting of misuse and issues that could lead to misuse.
b. For internal use, a summary of possible limitations of evaluations or safeguards relevant to risk assessments, possible deficiencies within the organization's risk management processes, and any detected errors in disclosures related to misuse risk.

Practice 6.4: Provide safe harbors for third-party safety research.

Create protections for third-party researchers to encourage research on reducing misuse risk.

Recommendations for Practice 6.4:

1. Publish a clear vulnerability disclosure policy for model flaws that outlines how such flaws should be shared with the developer and the public, and how the organization will respond to reported flaws.[29]
2. Publish a clear safe harbor policy that commits to not pursuing legal action against or restricting the accounts of external safety researchers that act in good faith and comply with the vulnerability disclosure policy.[30]
3. Consider providing support and accommodations for vetted external researchers' interactions with the model, such as providing researchers with access to models with fewer safeguards to conduct post-deployment red-teaming exercises.

Documentation for Practice 6.4:

a. A vulnerability disclosure policy.
b. A safe harbor policy.
c. A summary of the organization's commitment to not pursue legal action against third-party researchers acting in good faith.
d. A description of the organization's process for providing vetted researchers with access to models with fewer safeguards.[31]

Practice 6.5: Create bounties for issues related to misuse risk.

Establish a bounty program to incentivize researchers to identify model flaws.

Recommendations for Practice 6.5:

1. Establish a program to incentivize researchers to find model flaws and disclose them according to the vulnerability disclosure policy established in Practice 6.4.[32]
2. Consider referring to norms and best practices of existing bug bounty programs, such as those instituted by software vendors, to guide program development.

Documentation for Practice 6.5:

a. The terms of the bounties and details regarding the process for submitting model flaws.

Objective 7: Disclose misuse risk management practices

Provide transparency to the public and relevant entities about the organization's processes for addressing misuse risk to facilitate understanding, accountability, collaboration, and scientific advancement.

Practice 7.1: Publish transparency reports.

Regularly publish transparency reports that include key details regarding misuse risks, how those risks are assessed and managed, and roadmaps for safety and security improvements.

Recommendations for Practice 7.1:

1. Share the methodology and results of pre- and post-deployment evaluations of model capabilities, risks, and mitigations, including as much detail about the data and evaluation methodology as can be disclosed without introducing risks to public safety.[33]
2. Share details regarding the safeguards in place for the model, including how they are applied across different distribution channels, with as much detail as can be shared without rendering the safeguards ineffective.[34]
3. Share information about data used to build the model that is relevant to assessing the misuse risk, such as criteria for data filtering, criteria for data selection, and data sources, when possible.[35]
4. Share steps that downstream developers and deployers of AI systems that integrate the foundation model should take to manage misuse risk.
5. Share relevant details about the internal organizational processes used to create risk assessments and to make deployment and development decisions.
6. Make the transparency reports publicly available, update them on a regular basis (e.g., with each new major version of the model), and include key information related to misuse risk.[36]

7. Consider aligning transparency reports with existing public AI safety commitments, such as those made to the White House Voluntary Commitments, the HAIP G7 Code of Conduct, and the Seoul Frontier AI Safety Commitments.[37]

Documentation for Practice 7.1:

a. A public transparency report.

Practice 7.2: Disclose information about risk management practices.

Share detailed information on risk management practices with other developers, downstream deployers, and other entities across the AI supply chain, as appropriate,[xi] to advance the science of AI safety.

Recommendations for Practice 7.2:

1. Share information covering the practices used to achieve the objectives listed in this document, including at least as much detail as described in the documentation sections for each practice.
2. Share information about threat profiles with other entities across the AI supply chain and the public to develop a joint knowledge base and save resources.

Documentation for Practice 7.2:

a. A summary of the list of stakeholders that receive information about risk management practices and the types of information disclosed to them.

Practice 7.3: Report misuse incidents.

Based on existing best practices and an adequate consideration of the benefits and risks of disclosing certain information, disclose incidents of misuse.

Recommendations for Practice 7.3:

1. Define the category of misuse events to report.[38]
2. Collate verified reports of misuse in a commonly used and standardized format.[39]
3. Share verified reports of misuse with relevant third parties, such as AI incident databases and other model developers.[40]

Documentation for Practice 7.3:

a. A description of the incident reporting process, including an example of a type of incident that might be reported and to whom it would be reported.

[xi] There may be some situations where sharing details of threat profiles widely is not appropriate, such as sharing classified or highly sensitive information that is not widely known or disseminated. Consider the trade-offs between information sharing risks due to the exposure of sensitive data and the benefits of transparency when determining the recipient(s) and level of detail of documentation.

Appendix A. Glossary

Artificial Intelligence (AI)
A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.[41]

AI Red-Teaming
A structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. AI red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.[42]

Distribution Channel
The various ways in which a model could be distributed, including, but not limited to, public release of the model weights, access to the model via an API supplied by a cloud service provider, access to the full model on open-source repositories, or access to a fine-tuned or otherwise augmented version of the model from a third-party deployer.

Dual-Use Foundation Model or Foundation Model (for the purposes of these guidelines only)
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: substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear weapons; enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber-attacks; or permitting the evasion of human control or oversight through means of deception or obfuscation. Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.[43]

Fine-Tuning
An approach in which the parameters of an already trained model are adjusted by training on new data. Fine-tuning is often used to adapt a model to a particular task, or to mildly modify a model's behavior.

Margin of Safety
The buffer that a developer may choose to leave between the level of risk associated with a model measured pre-deployment and the actual level of risk presented by its particular deployment to account for uncertainty in capability evaluations and risk estimations. This margin could be, for instance, accounting for the additional time, resources, and knowledge that malicious actors may have when assessing the amount of risk measured in a red-teaming exercise.

Misuse Risk
A risk that an AI model will be deliberately misused to cause harm. Pursuant to the NIST AI Risk Management Framework (AI RMF),[44] risk is defined as the composite measure of the probability that a harm occurs and the magnitude (or degree) of the corresponding harm.

Model Flaws
Issues in a model that could lead to misuse risk, such as security vulnerabilities and other model safety issues, including weaknesses in model safeguards.

Model Performance
A model's capability and adaptability to a new task or unseen data, which may, but does not necessarily include, a model's accuracy.

Unauthorized Access
Obtaining without permission information that substantially aids an actor in recreating an AI model or its capabilities that may introduce misuse risk, such as the code or weights that define the model, or bypassing security measures designed to prevent access to the model itself such that an actor can utilize AI model capabilities in an unauthorized manner. Depending on the stage in the AI lifecycle and how the model is deployed, categories of unauthorized access may have differing degrees of relevancy.

Proxy Models
Deployed models that may have similar capabilities to the model and thus be most pertinent to estimating the risks of misuse.

Threat Profile
A threat profile consists of: (a) the malicious task(s) that the threat actor might accomplish using the model, (b) the threat actor or actors who might use the model to cause this type of harm, (c) the way(s) in which they could use the model to accomplish this task, and (d) the mechanism(s) by which this could cause harm.

Appendix B. Example Methods to Mitigate Misuse Risk

The available mitigations against the misuse of foundation models continue to evolve and expand. Current literature suggests that mitigations' effectiveness can vary widely, and reliance on them to prevent misuse should be based on empirical evidence of the mitigations' efficacy, particularly as some may incur tradeoffs in areas like cost, transparency, researcher access, and user privacy. The following table provides a non-exhaustive survey of several categories that organizations can consider:

Table 1. Example Risk Mitigations.

Risk Mitigations	*Possible Implementation Methods*
Improve the model's training.	• Filter training data to exclude examples that could result in capabilities that increase the likelihood of misuse, such as biological sequence data, or known CSAM/NCII images.[45] • Apply refusal training and similar training techniques to enhance a model's ability to refuse harmful requests.[46] • Train models with approaches that make it more difficult for subsequent fine-tuning to remove safeguards.[47] • Consider employing machine unlearning approaches, such as exact unlearning, approximate unlearning, zero-shot learning (ZSL) and fast and efficient unlearning, to reduce the model's knowledge of harmful information.[48]
Detect and block attempted misuse.	• Add additional infrastructure around the base model to monitor for and detect behavior which may constitute misuse—such as with algorithmic classifiers for misuse. This infrastructure should be supported, as appropriate, by human review.[49] • Once detected, block, modify, or otherwise limit unsafe queries and responses, and place limitations on users and organizations engaging in misuse or attempting to circumvent safeguards. • For available weights, maintain a feedback mechanism to learn of misuse incidents and work with model hosting platforms to address distributions that are intended to increase misuse.
Reduce or limit access to the model's capabilities.	• Limit access to the model, or to versions or features of the model that display particularly high-risk capabilities, to users with lower risk of misuse and contexts that are easier to monitor.[50] • Reduce access to the model reactively when misuse is detected to limit further misuse, such as by rolling back a model to a previous version or discontinuing availability of a model, if a model displays sufficiently extreme misuse risk while it is in production. • Consider limiting internal access to the models' weights within an organization. • Consider when it is appropriate to allow the model to be fine-tuned via API, which also can reduce the availability of safeguards, such as by letting users train the model on data related to dangerous tasks or reduce how often the model refuses dangerous requests.[51] • Iteratively deploy systems so as to observe real-world capabilities and monitor for misuse from weaker systems. • For models that pose sufficiently severe risk, consider restricting and monitoring access even by individuals directly involved in its development and evaluation. Internal protections can reduce the risk that an organization's own employees or contractors are able to misuse models; these protections can include logging

	employee interactions with models, restricting model access to a subset of employees, requiring multiple employees to access a model together, and ensuring that all employee interactions with the model are accompanied by the model's other relevant safeguards.
Release the model in stages.	• Consider deploying a model via an API to understand its impacts before making its weights available, such as by making them available for download by the public. • Consider the fact that once a model's weights are made widely available, options to roll back or prevent its further sharing and modification are severely limited.[52] • Consider deploying a model with a limited audience and gradually expanding access to the model to wider audiences (e.g. work with cybersecurity professionals to fix bugs or carry out large scale testing before a model is widely available).[53]
Collaborate across the supply chain to implement real-world protections.	• Collaborate with cybersecurity professionals to fix issues such as infrastructure vulnerabilities. • Collaborate with the scientific community to improve practices such as nucleic acid synthesis screening, a biosecurity practice that aims to control access to dual-use biological materials, rendering it more difficult to carry out a biological attack.[54] • Collaborate with other actors to improve filtering and protect against AI-enabled fraud.
Stop developing a model if it presents significant misuse risk.	• If other practically available safeguards are not sufficient to protect a model from misuse–including considering the risk of unauthorized access or internal abuse–then it may be appropriate to make significant changes to the development plan, including, if necessary, delaying the development of the model.

Appendix C. Key Characteristics of Measurement Tasks

This table provides a non-exhaustive overview of approaches to measure misuse risk to use in accordance with the recommendations provided in Practice 4.1. This table is based on the current state of the field, and organizations should recognize that these characteristics will likely change as the science of risk measurement evolves.

Table 2. Measurement Task Characteristics.

Measurement Environment *Measurement tasks can evaluate model capabilities in various environments, some of which are outlined below.*		
Environment type	**Description**	**Examples**
1. ***Question & answer***	The model is given static prompts (whether text, images, or other modes), and its response is scored.	- The model is run on MMLU, GPQA, and other widely used, public benchmarks.[55]
2. ***Question & answer with tools***	The model is given static prompts, but it can interact with a computational environment containing relevant tools or workspaces to draw conclusions and inform a response. Note that the inclusion of tools may provide a better indication of system performance under realistic conditions.	- When answering mathematical reasoning questions, the model is provided with a scratchpad for Chain-of-Thought reasoning and the ability to write and run Python code. - When answering biological questions, the model is provided with relevant domain-specific tools that a human would use in their work, such as bioinformatics tools.
3. ***Computer environment***	The model is given a task to complete and access to versions of standard OS utilities such as shells, interpreters, and a text editor, to interact with computer systems or other computers over a network. These evaluations can be automatic while still closely matching some realistic deployments.	- A capture the flag challenge in which the model runs arbitrary code on a local machine to attempt to compromise a remote server. - A machine learning challenge that provides the model with a training data set and the system runs within a docker container to submit a trained model that is scored using a hidden test set.
4. ***Human interaction***	The model exchanges messages with humans, which can be accompanied also with interactions with computer environments and tools.	- The model negotiates with a human participant as part of a controlled experiment, and the participant grades model performance on a set scale. - A biosecurity expert asks a series of questions to the model to simulate an interaction with a malicious actor. The expert grades the model responses using a provided rubric.

5. ***Physical environment***	The model interacts with physical systems. Evaluations usually involve physical experiments in which the model controls physical actuators (or gives advice to humans to take physical actions) based on observations of the environment.	- The model advises a human on how to carry out a procedure in a biology lab. A human carries out the instructions and the results are assessed. - The model controls a drone to evade obstacles and reach a target.
6. ***Simulated environment***	When conducting physical experiments is expensive, it may be possible to collect approximate results by simulating a realistic environment and measuring model capabilities in that simulation.	- The model advises a human on carrying out lab protocols in a virtual reality simulation of a biology lab. - The model interacts with a computer environment in which the role of human users is simulated by another model.
Grading Procedure *Measurement tasks can evaluate model capabilities using various grading procedures, both qualitative and quantitative, some of which are outlined below.*		
Grader	**Description**	**Examples**
1. ***Known ground truth***	Reponses to questions are graded by a simple and unambiguous rule for determining which responses are correct (or how to award partial credit). Note that this method only works for specific questions (see non-exhaustive examples).	- Multiple-choice questions can be graded by comparing an answer to the ground truth. - Numerical questions can be graded by comparing numerical values.
2. ***Interpreting results based on a rubric***	A subject matter expert reviews model output to judge whether it is correct or whether analogous responses would be useful to a malicious actor in a realistic misuse context. These reviews may also be automated by an AI system equipped with a task rubric, although automated grading can introduce further error.	- The model is tasked with producing detailed instructions for carrying out a real-world task, and an expert judges whether those instructions are precise and accurate enough to be useful for a malicious actor. - The model carries out multiple steps of a simulated offensive cyber operation, and an expert reviews its behavior to assess how well the model might perform in a more realistic setting.
3. ***Environmental objective***	The model is graded in an interactive environment based on whether it has a certain effect on the environment.	- In a software engineering evaluation, the model can be scored based on whether it writes software that passes unit tests and how many resources it consumes while doing so. - In a laboratory uplift experiment, the model that is helping novices troubleshoot a lab protocol can be scored based on physical

		measurements of whether those protocols were successful.
Model Involvement in Measured Task *Measurement tasks can evaluate misuse risks that involve various levels of model autonomy, such as the model automating an end-to-end task or the model providing a human marginal 'uplift' on an existing task, some of which are outlined below.*		
Evaluated system	**Description**	**Examples**
1. ***Fully autonomous system***	An AI system, which incorporates a foundation model, solves an evaluation task fully autonomously. Although most misuse tasks involve an instruction and/or other aid from a malicious actor to the AI system they are using as a tool, for some misuse scenarios there may still be evaluated subtasks that are fully automated.	- A capture the flag challenge in which the model autonomously identifies vulnerabilities and develops an exploit. - A question-answering task in which the model uses tools and web research to write instructions for a complex task that are graded by an expert.
2. ***Human + AI system activity ("uplift study")***	A human uses an AI system, which incorporates a foundation model, as an assistant to accomplish an evaluation task. For many tasks there is no simple decomposition of a task into subtasks that are performed by an AI system and subtasks performed by humans. In these settings, evaluating the performance of a human-AI team is the only way to obtain a directly meaningful indicator of performance. Such evaluations may consider humans with limited expertise being able to achieve previously unobtainable outcomes by using an AI system.	- A capture the flag challenge in which a human uses an AI tool to help them more quickly identify vulnerabilities and develop an exploit to extract a flag. - A question-answering task in which human novices use an AI assistant to write instructions for a complex task that are graded by an expert. - A laboratory uplift study in which human participants use an AI system as an assistant to propose and troubleshoot protocols to achieve a real-world goal.
Conclusions Drawn from Measurement *Different evaluation tasks provide varying degrees of information for which conclusions about the model can be drawn. For instance, some measurement methods, such as monitoring the impact of the model in the wild, can inform organizations' understanding of a model's absolute performance, whereas other assessments should be used only to assess performance relative to another model or a human baseline*		
Conclusion Drawn	**Description**	**Example**
1. ***Absolute performance on a task that is representative of real-world risk***	Some measurements are sufficiently representative of real-world misuse scenarios. For these tasks, it may be possible to draw a direct correspondence between the measured level of performance on the task and the model's potential real-world impact.	- The model is used to identify vulnerabilities and create and execute exploits for real-world software. These results may provide direct evidence of the model's usefulness in automating an offensive cyber task. - Novices use the model to troubleshoot lab protocols

		similar to those needed to develop a bioweapon. These results may provide direct evidence of the tool's usefulness in a malicious biological workflow. - The model's ability to identify vulnerabilities and create exploits for real-world software is related directly to a model's ability to enable that misuse risk in the wild.
2. ***Performance on a task relative to another model or a human baseline***	Some measurements may be less directly comparable to a real-world misuse scenario. Instead, they involve capabilities that are similar enough to provide some useful signal, particularly about the relative performance of different models (or the comparison between model and human performance). Such measurements may provide an indicator of whether the model may significantly alter misuse potential beyond existing baselines. Combined with other evidence about the misuse potential of a proxy model or human baseline, this measurement can give evidence about the misuse potential of the model.	- Comparing the performance of two models on cyber capture the flag challenges to get an indicator of whether one of them is likely to be significantly more useful for vulnerability exploitation than the other.

Relationship to Misuse Scenario

The relationship between the measurement task and the actual misuse task may vary, which can impact how useful the measurement is for assessing real-world misuse risk and whether the measurement is better suited to absolute or relative risk assessment.

Relationship to misuse task	Description	Examples
1. ***Close match for misuse task***	In some cases, it is possible to safely evaluate tasks from essentially the same distribution that would be used by a malicious actor, or to monitor real-world misuse.	- An authorized attempt to compromise a real IT system. - Monitoring the impact of AI-generated synthetic content in the wild.
2. ***Safe proxy for misuse task***	In some cases, it is possible to modify a task in a way that reduces the risk of carrying out the measurement without changing the capabilities that would be required to succeed at the dangerous variant of the task.	- Measuring a model's ability to help a human synthesize a safe biological agent with similar properties to a dangerous bioweapon.
3. ***Subtask of misuse task***	Sometimes it is easier or safer to evaluate individual steps or tasks in a more complex misuse task or chain of tasks, and weak	- Having novices use the model to carry out particular steps of dual-use biological protocols

	performance on an individual subtask can provide a strong indication that a model will not perform well at a more complex end-to-end task.	that are particularly challenging or inaccessible. - Asking the model to identify and fix a simple flaw in a piece of malware.
4. ***Requires similar capabilities to misuse task***	Tasks that experts consider to involve similar skills to a misuse task can also be used to measure misuse risk. The quality of inference depends on the degree of similarity. When tasks are only distantly related, then the evidence can only be used in settings where large error is acceptable (such as establishing a conservative bound on capabilities).	- Evaluating the model's ability to generate explicit videos of adults to provide partial evidence of their ability to do so for minors. - Studying the benign usage of models in real-world biological laboratory settings.
5. ***Simpler than misuse task***	In many cases, experts may have some confidence that one task is significantly easier than another loosely related task, even if they are uncertain about the exact relationship. This can provide evidence that an AI system does not pose a significant misuse risk even if there are important differences between evaluation and misuse tasks.	- Evaluating the model's general scientific capabilities using a question-answer dataset designed to be much easier than autonomously facilitating novel scientific discoveries. - Evaluating the model's ability to identify security-critical flaws in short snippets of code designed to be much easier than finding vulnerabilities in realistic software.

1

Appendix D. Application of NIST AI 800-1 to Chemical and Biological Misuse Risk

This appendix[xii] outlines additional considerations relevant to identifying, measuring, and mitigating risks associated with the misuse of foundation models for the development or use of dangerous biological or chemical agents. This appendix's goal is to complement the broader set of objectives and practices outlined in the main body of the guidance by assisting developers, deployers, and other actors across the AI supply chain in understanding domain-specific considerations that can assist in implementing misuse risk management practices. The recommendations in this appendix are primarily relevant for foundation model developers and deployers, as well as third-party evaluators that work with developers and deployers to help assess chemical and biological misuse risks. Other actors' roles are discussed in Section 3 of the main document.

The assessment of AI misuse risks, particularly for foundation models, is a rapidly evolving field requiring expertise across multiple domains including cybersecurity, biosecurity, and national security. The guidelines in this appendix represent initial considerations to assist organizations in approaching these complex challenges. These documents are being released for public comment and further review. U.S. Government (USG) entities, along with state, local, tribal, and territorial government actors and law enforcement agencies, possess unique insights, expertise, and access to sensitive information regarding threat actors, historical incidents, and emerging scenarios that are essential for comprehensive risk assessment in the domain of chemical and biological misuse. This expertise can complement private organizations' technical understanding of foundation models' capabilities, and future work may explore how public and private expertise and risk assessments can work together.

D.1 Identifying Chemical and Biological Misuse Risk

This section expands on Objective 1 ("Identify potential misuse risk") for identifying chemical and biological (CB) misuse risks from foundation models. The potentially high-consequence and cascading effects of CB agents, particularly novel agents and/or toxins and transmissible biological agents, necessitate careful identification of potential risk.[56] As outlined in Practice 1.3, organizations should assess both the likelihood of successful misuse and the impact of possible consequences. Risk assessment in this domain represents an interplay between organizations and government entities, with each bringing distinct and complementary capabilities and insights to the process. While organizations such as developers and deployers can assess a foundation model's technical capabilities, USG entities can provide critical context about threats, actors, and scenarios that enable comprehensive risk evaluation.

CB risk assessment involves analyzing the intersection of two rapidly evolving fields: AI and biotechnology. Risks may emerge not just from improvements in foundation models, but from how these models interact with new biotechnology tools and techniques.

In assessing CB misuse risks, organizations such as developers, deployers, and third-party evaluators can focus on evaluating a foundation model's technical capabilities and how these capabilities might affect

[xii] Note: While the concepts in this appendix align with the main text of 800-1, some section numbers and cross-references may not match exactly between the documents.

existing barriers to misuse, while USG entities can provide organizations with appropriate threat profiles for evaluation purposes. For emerging risks without extensive historical precedent, organizations can work with USG entities to understand how capability changes could affect different threat scenarios and identify potential defensive measures.

D.1.1 Establishing Threat Profiles: Actors and Scenarios

When establishing and prioritizing CB misuse scenarios for assessing foundation model capabilities, organizations typically benefit from working with subject matter experts (SMEs). SMEs can help analyze how a model might enhance different actors' abilities to accomplish measurable CB tasks that could enable harm. Such an analysis might focus in particular on how a foundation model could reduce capability barriers for less sophisticated actors. To aid this work, organizations can build on existing frameworks for assessing life sciences dual-use research of concern[57] to incorporate AI-specific considerations,[58] such as by documenting when and how specific measurable chemical and biological misuse risks may arise (see Table D.4). Where they already exist, organizations may use established frameworks for identifying AI-specific CB misuse risks to complement the guidance in this document.

As outlined in Practice 1.2, when establishing CB misuse scenarios, organizations can identify representative threat profiles to help categorize and prioritize the risk space, focusing on those most relevant to a specific foundation model's capabilities. While the distribution of such a list requires careful consideration of potentially sensitive information, it helps ensure no major categories of risk are overlooked and allows for more transparent justification of which scenarios to prioritize for deeper analysis.

The following subsections outline relevant factors and example methods that organizations can consider when analyzing potential misuse scenarios and threat profiles, including actor categories, scenario types, and key barriers that foundation models might aid an actor in overcoming. Detailed threat profiles are primarily necessary when a foundation model may advance the frontier of CB capabilities available to malicious actors. In other cases, and for other capabilities, organizations can typically rely on simpler risk assessment approaches, such as comparisons to proxy models as outlined in Practice 1.1, or existing frameworks.

Threat Actor Categories. To enable better scoped and more effective risk assessment and mitigation strategies, organizations can distinguish between state actors (who often have access to multiple expert scientists, extensive resources, and specialized facilities, and may pursue strategic national objectives) and non-state actors (who typically face various resource and expertise constraints and may have different motivations). To facilitate more consistent assessment across organizations, the following illustrative non-state actor profiles could be considered:

- Organized violent non-state group with multiple members, some funding, and potential access to relevant facilities.
- Individual with advanced scientific expertise (e.g., PhD or equivalent experience in a relevant biological subfield).
- Lone actor or small group with malicious intent but limited expertise and resources.

For each actor category, capabilities and sophistication can be assessed across multiple dimensions, such as:

- Scientific expertise and laboratory experience (including quality and quantity of research experience, tacit knowledge, and technical support networks).
- Available time and operational constraints (including both self-imposed and external timeline pressures).
- Ability to acquire necessary materials (including access to facilities, suppliers, and financial resources).
- Ability to access and use foundation models (including elicitation capabilities such as tooling development, potential for model weight exfiltration, fine-tuning on private datasets, and circumvention of safeguards).

Organizations may also consider assessing both deliberate misuse by malicious actors and unintentional harm from researchers or organizations lacking sufficient safety expertise or oversight.[59]

Threat Scenarios. To better understand the range and severity of potential harms, organizations can consider threat scenarios across different agent types and intended outcomes. High-level examples include, but are not limited to:

- Acquisition and/or release of a known transmissible biological agent that could seed an epidemic or pandemic.
- Acquisition and/or release of a known chemical or non-transmissible biological agent over a high-density or large population area or in a transportation hub, urban event, or other congregate setting.
- Design, acquisition, and/or release of an enhanced or novel chemical or biological agent.

Organizations can use the above high-level threat scenarios as a starting point for more specific and comprehensive CB threat scenarios. These could be similar to the list of example biological and chemical threats in the Homeland Security Planning Scenarios.[60]

Barriers and Model Capabilities. To identify specific ways that foundation models might increase risk, organizations can assess how foundation models might help actors overcome existing barriers, such as:

- **Technical Barriers:** Technical barriers relate to scientific challenges in the end-to-end synthesis, acquisition, and delivery of chemical or biological agents. Key examples include tacit knowledge requirements for agent-specific synthesis and experimental manipulation, genome design and assembly challenges, interdisciplinary expertise gaps, and laboratory troubleshooting capabilities. Foundation models may reduce these barriers through detailed guidance and assistance, including in multiple modalities.
- **Operational Barriers:** Operational barriers involve logistical challenges such as facility requirements, equipment and material limitations, financial and time constraints, headcount requirements, and industry screening of customers or orders. While models cannot directly address physical infrastructure needs, they may help actors identify minimum necessary requirements, aid in execution of tasks within a laboratory, suggest alternative approaches to achieving the same capabilities with different methods or materials when preferred options are unavailable, or explore ways to circumvent existing biosafety[61] and biosecurity[62] controls.
- **Motivational Barriers:** Motivational barriers include the initial ideation of using CB agents as weapons for specific goals, perceived difficulty of developing and using CB agents as weapons, as well as psychological and normative barriers such as fear of attribution or international norms. While some actors may be more willing to overcome these barriers based on their

beliefs or objectives, model assistance could make complex processes appear more manageable and reduce the need for collaborators. Organizations may consider how their models' responses to CB-related queries might affect these barriers.

D.1.2 Risk Assessment

After identifying key threat profiles, organizations can identify which scenarios warrant detailed assessment. The process of assessing CB misuse risks requires close coordination between organizations and USG entities. Organizations' technical understanding of foundation models complements USG expertise regarding threats and scenarios. To ensure efficient use of assessment resources, organizations may first identify a set of representative scenarios spanning different scales of potential harm (e.g., varying levels of health, economic, or national security impacts), then evaluate these scenarios using the following three factors:

- **Severity of Potential Harm**: The magnitude and scope of potential harm if an actor successfully executes the threat scenario. This may be particularly important for scenarios involving transmissible biological agents that could cause cascading effects beyond initial release or agents for which there are no reliable medical countermeasures. Impacts can span across multiple categories, including but not limited to: human health effects (i.e., illnesses and deaths); impacts to animal, plant, and environmental health; harm to U.S. national security or economic security; and broader societal disruption. Organizations can consider how different combinations of reduced barriers might affect potential harm.
- **Counterfactual Effect**: The marginal increase in risk that foundation models might create beyond baseline capabilities, which can vary across actors. Some threat actors may already have a high likelihood of success through certain pathways (such as synthesis of known pathogens in professional laboratory settings or acquisition of industrial chemicals through legitimate suppliers), making the additional impact of access to a foundation model's capabilities less significant for these scenarios.
- **Likelihood of model enabling actors to overcome barriers**: The likelihood that foundation models will develop capabilities that enable actors to overcome specific barriers within the assessment timeframe. This varies significantly by barrier type – for instance, models may more readily help overcome technical barriers like laboratory protocol development than operational barriers like acquiring specialized equipment. Organizations can incorporate both technical feasibility and anticipated model development trajectories when assessing this likelihood.

For scenarios identified as high priority through the above analysis, organizations can assess how model capabilities might affect technical success rates and reduce existing barriers across different types of actors. When working with relevant USG entities, additional context may be available to inform these assessments, including:

- Threat actor profiles and characteristics.
- Technical and operational factors that influence success rates.
- How capabilities might affect different categories of actors.

This collaborative approach enables organizations to contribute their deep understanding of model capabilities while drawing on USG expertise regarding threat scenarios, actor profiles, and assessments of potential threat actors and likelihood of misuse attempts. Together, this allows for more comprehensive evaluation of how a foundation model's capabilities might affect risk within specific scenarios.

These assessments can consider:

- Baseline risk.
- Marginal risk from model access (how a model might expand actor pools, improve success rates, or increase potential impact through enhanced capabilities or novel techniques).
- The impact of different combinations of overcome barriers on risk (e.g., overcoming technical but not operational barriers may limit potential impact).
- Existing defensive measures and controls (like biosurveillance systems and nucleic acid synthesis screening) that are publicly documented and how foundation models might enhance such measures and controls.[63]

It is useful for organizations to draw on historical CB incidents and SME expertise to inform these assessments. The group of SMEs is often most valuable when it has demonstrated expertise in both the relevant scientific and technical domains (virology, synthetic biology, machine learning, etc.) as well security (biosecurity, national security, etc.). This can include using structured expert elicitation methods, such as the Delphi method, and noting areas of expert agreement and disagreement. While precise quantification may be challenging, organizations can document:

- Key assumptions and evidence about foundation model capabilities.
- How capabilities vary across deployment contexts.
- Relative weighting of different risk factors and treatment of uncertainties.

These assessments can be updated by organizations with access to relevant SME expertise as new evidence emerges about model capabilities through observation of legitimate use (e.g., model assistance for beneficial research) and malicious misuse attempts.

Information Sharing Considerations. After conducting CB risk assessments, organizations may carefully weigh transparency benefits against publicizing potentially sensitive information, making sharing decisions on a case-by-case basis. Organizations may consult with USG entities regarding potentially sensitive findings. Examples of information that may be useful to share publicly include:

- The general framework used for creating threat scenarios or profiles.
- High-level categorization of threat actors.
- The dimensions considered in capability assessment (scientific expertise, material acquisition, model access).
- The key factors used in scenario prioritization (counterfactual effect, likelihood, severity)
- General approach to risk quantification methodology.
- Descriptions of how expert panels are sourced and structured.

Some examples of information that may require careful consideration before sharing and may be more appropriate to share with a limited group include:

- Detailed threat scenarios or profiles.
- Quantitative or qualitative risk estimates for specific scenarios.
- Novel risk pathways identified during assessment.
- Detailed information about routes to CB acquisition.
- Detailed information about routes to harm, including specific steps in the CB attack chain (e.g., synthesis, weaponization, delivery).
- Information about specific pathogens.

D.1.3 Key Opportunities and Challenges for Risk Assessment

Opportunities:

- Laboratory processes' structured nature and clear technical requirements enable systematic analysis of the impact of assistance from foundation models
- Existing life sciences dual-use research frameworks provide valuable precedent for evaluating dangerous capabilities of foundation models
- Risk assessment methodologies can effectively break down complex scenarios into analyzable components

Challenges:

- The complex interplay between actor sophistication levels and their ability to overcome various barriers makes mapping model performance on a task to actual risk difficult
- Limited historical data on CB attacks complicates baseline risk estimation
- Balancing scientific transparency with security considerations may require careful consideration given the benefits of transparency and the potential risks of disclosing novel information regarding CB risk pathways

D.2 Evaluating Chemical and Biological Capabilities Relevant to Misuse Risk

Building directly on Section 5, Objective 4 ("Measure misuse risk associated with model deployments") and the measurement characteristics in Appendix C, this section provides detailed guidelines for evaluating CB misuse risks. Organizations can refer to Appendix C's framework while implementing the CB-specific guidelines provided here.

D.2.1 CB Evaluation Types and Characteristics

When designing foundation model evaluations for CB misuse risk, organizations can account for the rapid pace of advancement in both AI and biotechnology – particularly how emerging tools and techniques might create new risk pathways. The following evaluation types provide a framework that can be adapted as capabilities in both domains continue to advance.

Building on Appendix C, organizations can employ multiple complementary evaluation approaches to build a more comprehensive understanding of potential misuse tasks while maintaining documentation of methodologies and results. This section focuses on evaluating foundation models on CB assistant tasks, especially in laboratory contexts. Other capabilities, such as those of specialized chemical and biological AI models[64] like protein design models, are also important potential sources of risk, but evaluating the capabilities of such models typically involve other measurement approaches.

Common evaluation types for CB misuse risk include the following:

- **Automated Benchmarks**: These are standardized question-answer sets that enable rapid, large-scale testing of model capabilities. For broad coverage of relevant CB misuse risks and measurable tasks, automated benchmarks typically cover a broad range of CB topics with dual-use applications – usually comprising hundreds if not thousands of questions. Questions can be

multiple-choice or open-ended, with open-ended questions better resembling real-world human-AI interactions.

- **Assistant Task Evaluations**: These evaluate a model's ability to provide actionable guidance when asked to accomplish specific tasks. The tasks can range from short-form focused queries (e.g., "How would you modify this DNA plasmid sequence in X way?") to long-form complex multi-step processes (e.g., "Provide a complete protocol for the reverse genetics synthesis of virus Y"). Task-based evaluations are intended to approximate actual human-AI interactions better than automated benchmarks.
- **Expert Model Assessment**: This evaluation approach involves qualitative exploration of model capabilities by SMEs (e.g., virologists, molecular biologists, biodefense experts) to help identify potential areas of concern. Expert assessment typically takes two forms, which can be combined: (i) SMEs evaluating model responses to assistant tasks for technical accuracy and real-world feasibility, and (ii) SMEs engaging in exploratory dialogue with models to simulate how malicious actors might attempt to elicit harmful information. Organizations can consider using multiple expert teams tailored to each representative CB pathway being assessed.
- **Uplift Studies**: These evaluations assess how foundation models enhance human capabilities by comparing the performance of humans with and without model assistance. While assistant task evaluations measure model outputs directly, uplift studies measure success at specific overarching tasks by measuring how effectively humans can use models to accomplish these tasks. They focus on the combined human-AI results rather than the model results alone. While expert model assessment can help identify potential risks, uplift studies generally provide more reliable evidence about real-world capabilities since they directly measure what humans can accomplish with model assistance. They include three main types: task operational uplift studies (focusing on computational or web-based tasks), end-to-end operational uplift studies (assessing performance across the entire ideation-to-release spectrum without physical implementation), and full laboratory uplift studies (involving wet lab work). End-to-end operational studies, including several recent studies,[6566] evaluate how dual-use foundation models' impact performance across different stages, while laboratory studies provide the most comprehensive but resource-intensive assessment. Laboratory studies should be carefully designed to understand safety risks while adhering to all applicable oversight frameworks, policies, treaties, and regulations.

Relationship to Misuse Tasks: When selecting evaluation approaches, organizations can consider how closely they correspond to dual-use scenarios that could lead to realistic CB misuse scenarios. The relationship between evaluation tasks and misuse tasks in the CB domain typically falls into one of five measurement categories, as outlined in Appendix C: close match tasks, safe proxy tasks, critical subtasks, tasks requiring similar capabilities, or tasks simpler than misuse tasks. In applying this framework to CB evaluations:

- **Close-match tasks and safe proxy tasks** are the most direct measurement but are often difficult to measure given safety and security considerations as well as laboratory requirements. Safe proxy tasks, such as synthesizing benign viruses with similar technical characteristics to dangerous ones, can provide comparable insights while reducing risk. Organizations should ensure any such evaluations comply with all applicable regulations and oversight requirements.
- **Critical subtasks,** and **tasks requiring similar capabilities,** are particularly relevant in CB evaluations as a second-best alternative to close-match and safe proxy tasks.

- **Tasks strictly simpler than misuse tasks** often have limited utility in certain risk scenarios as state-of-the-art foundation models often match the performance of human experts on these tasks.[676869] There are exceptions to this, especially with respect to CB agent design and *in silico* testing.

Close-match measurement through laboratory work typically provides the most reliable evidence of model capabilities. However, laboratory studies present significant challenges - including cost, the need for specialized facilities and expertise, and safety and security considerations. Moreover, for tasks relating to synthesis of dangerous CB agents, organizations typically find it valuable to pursue safe proxy tasks rather than close-match tasks. When direct laboratory evaluation is not feasible, organizations can employ two key alternative measurement approaches:

1. **Comparable Task Assessment** (corresponding to safe proxy tasks, critical subtasks, and tasks requiring similar capabilities in Appendix C): CB evaluations usually involve safe comparable tasks or subtasks that maintain technical similarity while reducing evaluation-related risk. Organizations can validate the relationship between comparable or similar task and actual task performance through scientific and national security expert assessments,[70] document assumptions about how comparable task performance relates to real-world capabilities, and consider multiple comparable tasks to triangulate capability assessment. For example, when assessing capabilities related to viral agent synthesis, organizations might evaluate performance on synthesizing harmless viral proxies that maintain technical similarity while reducing biosafety and biosecurity risks of performing viral agent synthesis in a laboratory setting.
2. **Proxy Model Comparison** (corresponding to the relative risk measurement approach in Appendix C): Organizations can compare their model's performance against existing widely used models whose risks and capabilities are better understood. For example, if model A has been tested in laboratory uplift studies and found to provide sufficient uplift for viral synthesis, a developer might conclude that model B presents a manageable risk by finding that it has significantly lower performance than model A on question-answering tasks related to laboratory troubleshooting. This allows organizations to leverage existing knowledge about the relationship between different types of tasks while minimizing safety and security risks in their own testing. This approach is especially valuable for organizations with limited resources as it necessitates less extensive testing infrastructure.

D.2.2 Considerations for CB Evaluation Methodology

CB evaluation methodologies can incorporate several key elements across evaluation types to help ensure comprehensive and reliable assessment. To promote safety across the AI supply chain, these can be shared as part of transparency reports or system cards. Additional considerations for the following categories include:

Design of Automated Benchmarks and Assistant Task Evaluations:

- *Considerations*:
  - For multiple-choice questions, include carefully crafted options with plausible distractors to prevent models from using choice elimination strategies,[71] and validating format effectiveness through small-scale comparisons with open-ended versions. Open-ended questions better approximate real-world interactions, focusing on information integration, error identification, and novel knowledge application rather than factual recall alone.

    - Task-based evaluations typically use clear rubrics assessing technical accuracy, completeness, and real-world feasibility, often employing tiered scoring to account for partially correct responses. Where possible, assistant task evaluations can be validated against uplift studies to confirm that expert assessments correlate with real-world success rates.
- *Transparency*: Organizations can specify how they developed the questions, how they ensured quality control, the extent to which the questions or tasks reflect real-world workflows, comparisons of results against previous evaluation efforts when available, and other relevant factors.

Design of Expert Model Assessment:

- *Considerations*:
    - In expert model assessment, SMEs with relevant CB expertise (such as in virology, synthetic biology, and biosecurity) can assess the model's responses for accuracy and utility, providing a subjective assessment of how useful the model's response would be for an actor attempting to misuse the agent.
    - To maintain consistency and reduce subjective variation, organizations can establish structured evaluation formats and maintain consistent assessment criteria across evaluations while documenting methodology to enable comparison across models. Emerging automated expert assessment systems may complement human evaluation, though these approaches are still being developed for domain-specific assessment.
- *Transparency*: Organizations can specify the backgrounds of the experts, the rubric used for assessment of the utility and accuracy of model responses, and other relevant factors.

Design of Uplift Studies:

- *Considerations*:
    - Uplift studies generally require sample sizes sufficient to rule out concerning effect sizes and estimate improvements with clear error bounds. Studies can assess multiple metrics relative to control groups, as task completion alone may miss important changes in capability. Key metrics could include changes in protocol success rates, changes in protocol completion speed, and changes in participant confidence and motivation levels.
    - Time frames reflect evaluation goals – if comparing to expert performance, allow sufficient time for reasonable task completion; if comparing to malicious use scenarios, match expected threat actor time constraints. If time must be reduced (for example, condensing what would normally be a weeks-long project into hours), either decrease task scope or increase available time proportionally.
    - Participant expertise levels are meant to match threat actor profiles, with appropriate expert control groups for capability comparisons. Studies can provide time to allow participants to become familiar with prompt engineering and the model being studied, use initial skill assessment through preliminary tests rather than self-reporting (e.g. first-day skill calibration), and prevent information sharing between groups. Analysis may particularly focus on high-performing outliers, as these may represent significant risks even when mean improvements appear modest.
    - Organizations can consider different study designs to assess various aspects of capability enhancement. This includes: (i) open-ended studies where participants must achieve an

end goal without being given a protocol, and (ii) protocol-based studies where participants must identify and troubleshoot intentional errors or obstacles. Studies can also evaluate different team configurations, from individual actors to small groups (e.g., teams of 2-3 participants), as capabilities may vary significantly with team size and composition.
    - Laboratory uplift studies should be run for sufficient duration to accommodate multiple attempts per task, as participants typically show significant improvement after initial failures. These studies can incorporate both quantitative metrics and qualitative observations of participant behavior and problem-solving approaches, with particular attention to how participants interact with laboratory equipment and troubleshoot experimental issues.
- *Transparency*: Organizations can specify sample size calculations and power analyses, time allocation decisions, participant selection criteria, measures taken to prevent information sharing between groups, and other relevant factors.

Human Expert Baselines:

- *Considerations*:
    - Establishing robust human expert baselines is crucial for contextualizing model performance and assessing whether models are approaching or enabling non-expert users to approach expert-level performance.
    - Organizations can ensure high-quality expert participation by providing appropriate incentives and access to all relevant tools and web-based resources for a level of performance during evaluations that is reflective of human expert capabilities. Larger expert group baselines can provide more reliable results.
    - Expert grading should generally be blinded and assess multiple dimensions in addition to accuracy, such as completeness, time-to-completion, innovation, and detection avoidance. When assessing the ability of a foundation model to uplift the capabilities of human novices, organizations may consider establishing a novice baseline and a direct novice-using-foundation-model comparison.
- *Transparency*: Organizations can specify the backgrounds of the experts, how much time they were given per question or task, what prompting guidance they were given when they conducted such studies, and other relevant factors.

Tool Integration and Multimodality:

- *Considerations*:
    - Organizations can incorporate multiple modalities where relevant to CB tasks. For example, image analysis capabilities may be useful for laboratory troubleshooting, allowing models to interpret experimental results and identify equipment setup issues. While currently underexplored, voice interaction capabilities may also become relevant for hands-free model assistance during laboratory procedures.
    - Foundation models can be tested with tools human experts typically use, including literature search capabilities, bioinformatics software, and laboratory equipment interfaces. For example, joint U.S. and U.K. AI Safety Institute pilot studies demonstrated improved performance on DNA and protein sequence tasks when models had access to Python sandboxes and bioinformatics packages.[7273]

- Organizations can consider how different augmentations may enhance model capabilities, including:
  - Retrieval-augmented generation systems that enable access to domain-specific knowledge and reference materials
  - Additional inference time and compute resources that enable longer reasoning chains or additional attempts to respond to a request
  - Tool use and API access that allow models to interact with external systems
- Examples of specific tools or scaffolds which may be relevant include, but are not limited, to: literature search and retrieval-augmented-generation capabilities, bioinformatics software via APIs or graphical interfaces, sandboxed versions of commercial services like DNA synthesis providers, specialized biological or chemical AI models for safe proxy tasks, and laboratory equipment interfaces where appropriate.
- Once a system is deployed, users may discover additional ways to improve its performance or may use tools not contemplated during pre-deployment evaluations. This gap is best considered when interpreting evaluation results, and organizations may consider making forecasts about how performance may improve over time as available tools and agent scaffolds improve.[74]

- *Transparency*: Organizations can specify which tools or multimodal interfaces were made available to models and any limitations or constraints on tool or interface usage.

Real-world grounding:

- *Considerations:*
  - Organizations should ground their evaluations in empirical evidence when possible. For example, if conducting yearly laboratory uplift studies, organizations might design automated benchmarks that specifically test capabilities found to be critical in those studies. Similar approaches can be applied to web-based tasks like sequence ordering.
  - Organizations can consider clearly articulating why their chosen evaluation methods are relevant to assessing real-world misuse risk. This could include explaining how each evaluation relates to real-world CB tasks, and specifying which of the five measurement categories it falls into and why their results are informative about misuse risk. They may also justify why the level evaluation difficulty is appropriate (i.e., neither too easy nor unrealistically hard).
  - Organizations can include human participants where relevant, especially for longer, complex tasks where human-AI interaction could significantly impact performance.[75]
- *Transparency*: Organizations can specify how evaluation tasks were derived from real-world scenarios, what assumptions were made about task difficulty and complexity, and how human-AI interaction was incorporated into evaluation design.

Interpreting multiple evaluation results:

- *Considerations*:
  - Different evaluation approaches – even different configurations of the same evaluation – can yield varying results about a model's capabilities. For instance, a model might perform differently when using different tool configurations, testing the same task using different evaluation approaches, testing the same evaluation using different grading methods, or otherwise evaluating under different conditions. This variation makes interpretation of evaluation results challenging.

    - Given the potential magnitude of real-world harms in CB, organizations can attempt to measure tasks using diverse evaluation approaches, grading methods, and model configurations during their internal pre-deployment testing and carefully document these differences, erring on the side of caution when results conflict. Organizations can take steps to minimize false negatives in capability evaluations – that is, to avoid incorrectly concluding that a model lacks concerning capabilities.
- *Transparency*: Organizations can share evaluation results, including any apparent conflicts between different evaluation approaches and remaining uncertainties. Organizations can also share the number of runs conducted and inference parameters (temperature, top-p) used.

Careful consideration of potentially sensitive information and other security-relevant issues: Potentially sensitive information warrants particular attention when implementing and sharing the results of evaluations meant to assess CB misuse risks. Organizations can carefully weigh the scientific value of each measurement against the risk of revealing sensitive security-relevant information. This includes considering both direct technical details and how multiple pieces of information might be combined to cause harm. The ratio of beneficial to potentially harmful applications for any tested capability is important to consider when disclosing information publicly. Moreover, there are legal and regulatory frameworks, including export controls (including ITAR XIV[76]) and dual-use research of concern requirements (including but not necessarily limited to the USG Policy for Oversight of Dual Use Research of Concern (DURC) and Pathogens with Enhanced Pandemic Potential (PEPP)[77]), which may affect the what, where, and how of communicating about CB evaluations. Given security considerations, some information may be more appropriate to share with a limited group of organizations as well as government and non-government third-party evaluators.

D.2.3 Key Opportunities and Challenges for Measuring CB Risk

Current capabilities and measurement practices present several important opportunities and challenges for assessing CB misuse risk:

Opportunities:

- CB tasks typically enhance existing capabilities, enabling clear baseline comparisons with current resources (e.g., internet use, bioinformatics tooling) and expert performance
- Increasing standardization of evaluation frameworks enables systematic risk assessment through shared benchmarks and validated methodologies
- Multiple evaluation types allow effective triangulation of model capabilities and risks

Challenges:

- Novel capabilities, especially in biodesign, are difficult to assess without existing baselines or safe laboratory testing options
- Comprehensive risk assessment requires resource-intensive evaluation across the entire ideation-to-release pathway
- Balancing scientific transparency with security considerations may require careful consideration

D.3 Managing and Mitigating Chemical and Biological Misuse Risk

Building on the risk identification framework from D.1 and measurement approaches from D.2, this section expands on Objective 2 (“Plan for managing misuse risk”), Objective 5 (“Mitigate misuse risk

before deployment"), and Appendix B ("Example Safeguards Against the Misuse of Foundation Models") to apply specifically to CB misuse risks. Organizations can use their measurement results to select and calibrate appropriate safeguards for their context.

D.3.1 Considering Dual-Use Dynamics in Risk Management

CB capabilities in foundation models present complex dual-use dynamics that can be broadly categorized into three profiles:

- **Predominantly Harmful** capabilities with very limited legitimate applications
- **High-Impact Dual-Use** capabilities with significant legitimate applications but serious potential for harm, especially regarding transmissible CB agents
- **Mixed Dual-Use** capabilities with many legitimate applications and some potential for misuse

Based on specific risk assessments conducted as per Appendix D.1, organizations typically combine multiple safeguards from Appendix B to achieve risk levels acceptable to the organization. Understanding the relevant threat actors helps organizations implement proportionate security measures. Many CB misuse risks can be managed through basic API restrictions and user verification systems that keep out unsophisticated actors – for example, organizations may not need the highest levels of information security for protecting model weights unless their threat assessment identifies risks from sophisticated high-resource actors. These safeguards exist on a continuum rather than as binary measures, with different implementations requiring varying levels of expertise and resources to circumvent. Organizations can make evidence-based predictions about safeguard effectiveness against different threat actor profiles and document how these predictions inform their risk management decisions. Four safeguard approaches from Appendix B warrant particular attention in the CB domain:

1. **Improve the model's training**: Filtering pre-training data to exclude certain types of CB-related content, such as viral and toxin data, and implementing refusal training for harmful CB-related tasks may help reduce some risks while preserving beneficial capabilities. However, this approach has multiple limitations: models' performance at relevant tasks may generalize across biological domains even if not trained explicitly on certain CB agents, fine-tuning on excluded data can occur if model weights are exfiltrated by a malicious actor, and researchers who wish to use model capabilities for legitimate beneficial tasks may be impeded.
2. **Detect and block attempted misuse**: This is particularly important for "Predominantly Harmful capabilities" with minimal legitimate uses, such as queries relating to the circumvention of synthesis screening or dissemination of a CB agent. Relatedly, robustness to jailbreaking attempts is important, which can be tested through red-teaming simulating sophisticated jailbreaking attempts.
3. **Limit access to the model's capabilities:** Organizations can implement tiered systems where different user categories have access to different capability levels, with each tier combining appropriate access controls with other safeguards. For example:
    - **Mixed Dual-Use** capabilities may merit basic identity verification and monitoring
    - **High-Impact Dual-Use** capabilities may also merit institutional affiliation verification plus enhanced monitoring and detection systems
    - **Predominantly Harmful** capabilities may have highly restricted access with robust security measures, continuous monitoring, and comprehensive safeguards against misuse

4. **Collaborate across the supply chain to implement real-world protections**: Organizations may find it valuable to work with multiple types of suppliers and partners to strengthen security measures. For example, they could work with DNA synthesis providers to strengthen screening programs against AI-enabled circumvention attempts, coordinate with biological resource centers that maintain and distribute pathogen strains and pathogen-relevant cell lines to enhance verification of researcher credentials and intended use, partner with cloud laboratories to monitor for suspicious patterns in automated experiments, and collaborate with contract research organizations to identify concerning requests for specialized research services. This could include integrating AI model outputs with DNA synthesis screening through approaches like cryptographically signed certificates that capture metadata about design provenance and user intent.[78]

If an organization has found its foundation model to potentially enable meaningful risk, it is useful to document how implementing safeguards like those above will decrease risk prior to deploying the model. This can follow a similar structure as the Estimation of Marginal Risk process in D.1. In general, for capabilities that warrant higher security, organizations could consider implementing robust information security measures.[79]

D.3.2 Refining Risk Management Based on Real-World Evidence

The CB domain is dynamic and dual-use, with continuous advances in both legitimate research and potential misuse pathways from both increases in foundation model capabilities as well as advances in biotechnology. Organizations may gain significant insights about misuse risks after model deployment based on real-world usage patterns and research by third parties. This information can help refine threat assessments and mitigation strategies for both deployed and future models.

The table in D.4 categorizes common CB-relevant measurable tasks and their use profiles to help inform appropriate mitigation strategies. While this table presents capabilities as relatively standalone tasks, an alternative and potentially more useful approach is to map specific capabilities along key risk pathways of threat profiles.[80] For example, measurable tasks could be analyzed according to their role in different stages of a potential misuse pathway or mapped to specific technical processes like the laboratory steps involved in viral reverse genetics. This pathway-based analysis can provide additional insight into where and how different mitigation measures might be most effectively applied.

D.3.3 Key Opportunities and Challenges for Managing CB Risk

Opportunities:

- Structured scientific research frameworks enable clear implementation of access controls and monitoring systems
- Safeguards can be targeted to limit malicious access while enabling legitimate uses

Challenges:

- The dual-use nature of scientific knowledge complicates designing safeguards that do not overly restrict beneficial applications
- Severity of potential CB risks can require highly effective mitigations from initial deployment

Table D.4 Example Categorization of Potential CB Capabilities and Associated Tasks

Task Category	Use Profile	Potential Benefits	Potential Risks	Potential Measurable Tasks/Subtasks
Scientific Knowledge Integration	Mixed Dual-Use	Can enhance research efficiency and accelerate scientific discovery by making complex information more accessible and actionable	Can enable malicious actors to efficiently plan dangerous experiments and facilitate integration of sensitive CB information	• Hypothesis generation • Experiment planning • Literature review • Integration of information from multiple sources to form cohesive insights • New protocol development or detail-filling via extrapolation from other CB agents
Laboratory Assistance	High-Impact Dual-Use	Can improve laboratory efficiency, enhance experimental accuracy, and reduce human error	Can lower barriers for conducting complex and dangerous experiments by reducing required expertise	• Directing human usage of specialized biological or chemical tools/models • Generate step-by-step laboratory instructions • Multimodal laboratory troubleshooting • Help with lab setup to minimize failure modes
Laboratory automation	High-Impact Dual-Use	Can accelerate and reduce costs of beneficial biological research	Can automate or outsource key processes that could be exploited for harmful purposes	• Directly using specialized biological or chemical tools/models • Interfacing with cloud labs with minimal human input • Interfacing with and automating usage of bioinformatics software • Writing software/code for biological instruments
Agent Acquisition/ Synthesis	High-Impact Dual-Use	Can enable rapid and efficient production of beneficial agents for research and countermeasure development	Can provide detailed guidance for acquiring or synthesizing dangerous agents with reduced expertise requirements	• Generating information for how to acquire an agent from natural environment • Generating information for how to synthesize an agent in a laboratory • Generating information on how to verify that an agent has expected virulence/ pathogenicity
Design/*in silico* Testing of CB Agents[xiii]	High-Impact Dual-Use	Applications in understanding CB agent behavior and developing countermeasures	Can enable the deliberate modification of pathogens to increase their harmful potential	• Designs that enable the development of an agent described in the USG Policy for Oversight of DURC and PEPP • Ability to accurately predict results of new experiments

[xiii] This appendix describes potential risks produced by dual use foundation models; while out of scope for this appendix, additional considerations are needed for assessment of chem-bio AI models that could enable design and *in silico* testing of CB agents.

				• Discovery and analysis of novel chemical reactions • Design of pathways to produce toxic compounds
Weaponization/ Dissemination	Predominantly Harmful	Very limited legitimate applications	Can directly enable the weaponization and effective deployment of harmful agents	• Generating information for culturing of an agent in large quantities • Generating information for altering physical properties of agent re: formulation • Generating information for integration of an agent into a dissemination system • Generating information on agent behavior in various environmental conditions • Assistance with delivery via large-scale dissemination system
Biosecurity Circumvention	Predominantly Harmful	No legitimate applications except for assessment of models or systems via red-teaming	Can enable systematic circumvention of critical safety and security controls	• Technical circumvention of synthesis or customer screening • Generating information on obfuscating intent / avoiding law enforcement detection • Generating information on obfuscating markers useful for bioattribution

1

Appendix E: Application of NIST AI 800-1 to Cyber Misuse Risk

This appendix outlines additional considerations relevant to identifying, measuring, and mitigating risks associated with the misuse of models to assist with or automate the conduct of offensive cyber operations. Its goal is to complement the broader set of objectives and practices outlined in Section 5 of the main document by assisting organizations in understanding domain-specific considerations that can assist in applying practices for misuse risk management in this domain.[xiv] The considerations in this appendix are primarily relevant for foundation model developers and deployers, as well as third-party evaluators that work with developers and deployers to help assess cyber misuse risks; other actors' roles are discussed in Section 3 of the main document.

E.1 Identifying Cyber Misuse Risk

This section expands on Objective 1 ("Identify potential misuse risk") as it applies to cyber misuse risk, or the risk that actors might use a model to assist with or automate malicious offensive cyber activities, creating potentially significant impacts to public safety, national security, or economic security. It provides considerations to assist organizations in operationalizing key practices such as establishing threat profiles and assessing associated risks in the cyber domain.

E.1.1 Establishing Threat Profiles: Actors and Scenarios

Identifying Threat Actors: Given the diversity and number of cyber threat actors, organizations often cannot predict how each individual threat actor might misuse a model, nor can they fully ignore the distinctions between these actors when assessing potential model misuse.

Instead, organizations may benefit from identifying high-level groupings or types of actors who perform similar cyber activities and might therefore misuse foundation models' capabilities in similar ways, including by drawing from frameworks, standards, and concepts developed by cyber defenders to categorize threat actors for the purposes of cybersecurity risk management.[81] For example, based on factors including "motivation"[xv] and "sophistication",[xvi] defenders typically differentiate between nation state-associated hackers and financially motivated cyber criminals.[82] These groups may then be further sub-divided by sophistication – for example, differentiating between groups that use advanced capabilities such as zero-day exploits or custom-developed tooling versus those that rely on simpler access vectors and off-the-shelf tooling[83]. They may also be further subdivided by motivation – for example, financially motivated actors may have different goals such as ransomware and data extortion versus intellectual property theft; geopolitically motivated attackers might differ in goals such as espionage versus sabotage. Beyond these major profiles, other types of cyber threat actors might seek

[xiv] Note: While this appendix's concepts align with the main NIST AI 800-1 text, some section numbers and cross-references may not match exactly between the documents.

[xv] The reason or primary purpose for which a threat actor conducts offensive cyber actions. For example, see: An Introduction to the Cyber Threat Environment (2024), *Canadian Centre for Cybersecurity*, https://www.cyber.gc.ca/en/guidance/introduction-cyber-threat-environment.

[xvi] A concept which broadly encapsulates a group's level of access to expertise and resources that allow them to craft and use more bespoke and effective techniques.

to misuse model capabilities, including groups that sell offensive capabilities such as zero-days or initial access,[84] ideologically motivated non-state actors such as "hacktivists," or individual actors seeking to cause harm. Organizations can continually update threat actor profiles and prioritize additional distinctions in an evolving manner and as more details of AI-enabled cyber threats emerge from real-world misuse, as described in Practice 6.1.

Identifying Misuse Scenarios: Next, identifying cyber scenarios that would create significant potential impacts on public safety, national security, or economic security can help organizations identify capabilities that present serious misuse risks. Through this process, organizations can seek to connect particular model capabilities with specific, high-impact outcomes such as disruptions to critical infrastructure or large-scale theft of intellectual property.

Many high-impact cyber misuse scenarios may center around large increases in the scale or efficacy of different kinds of cyber attacks, such as increases in cybercrime, espionage activity, or attacks against critical infrastructure. The capabilities of foundation models could potentially increase the scale, volume, or efficacy of attacks in several ways, including:

- **Automation:** Allowing threat actors to perform more attacks with the same resources by automating time-intensive activities. For example, tasks such as compiling open-source information for attack planning or customizing malware for particular attack targets may be necessary but time-consuming for threat actors at present.
- **Attainment:** Allowing threat actors to increase their likelihood of succeeding at attacks, such as by adopting more sophisticated and effective techniques. For example, the cost of zero-day vulnerabilities on the gray market suggests that these capabilities are still relatively expensive to develop and may be out of reach for many actors. Likewise, targeted spearphishing emails may have higher success rates than generic phishing emails[85] but are more time-consuming to develop.
- **Accessibility:** Allowing a much wider range of actors to perform a particular kind of attack. For example, a model that was proficient and reliable at coaching non-expert operators through each phase of a cyber attack could potentially make the ability to perform such attacks accessible to a wider group of actors.

In addition to considering misuse scenarios based on large increases in the scale of existing activities, organizations may identify other high-impact misuse scenarios by studying significant past cyber attacks or wargaming with SMEs. For example, a review of cybersecurity history suggests that several of the costliest historical cyberattacks were caused by worms,[86] which provides evidence that this may be a high-impact misuse scenario. Organizations can refine misuse scenario assessments over time by integrating them with other risk management practices (such as those discussed in Section E.3); for instance, misuse scenarios may inform an organization's practices for monitoring or threat intelligence gathering, and, in turn, information from ongoing monitoring can be used to prioritize and update identified misuse scenarios.

Identifying Relevant Model Capabilities: Existing cybersecurity resources, such as taxonomies of attacker tactics, techniques, and procedures (TTPs),[87] models of the cyber kill chain,[88] and advisories on specific threat actor TTPs[89] can help organizations identify specific steps or activities in offensive cyber operations where foundation models' capabilities could be misused to increase a threat actor's

capabilities beyond what existing tools currently provide.[xvii] One particularly helpful frame may be to look for how a model's capabilities may help reduce current "bottlenecks" for a particular threat actor group or type of cyber activity, including those that relate to the scenarios of automation, attainment, and accessibility described above.

The increasing use of models as *agents* may also help automate larger offensive cyber workflows in addition to assisting with completion of discrete tasks. Organizations may consider risks both from capabilities at specific high-value tasks and from groups of capabilities that could enable models to autonomously execute multi-step workflows or even entire attack chains. Economic models of cyber crime may also be useful in identifying misuse scenarios,[90] such as to identify cyber activities or targets that are currently too low-return for expert human cyber operators to invest in exploiting at scale but that might be more often exploited if automation with foundation models lowered the costs of doing so significantly.

To identify the particular capabilities implicated in identified misuse scenarios and to inform the design of appropriate evaluations, organizations may consider:

- *Examples of specific real-world tasks and workflows that correspond to the misuse scenario.* Within a general capability area, there may be important gradations in difficulty: for example, within the general capability area of vulnerability discovery and exploitation, there is a significant difference in both difficulty and impact between common web vulnerabilities such as SQL injections and vulnerabilities in hardened software systems such as operating systems, firewalls, or identity and access management services. Specific, real-world example tasks and workflows can support the design of tailored capability evaluations.[91]
- *How the model would be used by the threat actor in this scenario.* Different threat scenarios may involve the use of a model in a *human uplift* setting, e.g. to aid a human operator; an *autonomous* setting, in which a model-agent system completes a particular action with little human oversight; or somewhere in between. Ideally, pre- and post-deployment risk evaluations can be designed to mimic this expected usage context. Likewise, evaluation design and interpretation may depend on questions about the expected level of expertise of the threat actor or the resources they are willing to expend to complete a particular task.
- *The level of model performance necessary for the threat scenario.* A foundation model may need a particular level of accuracy, performance, or reliability for a particular threat scenario. In some cyber misuse contexts, a model may be able to attempt a task many times (e.g., trying many inputs to a potentially vulnerable web app or debugging an exploit against a local target) and in others a model might need high accuracy on each attempt (e.g., identifying vulnerabilities that must then be explored and validated by a human operator, or using an exploit against a closely monitored remote target). There are varied ways to define and measure performance for a particular task, but two common approaches include:

[xvii] Threat actors currently use a variety of software tools to automate parts and processes of conducting cyber attacks (for example, see Ransomware Rebounds (2024). *Google.* https://cloud.google.com/blog/topics/threat-intelligence/ransomware-attacks-surge-rely-on-public-legitimate-tools); organizations may consider these existing tools in identifying existing gaps and workflows where AI capabilities could provide uplift above and beyond what is already available through existing software-based tools.

    - Directly defining an objective measure of the model's ability to obtain a binary or continuous outcome – for example, whether a model can successfully identify and exploit a specific kind of vulnerability.
    - Defining a way to compare the performance of the model to the performance of humans – for example, whether a model can write a phishing email that is more persuasive than a phishing email written by an expert human.
- *The types of real-world defenses with which a model might interact in a threat scenario.* Defensive systems, from intrusion detection systems to email filters, will play a key role in mediating the potential cyber misuse risks of model capabilities. Considering the dynamics of these interactions – including a model's adaptability to overcome such defenses, and how defenses might adapt in turn – can help inform risk assessments and the design of appropriate evaluations.

E.1.2 Risk Assessment

Best practices for cyber misuse risk assessment are still nascent. Existing frameworks for cyber risk assessment are primarily intended for assessment of cyber risks *to* an organization; however, these documents offer approaches to selecting and operationalizing qualitative and quantitative cyber risk assessments that organizations may adapt in seeking to assess cyber misuse risks.[92]

A widely used formula for quantitative risk assessments, both in cyber and in other domains, is to multiply the potential impacts of a negative outcome times the likelihood of such an event occurring. This approach may be well-suited for cyber misuse scenarios that focus on a single, high-impact event, but less tractable for scenarios that involve increasing the scale of current cyber activities across the ecosystem, where modelling the likelihood of each individual event may be prohibitively challenging. Instead, organizations may adapt this approach to model how a model capability might increase current or baseline risks associated with particular kinds of cyber activity.

Estimating Baseline Risk: Rather than separately estimating the impact and likelihood of cyber events, organizations may estimate a baseline level of risk based on the current costs or harms created by a particular category of cyber activity. A range of actors already aggregate and report data about cyber attacks and associated costs, which can help inform these estimates.[93] Organizations should consider using data and estimation techniques to assess costs that are both *comprehensive* and *specific* to the threat profile. Assessing comprehensive costs for certain kinds of cyber activities may require considering not only direct losses, but broader costs such as indirect losses (e.g., revenue loss due to downtime or declining consumer trust) and the costs of implementing defenses.[94] Some important harms may be challenging to estimate in dollar amounts, such as harms to US national security interests or privacy harms to individuals. Organizations might seek to estimate a dollar value for these harms using related data sources (e.g., on the costs that companies or governments currently pay to prevent these harms[95]) or use non-monetary variables to represent current baselines. In general, where high-quality data on baseline impacts is unavailable or where available data suggests a large range of possible values, organizations could use other processes to generate and refine estimates such as soliciting and aggregating independent expert estimates.

Estimating Misuse Risk: After estimating baseline risk for a category of cyber activities, organizations might estimate how a particular model capability (or a combination of capabilities) could modulate this

risk. There is not yet consensus on a single best approach to such risk modelling. One example could be using concepts of automation, attainment, and accessibility to estimate how a particular capability might scale threat actors' baseline output: for example, a model capability that provided automation could scale output by increasing the number of attacks that the average threat actor can attempt in a given time period; a capability that increased attainment could scale output by increasing the success rate of the average attack attempt; or a capability that increased accessibility could scale output by increasing the overall number of threat actors attempting attacks. For example, a model capability that allowed the average threat actor to automate a task that currently takes 20% of operation time could theoretically increase the number of attacks each actor attempts in a given time period by 25% (assuming perfect efficiency), resulting in a new estimated impact of 1.25*(baseline). Then, the misuse risk associated with that model capability would be given by the difference between the new estimated risk and the baseline risk, or 0.25*(baseline).

In reality, even “simple” relationships like the impacts of automation are unlikely to have purely linear effects because of relationships between variables – for example, increasing the average number of attacks that attackers can perform in a given time period might have sublinear effects because each marginal new target may be more difficult to compromise than the previous average, or superlinear effects because cheaper attacks could become appealing to a larger number of attackers. Organizations can seek to identify and to correct for the direction of these effects or to use modelling approaches that account for these relationships in their final estimates.

Estimating Misuse Risk from Rare, High-Impact Events: For cyber scenarios that center around infrequent but high-consequence events, organizations might use a more traditional risk estimation formula such as impact multiplied by likelihood – or, to estimate marginal risk, impact multiplied by the increase in likelihood as the result of the model capability. Estimates of the potential impact and likelihood (with and without the model capability) could be created by aggregating multiple independent expert assessments or drawing from case studies of significant past cyber attacks.

Transparency: The process of estimating the potential impacts of cyber capability misuse is still nascent and requires organizations to grapple with significant uncertainty. Public transparency about this process and its results can enable other parties – in particular the broader community of cybersecurity defenders, experts, and researchers – to offer feedback on specific assumptions and generally facilitate collaboration towards a stronger consensus on methodological best practices. Where possible, organizations might share full insights into their process and resulting estimates; where these analyses are too sensitive to share in their entirety, publishing information about risk assessment methodology, data sources, or open questions could still be useful to the broader field, as described in the documentation associated with the Objectives above and in Objective 7 in particular.

E.1.3 Key Challenges & Opportunities for Identifying Cyber Misuse Risk

- **Cyber-relevant model capabilities span a diversity of activities, actors, and contexts.** A single cyber operation may implicate diverse skillsets, activities, and expertise, from the ability to craft a persuasive phishing email to specific technical knowledge of the features of networking protocols or endpoint detection and response software. These activities vary across actors from “script kiddies” to well-resourced “advanced persistent threats”, and actors’ varied motivations, resources, and needs shape how they might misuse (and benefit from the misuse of) model capabilities. The cybersecurity field has a significant body of work characterizing this threat

environment for defenders; by building atop these foundations and engaging with the broader cybersecurity community, organizations can ground their practices in considerable prior work.

- **Foundation model-enabled cyber threats will not be static.** As defensive measures improve, threat actors will adapt their methods and seek new paths to exploitation, creating a moving target for cyber risk assessments. Organizations may need to continually update their understanding of attacker tradecraft and model capabilities, leverage threat intelligence, and refine their risk assessments in an iterative manner to keep pace. However, this ongoing process will create significant opportunities for organizations to closely integrate risk assessments with risk management practices, and to learn from and contribute back to the broader cyber defense ecosystem over time.
- **Cyber is a domain where the potential for model-enabled scale may be particularly powerful.** Foundation models' advantages with respect to automation may be particularly impactful in digital contexts like cyber, where activities – once automated – have the potential to be scaled arbitrarily without a bottleneck on human operators taking actions in the physical world. Automation-enabled scaling of particular cyber attack techniques or workflows could increase attackers' efficacy, alter the economics of offense and defense in the cyber ecosystem, and overwhelm defenders' resources. Organizations should consider the effects of scale in identifying high-impact scenarios for misuse of cyber-capable models.

E.2 Evaluating Cyber Capabilities

Building on Objective 4 ("Measure the Risks of Misuse") and the measurement characteristics outlined in Appendix C, this section provides considerations for measuring cyber misuse risks in the context of pre- or post-deployment model evaluations, including an overview of evaluation types in the cyber domain and specific methodological considerations for cyber capabilities evaluations.

E.2.1 Cyber Evaluation Types

Direct Measurement: The digital nature of the cyber domain permits organizations to directly measure many cyber capabilities through evaluations in which models interact with real or realistic information systems and resources in isolated, simulated, or otherwise non-harmful contexts. For example, evaluators could curate a collection of vulnerable open source codebases[96] or codebases with synthetically injected vulnerabilities[97] to measure a model's ability to identify vulnerabilities, or use "cyber ranges" or synthetic network environments to test whether a model can take actions such as establishing persistence, moving laterally, or evading common defenses.

Direct measurements can be designed as automated evaluations, in which a model or model-agent system's performance is automatically scored according to pre-defined criteria. In the cyber context, examples include:

- Environmental objectives, such as capture-the-flag style evaluations,[xviii] in which a model-agent system interacts with a vulnerable computer system or resource to find a hidden "flag" and passes or fails the task based on whether it successfully submits the flag within a certain number of messages.

[xviii] Note that capture-the-flag style evaluations is not necessarily synonymous with evaluations based on challenges from real-world "Capture the Flag" competitions; tasks designed specifically for the purposes of AI evaluation may still use a capture-the-flag style scoring mechanism for ease of automatic grading.

- Ground truth scoring, such as for vulnerability identification evaluations where model-identified vulnerabilities are compared to a known ground truth of vulnerabilities present in a particular code sample.[98]

Designs for automated direct measurements are less mature in capability areas like social engineering, where the need to assess interactions with humans creates additional methodological challenges. Direct measurement on real cyber tasks *without* automatic scoring – for example, having an expert use a model to try to complete a real cyber task – can provide valuable evidence about model capabilities, especially in the context of use by a human operator. However, standardization of best practices for qualitative judgements of model performance and methods for controlling for the effects of operator contributions are still nascent. Thus, these evaluations may be useful for exploring areas of capability and risk but less suited for use in performance comparisons.

Performance Comparisons: Performance comparisons help contextualize cyber capability measurements in comparison to existing models or human experts to support marginal risk assessment. For example, challenges from university- or professional-level "Capture the Flag" competitions are widely used as cyber benchmarks[99] because they provide self-contained environments to test vulnerability discovery and exploitation skills. These challenges differ from real-world vulnerabilities – in that they involve smaller and simpler artifacts and are designed to be solvable, entertaining puzzles – preventing direct mapping from models' performance on these challenges to estimates of capability at real-world vulnerability discovery and exploitation. However, because these challenges require similar skills, they can be useful in performance comparisons to existing models or to human experts: if a new model underperforms publicly available models on CTF challenges, it provides some evidence that it is unlikely to be significantly more capable at real-world vulnerability discovery; if a model cannot solve CTFs that take human experts a relatively short amount of time, it provides some evidence that the model is unlikely to succeed at real-world vulnerability discovery tasks that take human experts a much greater amount of time. For these kinds of performance comparisons, organizations should validate key assumptions about the relationship between the evaluation tasks and the real-world tasks of concern, such as by soliciting opinions from multiple independent experts.

E.2.2 Considerations for Cyber Evaluations

Breadth and Depth: When selecting evaluations for a particular cyber capability, organizations may consider both *breadth* and *depth* of the evaluations:

- **Breadth:** whether evaluation tasks comprehensively cover the distribution of real-world tasks of concern, especially if there is significant variety in these tasks. For example, assessing a model's ability to find memory management vulnerabilities may provide weak evidence about its ability to find design flaws in identity and access management infrastructure and vice versa.
- **Depth:** whether a benchmark has *headroom* – room for a model to improve its score without saturating the benchmark – and gradations of difficulty sufficient to track progress in model capabilities over time. Organizations may be able to add additional gradations in difficulty through evaluation design, for example, by varying the amount of information provided to a model about the vulnerability that it needs to exploit in a capture-the-flag challenge. Human baselines (e.g., measures of how long the tasks in a benchmark take human experts to complete) can help evaluators assess the relative difficulty of different tasks as well as support in conducting performance comparisons.

Elicitation: Especially for agent-based evaluations like capture the flag-style evaluations, *elicitation* is a key methodological consideration. Elicitation refers to efforts to maximize the performance of a model on a task through methods such as improving prompts, using more performant agent scaffolds, or fine-tuning. In the cyber evaluation context, elicitation can help ensure that a model is measured on tasks in a way that comparably effective to how it could be used by threat actors in the misuse scenario the evaluation is designed to measure. Research results and observations of the performance of tailored agents for cyber applications suggest that elicitation can significantly increase model performance.[100]

- Organizations may use a process for elicitation prior to or during an evaluation to select an agent scaffold that is optimized for both the model being evaluated and the specific evaluation tasks. Organizations can use a "development set" of similar tasks that is separate from the "test set" of evaluation tasks to experiment with agent architectures and prompts, review transcripts to address causes of failure, and select the best-performing agent scaffold.[101]
- Access to appropriate cybersecurity tools and utilities such as a code interpreter, debugger, decompilation tool, file editing tool, or web browsing tool may significantly improve agent performance on cyber tasks.[102] Useful practices may include providing agents with tools comparable to what a human expert would use to complete a task, reviewing task transcripts to identify tools that agents attempt to use during task runs and pre-installing them in the agent environment, and informing agents of available tools in initial prompts.
- Organizations can form evidence-based hypotheses about the magnitude of potential post-release performance gains through additional elicitation and factor this into assessments of model capability. If a developer believes that existing agent scaffolds may not be well-suited for the evaluated model, such as if it is paradigmatically different from previous state-of-the-art models, they may use a larger estimate.
- Organizations can observe how models are applied and customized to real-world cyber tasks and applications in practice and seek to use similar designs and methods during evaluations.

Running Cost-Aware Evaluations: Evaluators may consider documenting or controlling for costs when running cyber evaluations, either to approximate the costs that a threat actor might be willing to pay for a particular misuse case (for direct measurement) or to support performance comparisons between models or with human experts. There are several ways that additional inference compute can improve model performance, including allowing models longer chains of thought or more tool calls, or running more separate trials and combining the results such as by taking a majority vote or using the best outcome (for tasks like vulnerability discovery, where it is possible to verify success).[103] Organizations may use these methods to help equalize inference costs across models as appropriate.

Contamination: Since many cyber evaluations are based on public data, evaluators should be cognizant of contamination risks. Contamination could arise not only from public cyber benchmarks in training data but also from other publicly available information about evaluation tasks – such as descriptions of past CVEs or write-ups of CTF challenges – present in training data or queried by an internet-connected agent at evaluation time. Evaluators may seek or develop benchmarks that rely on private data and use held-out test sets to reduce the risk of contamination, and organizations should avoid overlap between data used to train a model and data used in capability evaluations, including by respecting canary strings, as described in Practice 4.1[104]

Standardization and Transparency: Organizations may consider how to design and document cyber evaluations to allow for consistent comparison over time and across different models. Standardization of metrics and testing procedures, documentation of key methodological decisions, and reporting of supporting results such as ablation experiments for elicitation techniques can support the cross-comparison and replication of results within and outside of an organization over time.

E.2.3 Key Challenges and Opportunities for Measuring Cyber Capabilities

- **Models will be applied and customized to cyber tasks in ways beyond what can be measured in a time-bound evaluation setting.** Attackers and defenders are constantly evolving, experimenting, and engineering to gain the upper hand in the cyber domain. Once a model is released, both specific threat actors and the broader cybersecurity community will experiment in applying it to new cybersecurity tasks and may improve its performance through customized scaffolds and new methods of elicitation. Evaluators cannot realistically recreate this decentralized development process within the timebound context of an evaluation, creating challenges in mapping from evaluation results to estimates of real-world capability; better methods for estimating the size of this effect are still needed. Evaluators can gain valuable insight from observing applications of deployed models, which can help inform analysis of cyber misuse risks and improve methods for capability measurement.
- **Designs, datasets, and standard tooling for cyber evaluations are still being developed.** While the digital nature of the cyber domain has advantages in allowing the direct evaluation of models on a variety of real or realistic cyber tasks, the creation of datasets, tools, and resources for cyber evaluations is still in its early days. Benchmarks have been publicly released that are comprised of CTF challenges or based on real vulnerability data have been publicly released, but there are fewer purpose-built benchmarks for simulating entire networks and attack chains or evaluating capabilities in social engineering. Standard frameworks exist for running agent-based evaluations[105] but they are not universally adopted. And additional auxiliary tooling, such as tooling to assist in performing elicitation or reviewing task transcripts for causes of failure, is still being developed.

E.3 Managing and Mitigating Cyber Misuse Risk

This section expands on Objective 2 ("Plan for Managing Misuse Risk"), Objective 5 (“Mitigate Misuse Risk Before Deployment”), and Appendix B (“Example Safeguards Against the Misuse of Foundation Models”). It outlines some key considerations in operationalizing mitigations for cyber misuse risks and calls attention to several as-yet unsolved challenges in this area.

E.3.1 Considering Dual-Use Dynamics in Risk Management

Reducing Compliance with Harmful Requests: The dual-use nature of many cyber tools and the close resemblance – in the absence of additional context about a user’s intent or authorization – between malign and benign cyber tasks creates challenges for mitigations like safety training for refusals or auxiliary request filtering that aim to reduce compliance with harmful requests at inference time. A threat actor might disguise a request to craft a phishing email as assistance in drafting an outreach

email, ask a model to identify vulnerabilities in code in the guise of a developer seeking to address them, or request distinct commands or pieces of code that are malicious only when combined. Organizations may not be able to fully prevent compliance with such misuse requests without also blocking certain benign use cases. While reducing helpfulness with overtly malicious cyber tasks may still be an important component of broader cyber misuse risk management, organizations may need to use these approaches in concert with other mitigations.

User Accounts: For models offered through an API, requiring user accounts and monitoring longer-running usage patterns for signs of offensive cyber misuse can allow organizations to identify and block usage by cyber threat actors.[106] However, this monitoring may still face similar challenges in differentiating cyber misuse from legitimate use cases such as security research.

Staged Releases and User Vetting: For highly cyber-capable models, organizations may employ staged releases to advantage defenders' access to a model ahead of attackers': for example, a developer could make a model with significant vulnerability discovery capabilities available to verified organizations and developers before allowing access by unverified users, allowing them to use the model to find and fix some vulnerabilities before attackers can use the model to discover and exploit them. However, staged release strategies require realism about the broader cybersecurity ecosystem: adoption and integration of defensive tools takes time, slow patching is a known challenge in cybersecurity,[107] and significant numbers of legacy systems can no longer receive upgrades and security patches even if vulnerabilities are found.[108]

Modifying Outputs to Reduce Misuse Potential: In certain cases, organizations may be able to modify model outputs to asymmetrically reduce the utility of the outputs for threat actors without needing to differentiate between malign and benign uses at inference time. For example, placing a visible watermark or indicator on an AI-generated video of a person's likeness could make it less useful for social engineering attacks.[109]

E.3.2 Using Identified Threat Actors to Inform and Evaluate Risk Mitigations

In the cyber context in particular, threat actors who might wish to misuse a model for cyber offense may be willing and able to use offensive cyber techniques to gain access to models and circumvent protections for the purposes of misuse. Thus, information security measures and adversarial testing may play significant roles in the cyber misuse management strategies of developers and deployers.

Organizational Security Practices: Organizations cyber misuse risk management plans should include consideration of whether organizational information security measures are sufficient to prevent unauthorized access to models by identified cyber misuse actors (i.e., potentially up to and including nation-state threat actors) wherever unauthorized access would subvert key assumptions about misuse risk management, such as that model access is controlled through an API. Beyond employing general, industry-standard cybersecurity best practices, organizations may consider additional security practices specific foundation models, as described in Practices 3.1 and 3.2.

Protections for Deployed Models: When applying misuse risk mitigations for deployed models such as user account-level protections or measures to reduce model compliance with malign requests,

organizations should evaluate the efficacy of these measures considering the resources and sophistication of the adversaries they are intended to protect against.

Red-Team Testing: Traditional cybersecurity red-team testing of organizational networks and deployment infrastructure, as well as AI-specific adversarial red-teaming of model safeguards, can assist in validating the robustness of these measures and closing exploitable gaps.

E.3.3 Refining Risk Management Based on Real-World Evidence

Cybersecurity is a fast-evolving and adversarial environment, and organizations will receive significant evidence based on real-world use of their foundation models and proxy models that can be used as a part of risk management practices.

Monitoring for Misuse: If offered through an API, organizations can monitor deployed models for misuse to apply mitigations such as restricting user accounts or updating systems to reduce compliance with malicious requests. Regardless of deployment strategy, organizations can monitor other sources of public information such as forums, news reports, and cybersecurity advisories for evidence of offensive cyber misuse of their foundation models or similar models. This information can be useful for updating safeguards for previously released models and for refining risk assessments, measurements, and mitigations for future models.

Tracking Cyber Capabilities and Applications: Organizations can stay up to date on the work of other companies and researchers to understand how users have customized and applied models for real-world cyber tasks and applications, including where scaffolds or other methods of elicitation have improved model performance on these types of tasks. This information can refine assessments of model capability in real-world settings as compared to measurements in pre-deployment evaluations and feed into the design of future evaluations such as by providing new methods of elicitation.

Planning for Adaptive Defenses: Organizations can seek to prioritize the development of layered and adaptable defensive mechanisms that are designed to be updated over time based on evolving information about the threat landscape and attacker methods. For example, organizations might create organizational structures and technical methods to rapidly update misuse detection tools based on ongoing threat intelligence, or develop AI systems to assist in rapidly identifying and responding to new patterns of misuse.

Transparency: Publishing information on identified cyber misuse of models can assist the broader ecosystem in understanding real-world cyber misuse risks. Best practices for detecting and disrupting cyber misuse of foundation models are still nascent, and sharing methodologies, information, or tools relating to cyber misuse detection – with the broader public where possible without compromising their efficacy, and with smaller and more trusted groups for more sensitive information – can assist other organizations in maturing their own detection and disruption methods.

E.3.4 Key Challenges and Opportunities for Mitigating Cyber Misuse Risk

- **Dual-use dynamics are fundamental to cyber misuse risk management but not yet fully characterized.** The dynamics of the offense-defense interplay in cyber mean that most capabilities will be, in some sense, dual use. However, the specific dynamics of this interplay, including whether and how defensive uses can counteract offensive ones, will vary based on capabilities and the broader characteristics of the cyber ecosystem. While early theoretical work has explored this space,[110] further study is needed. These dynamics will challenge organizations to design mitigations that can reduce the misuse potential of a model without impacting benign use cases but may also create opportunities to deploy capabilities for defensive purposes.
- **Reliably differentiating cyber misuse from legitimate use cases may be challenging.** The use of a model to write software, draft emails, and even to actively probe systems may not be easily distinguished between beneficial activities like security research and threat actor misuse without additional context, presenting challenges for designing and implementing mitigations such as refusals, request filtering, and user account-level interventions. Monitoring additional sources of public information and providing reporting mechanisms may help provide some additional context to distinguish security research from offensive cyber operations, and further transparency around practices and open questions in this area can assist in the maturation of consensus best practices.[111]
- **Evolving attacker techniques will require adaptive defenses.** Threat actors are likely to continually adapt their techniques to circumvent established safeguards, rendering once-effective controls quickly outdated, and organizations may need to continually evolve their risk mitigations in tandem with attacker methods. Organizations can seek to create adaptive mitigations that are designed to be updated over time, based on feedback loops from real-world incidents and threat intelligence, to help maintain the effectiveness of safeguards over time.

---

[1] Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence (2024). https://www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/.

[2] Anwar, U. et al., (2024) Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *arXiv*. https://arxiv.org/pdf/2404.09932. For information on the rate of progress of foundation models, see Maslej, N. et al., (2024) The AI Index 2024 Annual Report. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University*. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf; Ho, A. et al., (2024) Algorithmic progress in language models. *arXiv*. https://arxiv.org/abs/2403.05812; and Sevilla, J. et al., (2024) Training Compute of Frontier AI Models Grows by 4-5x per Year. *Epoch AI*. https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

[3] Vogel, M. et al., (2024) Findings & Recommendations: AI Safety. *National AI Advisory Committee*. https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf.

[4] Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023) *The White House.* https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[5] Dual-Use Foundation Models with Widely Available Model Weights Report (2024). *National Telecommunications and Information Administration*. https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report ; Kapoor, S. et al., (2024) Position: On the Societal Impact of Open Foundation Models.

Proceedings of the 2024 International Conference on Machine Learning. https://openreview.net/pdf?id=jRX6yCxFhx.

[6] Blueprint for an AI Bill of Rights (2022) *The White House*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[7] NIST AI 600-1 AI Risk Management Framework: Generative Artificial Intelligence Profile. *National Institute of Standards and Technology*. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

[8] Srikumar, M. et al., (2024) Risk Mitigation Strategies for the Open Foundation Model Value Chain. *Partnership on AI*. https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/; Heim, L. et al., (2024) Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. *arXiv*. https://arxiv.org/pdf/2403.08501; and Gorwa, R. et al. (2023) Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries. *arXiv*. https://arxiv.org/abs/2311.12573.

[9] Ibid.

[10] Constanza-Chock, S. et al., (2022) Who Audits the Auditors? Recommendations from a field scan of algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. https://dl.acm.org/doi/abs/10.1145/3531146.3533213.; Cen, S. et al., (2024) AI Supply Chains. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4789403.

[11] Narayanan, A. and Kapoor, (2024) AI safety is not a model property. *AI Snake Oil.* https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.; T. et al., (2023) AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature*. https://www.nature.com/articles/s41592-023-02087-4.

[12] S. et al., (2024) Position: On the Societal Impact of Open Foundation Models. *Proceedings of the 2024 International Conference on Machine Learning*. https://openreview.net/pdf?id=jRX6yCxFhx.

[13] Hoffmann, J. et al., (2022) Training Compute-Optimal Large Language Models. *Proceedings of the 2022 Conference on Neural Information Processing Systems*. https://dl.acm.org/doi/10.5555/3600270.3602446.

[14] Besiroglu, T. et al., (2024) Chinchilla Scaling: A replication attempt. *arXiv*. https://arxiv.org/pdf/2404.10102.

[15] Anwar, U. et al., (2024) Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv. https://arxiv.org/pdf/2404.09932.

[16] Upadhayay, B. and Behzadan, V. (2024) Sandwich Attack: Multi-language Mixture Adaptive Attack on LLMs. *arXiv*. https://arxiv.org/abs/2404.07242; Zou, A. et al., (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*. https://arxiv.org/pdf/2307.15043.

[17]Adapted from OECD Digital Economy Papers, (2022) OECD Framework for the Classification of AI Systems. *OECD.* https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html.

[18] Kapoor, S. et al., (2024) Position: On the Societal Impact of Open Foundation Models. *Proceedings of the 2024 International Conference on Machine Learning*. https://openreview.net/pdf?id=jRX6yCxFhx; and Mouton, C. et al., (2024) The Operational Risks of AI in Large-Scale Biological Attacks. *RAND*. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[19] Bernardi, J. et al., (2024) Societal Adaptation to Advanced AI. *arXiv.* https://arxiv.org/abs/2405.10295.

[20] NIST AI 100-1 AI Risk Management Framework. *National Institute of Standards and Technology.* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf; NIST AI 600-1 AI Risk Management Framework: Generative Artificial Intelligence Profile. *National Institute of Standards and Technology*. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf; and NIST SP 800-30 Guide for Conducting Risk Assessments. *National Institute of Standards and Technology.* https://csrc.nist.gov/pubs/sp/800/30/r1/final.

[21] Nevo, S., et al., (2024) Securing AI Model Weights. *RAND*. https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[22] NIST SP 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations (2020). *National Institute of Standards and Technology.* https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final; NIST SP 800-218A: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile (2024). *National Institute of Standards and Technology.* https://csrc.nist.gov/pubs/sp/800/218/a/ipd.

[23] Ibid.

[24] Carlini, N. et al., (2024) Stealing Part of a Production Language Model. *arXiv*, https://arxiv.org/pdf/2403.06634; Nevo, S., et al., (2024) Securing AI Model Weights. *RAND.* https://www.rand.org/pubs/research_reports/RRA2849-1.html.
[25] Casper, S. et al., (2024) Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv*. https://arxiv.org/pdf/2401.14446.
[26] Bran, A. et al., (2024) Augmenting large language models with chemistry tools. *Nature*. https://www.nature.com/articles/s42256-024-00832-8; Davidson, T. et al., (2023) AI capabilities can be significantly improved without expensive retraining. *arXiv*. https://arxiv.org/pdf/2312.07413; Measuring the impact of post-training enhancements. *METR*. https://metr.github.io/autonomy-evals-guide/elicitation-gap/.
[27] Jin, H. et al., (2024) JailbreakZoo: Survey, Landscapes, and Horizons in Jailbreaking Large Language and Vision-Language Models. *arXiv*. https://arxiv.org/pdf/2407.01599.
[28] O'Brien, J. et al., Deployment Corrections: An incident response framework for frontier AI models. *Institute for AI Policy and Strategy*. https://arxiv.org/pdf/2310.00328.
[29] Cattell, S. et al., (2024) Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities. *arXiv.* https://arxiv.org/abs/2402.07039; for an example of researchers pursuing coordinated disclosure practices, see Nasr, M., et al., (2023) Scalable Extraction of Training Data from (Production) Language Models. *arXiv*. https://arxiv.org/pdf/2311.17035.
[30] Longpre, S. et al., (2024) A Safe Harbor for AI Evaluation and Red-Teaming. *arXiv*. https://arxiv.org/pdf/2403.04893.
[31] Bucknall, B. et al., (2023). Structured Access for Third-Party Research on Frontier AI Models. *Center for the Governance of AI*. https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models.
[32] Wan, S. et al., (2024) Bridging the Gap: A Study of AI-based Vulnerability Management between Industry and Academia. *arXiv*. https://arxiv.org/abs/2405.02435.
[33] Phuong, M. et al., (2024) Evaluating Frontier Models for Dangerous Capabilities. *arXiv*. https://arxiv.org/pdf/2403.13793; Anthropic (2024) Claude 3.5 Sonnet Model Card Addendum. *Anthropic*. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf; OpenAI (2023) GPT-4 System Card. *OpenAI*. https://cdn.openai.com/papers/gpt-4-system-card.pdf
[34] Bommasani, R. et al., (2024) The Foundation Model Transparency Index v1.1: May 2024. *arXiv*. https://arxiv.org/pdf/2407.12929.
[35] Longpre, S. et al., (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv*. https://arxiv.org/abs/2310.16787.
[36] Voluntary AI Commitments. *White House*. https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf; Bommasani, R. et al., (2024) The Foundation Model Transparency Index v1.1: May 2024. *arXiv*. https://arxiv.org/pdf/2407.12929.
[37] Voluntary AI Commitments (2023). *The White House.* https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.; Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (2023). *G7 2023 Hiroshima Summit.* https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05_en.pdf.; Frontier AI Safety Commitments (2024). *AI Seoul Summit 2024.* https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.
[38] McGregor, S. (2020) Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *arXiv*. https://arxiv.org/pdf/2011.08512.
[39] One example of such a format is the OECD AI Incidents Monitor (AIM).
[40] Turri, V. et al., (2023) Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. *AAAI/ACM Conference on AI Ethics and Society*. https://doi.org/10.1145/3600211.3604700.
[41] The term "artificial intelligence" or "AI" has the meaning set forth in 15 U.S.C. 9401(3).
[42] Executive Order 14110.

[43] Executive Order 14110.
[44] NIST AI 100-1 AI Risk Management Framework. *National Institute of Standards and Technology.* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[45] Thiel, D., et al., (2023) Identifying and Eliminating CSAM in Generative ML Training Data and Models. *Stanford Digital Repository*. https://purl.stanford.edu/kh752sm9123.
[46] Yuan Y. et al., (2024) Refuse Whenever You Feel Unsafe: Improving Safety in LLMs via Decoupled Refusal Training. *arXiv.* https://arxiv.org/abs/2407.09121 .
[47] Henderson, P. et al., (2022) Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models. *arXiv*. https://arxiv.org/pdf/2211.14946.
[48] https://www.ingentaconnect.com/contentone/hsp/airwa/2024/00000003/00000001/art00005; Lynch, A. et al., (2024) Eight Methods to Evaluate Robust Unlearning in LLMs. *arXiv*. https://arxiv.org/pdf/2402.16835.
[49] Bommasani, R. (2023) Ecosystem graphs: The social footprint of foundation models. https://arxiv.org/abs/2303.15772.
[50] Seger, E. et al., (2024) Open-Sourcing Highly Capable Foundation Models. https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf
[51] Qi, X. et al., (2024) Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv*. https://arxiv.org/pdf/2310.03693.
[52] Kapoor, S. et al., (2024) Position: On the Societal Impact of Open Foundation Models. *Proceedings of the 2024 International Conference on Machine Learning*. https://openreview.net/pdf?id=jRX6yCxFhx.
[53] Solaiman, I. (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv*. https://arxiv.org/pdf/2302.04844.
[54] Framework for Nucleic Acid Synthesis Screening (2024). *National Science and Technology Council.* OSTP-Nucleic-Acid_Synthesis_Screening_Framework-Sep2024-Final.pdf (whitehouse.gov).
[55] Rein, D. et al., (2023) GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv*. https://arxiv.org/abs/2311.12022; Laurent, J. et al., (2024) LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv*. https://arxiv.org/abs/2407.10362; and Hendrycks, D. et al., (2021) Measuring Massive Multitask Language Understanding. arXiv. https://arxiv.org/abs/2009.03300.
[56] Department of Homeland Security (2024) Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats *Department of Homeland Security*.https://www.dhs.gov/publication/fact-sheet-and-report-dhs-advances-efforts-reduce-risks-intersection-artificial.
[57] The White House (2024) United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential. *The White House.* https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.
[58] Rose, S. et al., (2024) The near-term impact of AI on biological misuse. *CLTR.* https://www.longtermresilience.org/reports/the-near-term-impact-of-ai-on-biological-misuse/.
[59] Department of Homeland Security (2024) Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats. *Department of Homeland Security*. https://www.dhs.gov/publication/fact-sheet-and-report-dhs-advances-efforts-reduce-risks-intersection-artificial.
[60] Department of Homeland Security (2010) Federal Incident Management Planning. *Department of Homeland Security.* https://www.oig.dhs.gov/sites/default/files/assets/Mgmt/OIGr_10-58_Feb10.pdf
[61] Biosafety refers to the "practices, controls, and containment infrastructure that reduce the risk of unintentional exposure to, contamination with, release of, or harm from pathogens, toxins, and biological materials" (NIST Bioeconomy Lexicon: https://www.nist.gov/bioscience/nist-bioeconomy-lexicon).
[62] Biosecurity refers to the "security measures designed to prevent the loss, theft, misuse, diversion, unauthorized possession or material introduction, or intentional release of pathogens, toxins, biological materials, and related

information and/or technology" (NIST Bioeconomy Lexicon: https://www.nist.gov/bioscience/nist-bioeconomy-lexicon).

[63] Rose, S. et al., (2024) The near-term impact of AI on biological misuse. *CLTR.* https://www.longtermresilience.org/reports/the-near-term-impact-of-ai-on-biological-misuse/.

[64] Safety considerations for chemical and/or biological AI models. (2024) Federal Register. https://www.federalregister.gov/documents/2024/10/04/2024-22974/safety-considerations-for-chemical-andor-biological-ai-models.

[65] Building an early warning system for LLM-aided biological threat creation (2024) *OpenAI.* https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/

[66] Mouton, C. et al., (2024) The operational risks of AI in large-scale biological attacks: Results of a red-team study. *RAND.* https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[67] OpenAI o1 System Card (2024). *OpenAI.* https://cdn.openai.com/o1-system-card-20240917.pdf.

[68] Pre-deployment evaluation of Anthropic's upgraded Claude 3.5 Sonnet. (2024). *NIST.* https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet.

[69] Pre-deployment evaluation of OpenAI's o1 model (2024). *NIST.* https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model.

[70] Useful expert assessment typically draws on diverse expertise including: scientific research, biosecurity, law enforcement and national security, biodefense, as well as other relevant areas as specified in the USG Policy for Oversight of Dual Use Research of Concern https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.

[71] Laurent, J. M., et al., (2024). LAB-Bench: Measuring capabilities of language models for biology research. arXiv. https://doi.org/10.48550/arXiv.2407.10362.

[72] Pre-deployment evaluation of Anthropic's upgraded Claude 3.5 Sonnet. (2024). *NIST.* https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet.

[73] Pre-deployment evaluation of OpenAI's o1 model (2024). *NIST.* https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model.

[74] Updated responsible scaling policy (2024) *Anthropic.* https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy.

[75] OpenAI's CBRN tests (2024) *Planned Obsolescence*. https://www.planned-obsolescence.org/openais-cbrn-tests-seem-unclear/.

[76] Code of Federal Regulations Title 22, Part 121 (2024) *U.S. Government Publishing Office*. https://www.ecfr.gov/current/title-22/part-121.

[77] The White House (2024) United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential. *The White House.* https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.

[78] Carter, S., et al., Developing Guardrails for AI Biodesign Tools (2024) *Nuclear Threat Initiative.* https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/.

[79] Nevo, S., et al. A playbook for securing AI model weights (2024). *RAND.* https://www.rand.org/pubs/research_briefs/RBA2849-1.html.

[80] Building an early warning system for LLM-aided biological threat creation (2024). *OpenAI.* https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

[81] The NIST Cybersecurity Framework (CSF) 2.0 (2024). National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf; ISO/IEC 27001:2022 (2022). International Standards Organization. https://www.iso.org/standard/27001.

[82] Nation-State Cyber Actors (2024). *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/topics/cyber-threats-and-advisories/nation-state-cyber-actors; How Microsoft names threat actors (2024). *Microsoft*. https://learn.microsoft.com/en-us/defender-xdr/microsoft-threat-actor-naming; Naming Adversaries and Why It Matters to Your Security Team (2024). *CrowdStrike*. https://www.crowdstrike.com/en-us/blog/naming-adversaries-and-why-it-matters-to-security-teams/.

[83] Task Force Report: Resilient Military Systems and the Advanced Cyber Threat (2013). *Department of Defense Defense Science Board.* https://nsarchive2.gwu.edu/NSAEBB/NSAEBB424/docs/Cyber-081.pdf.

[84] Trends on Zero-Days Exploited In-the-Wild in 2023 (2023). *Google.* https://cloud.google.com/blog/topics/threat-intelligence/2023-zero-day-trends.

[85] Heiding, Fred et. al. Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects (2024). *arXiv.* https://arxiv.org/abs/2412.00586 ; Top Phishing Statistics for 2025: Latest Figures and Trends (2024). *StationX.* https://www.stationx.net/phishing-statistics/.

[86] The Untold Story of NotPetya, the Most Devastating Cyberattack in History (2018). *Wired.* https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/; Cyber-attack: US and UK blame North Korea for WannaCry (2017). *BBC.* https://www.bbc.com/news/world-us-canada-42407488 .

[87] ATT&CK (2024). *MITRE.* https://attack.mitre.org/.

[88] The Cyber Killchain (2024). *Lockheed Martin*. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

[89] Cybersecurity Alerts & Advisories (2024). *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/news-events/cybersecurity-advisories ; NSA Cybersecurity Advisories and Guidance (2024). *National Security Agency.* https://www.nsa.gov/press-room/cybersecurity-advisories-guidance/.

[90] Westland, Chris. A Rational Choice Model of Computer and Network Crime (1997). *International Journal of Electronic Commerce.* https://www.jstor.org/stable/27750812.

[91] XBOW Validation Benchmarks (2024). *XBOW.* https://github.com/xbow-engineering/validation-benchmarks/tree/main.

[92] Guide for Conducting Risk Assessments, SP 800 30-r. *NIST.* https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-30r1.pdf; An Introduction to the Factor Analysis of Information Risk (FAIR) (2017). *FAIR Institute.* https://www.fairinstitute.org/what-is-fair.

[93] The Cost of Malicious Cyber Activity to the U.S. Economy (2018). *The White House.* https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/02/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf; Federal Bureau of Investigation Internet Crime Report 2023 (2023). *Federal Bureau of Investigation.* https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-releases-internet-crime-report ; Net Losses: Estimating the Global Cost of Cybercrime (2014). *Center for Strategic and International Studies.* https://www.csis.org/analysis/net-losses-estimating-global-cost-cybercrime.

[94] Anderson, R, Barton, C, Böhme, R, Clayton, R, Gañán, C, Grasso, T, Levi, M, Moore, T & Vasek, M. Measuring the Changing Cost of Cybercrime (2019). *Workshop on the Economics of Information Security.* https://www.researchgate.net/publication/350793602_Measuring_the_changing_cost_of_cybercrime.

[95] Sec 15, Information Technology and Cybersecurity Funding, Budget for Fiscal Year 2025 (2024). *The White House.* https://www.whitehouse.gov/wp-content/uploads/2024/03/ap_15_it_fy2025.pdf.

[96] Chauvin, Timothee. eyeballvul: a future-proof benchmark for vulnerability detection in the wild. *arXiv.* https://arxiv.org/abs/2407.08708.

[97] AI Cyber Challenge (2024). *DARPA.* https://aicyberchallenge.com/.

[98] Chauvin, Timothee. eyeballvul: a future-proof benchmark for vulnerability detection in the wild.

[99] Zhang, Andy, et. al. CyBench: A benchmark for evaluating the cybersecurity capabilities and risks of language models (2024). *arXiv.* https://cybench.github.io/ ; Shao, M. et. al. NYU CTF Dataset: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security (2024). *arXiv.* https://github.com/NYU-LLM-CTF/NYU_CTF_Bench.

[100] Turtayev, Rustem et. al. Hacking CTFs with Plain Agents (2024). *arXiv.* https://arxiv.org/pdf/2412.02776; Abramovich, Talor et. al. EnIGMA: Enhanced Interactive Generative Model Agent for CTF Challenges (2024). a*rXiv.* https://arxiv.org/abs/2409.16165; Waisman, Nico and Dolan-Gavitt, Brendan. How XBOW found a Scoold authentication bypass (2024). *XBOW.* https://xbow.com/blog/xbow-scoold-vuln/.
[101] Guidelines for capability elicitation (2024). *METR.* https://metr.github.io/autonomy-evals-guide/elicitation-protocol/.
[102] Glazunov, Sergei and Brand, Mark. Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models (2024). *Google.* https://googleprojectzero.blogspot.com/2024/06/project-naptime.html.
[103] US AISI and UK AISI Joint Pre-Deployment Test, Anthropic's Claude 3.5 Sonnet (October 2024 Release) (2024). *US AI Safety Institute and UK AI Safety Institute.* https://www.nist.gov/system/files/documents/2024/11/19/Upgraded%20Sonnet-Publication-US.pdf; Kapoor, Sayesh and Narayanan, Arvind. AI leaderboards are no longer useful. It's time to switch to Pareto curves (2024). AI Snake Oil Blog. https://www.aisnakeoil.com/p/ai-leaderboards-are-no-longer-useful.
[104] Evaluations Canary (2024). Alignment Research Center. https://www.alignment.org/canary/.
[105] Inspect (2024). UK AI Safety Institute. https://inspect.ai-safety-institute.org.uk/ ; Vivaria (2024). METR. https://vivaria.metr.org/.
[106] Influence and cyber operations: an update (2024). OpenAI. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.
[107] Alexopolous, Nikolaos et. Al. How Long Do Vulnerabilities Live in the Code? A Large-Scale Empirical Measurement Study on FOSS Vulnerability Lifetimes (2022). *Usenix.* https://www.usenix.org/system/files/sec22summer_alexopoulos.pdf.
[108] CISA Insights: Remediate Vulnerabilities for Internet-Accessible Systems (2024). *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/sites/default/files/publications/CISAInsights-Cyber-RemediateVulnerabilitiesforInternetAccessibleSystems_S508C.pdf.
[109] NIST AI 100-4: Reducing Risks Posed by Synthetic Content (2024). *National Institute of Standards and Technology.* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf.
[110] Lohn, Drew and Jackson, Krystal. Will AI Make Cyber Swords or Shields? (2022) Center for Security and Emerging Technologies. https://cset.georgetown.edu/publication/will-ai-make-cyber-swords-or-shields/.
[111] Kapoor, Sayesh et. al. On the Societal Impact of Open Foundation Models (2024). *arXiv.* https://arxiv.org/pdf/2403.07918.