NIST AI 800-1

Initial 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.ipd

NIST AI 800-1
Initial 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.ipd

July 2024

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

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

The U.S. AI Safety Institute (AISI) at NIST is releasing this document as an Initial Public Draft for public comment.

Comments on NIST AI 800-1 may be sent electronically to NISTAI800-1@nist.gov with "NIST AI 800-1, 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).

Table of 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 guidelines for improving the safety, security, and trustworthiness of dual-use foundation models (hereafter referred to as "foundation models")[i] consistent with the National AI Initiative Act[ii] and Executive Order 14110.[iii] Specifically, it focuses on managing the risk that such models will be deliberately misused to cause harm. The ways that foundation models can be misused continue to evolve, but they include the risks that models will facilitate the development of chemical, biological, radiological, or nuclear weapons; enable offensive cyber attacks; aid deception and obfuscation; and generate child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII) of real individuals.

The rapid development of foundation models poses significant challenges to understanding their capabilities and misuse risks,[1] and this document provides a basis to identify, measure, and reduce these risks across the AI lifecycle. Misuse risks are not a product of a model alone—they result in part from malicious actors' motivations, resources, and constraints, as well as society's defensive measures against that harm.[2] As a result, the guidelines provided here address both technical and social aspects of these risks.

Building on the AI Risk Management Framework,[3] this document identifies best practices to map, measure, manage, and govern misuse risks from foundation models, as well how organizations can provide transparency into how they are managing these risks. This document focuses particularly on foundation models' initial developers, but actors across the lifecycle of a model all play a role in managing misuse risks.[4]

---

[i] 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." This definition is provided in a glossary at the end of this report, along with definitions of other key terms.

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

[iii] 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".

2. SCOPE

Risks. This document focuses on misuse risk from dual-use foundation models. Consistent with Section 3(k) of Executive Order 14110,[5] 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 emerging misuse risks, such as a foundation model facilitating the development of a novel biological weapon, as well as current harms from misuse, such as a foundation model generating CSAM or NCII.

This document does not address other important risks from foundation models, such as bias, discrimination, and hallucination, nor does it address all risks to public safety, including those that may arise from other types of AI models and systems.[6] Actors across the AI lifecycle should manage these risks as well, consistent with relevant guidelines, such as those provided in the *Blueprint for an AI Bill of Rights*,[7] as well as the NIST AI Risk Management Framework and its Generative AI Profile.[8]

Actors. The practices in this document are principally focused on the central role that foundation models' initial developers have in the supply chain for their models.[9] These developers contribute most to determining how their models are made available, the models' capabilities, and safeguards against their misuse. In some cases, model developers may share influence over these factors with an external partner, such as a cloud service provider that leads deployment of the model, and in such cases these partners also have expanded opportunities and responsibilities to manage the risks that a model may be misused.[10]

Other parties also play important roles in managing misuse risks, but they are not the focus of this document. They include downstream developers and deployers, third-party evaluators and auditors, civil society organizations, and government agencies.[11] Relevant stakeholders throughout the AI supply chain are encouraged to share information and collaborate to understand and mitigate misuse risks, including to integrate appropriate risk mitigations into downstream systems that rely on foundation models.[12]

3. KEY CHALLENGES IN MAPPING AND MEASURING MISUSE RISKS

Accurately identifying and measuring foundation models' misuse risks helps to build confidence in which potential harms are realistic and which practices are justified to mitigate them. However, methodological and scientific challenges can limit understanding and measurement of these risks. Organizations should take appropriate steps to address these challenges and strive to build an empirical basis to evaluate and mitigate risks. These challenges include:

1. **Foundation models are broadly applicable.** Models trained on a broad data distribution can often be applied across many different domains, including domains not explicitly considered by the developer. This makes it challenging to anticipate potential ways in which a model might be misused and complicates measuring or monitoring for misuse risk broadly.
2. **Capabilities do not clearly translate across domains.** A foundation model's performance on one task may not provide reliable evidence of its performance on another, even when the two tasks appear related.[13] 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.
3. **It is difficult to predict how scale will affect performance.** In many instances, a foundation model's performance can be improved by increasing the amount or quality of its training data, the quantity of compute used to build the model, or the number of parameters of the model.[14] But while these factors may be useful heuristics in some instances, their precision is limited and their relationship to any given risk is uncertain.[15]
4. **The relationship between a measured capability and its potential risk of harm is often unclear**. It remains challenging to determine the likelihood or severity of real-world harm through isolated testing. For many harms to public safety, domain knowledge and computer automation are not the only the limiting factors in carrying out an attack, which may also require physical infrastructure, distribution mechanisms, or complex interactions in the physical world.[16] 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.[17]
5. **Methods to evaluate safeguards are nascent**. Foundation model developers implement safeguards to protect their models from misuse, such as those outlined in Appendix B, but there is a lack of effective techniques for evaluating the adequacy of those safeguards under real-world conditions.
6. **Evaluating risk may depend on 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. In addition, organizations developing foundation models may not currently have the professional experts or staff,

nor the developed means of accessing external expertise, necessary to appropriately assess some misuse risks.

7. **It is difficult to accurately emulate threat actors and misuse**. Assessing misuse risk requires realistically profiling threats. This assessment includes understanding how an actor might exploit a model's capabilities to cause harm, or how they might circumvent safeguards that have been deployed to protect a model from misuse. This practice can be challenging if malicious actors are willing to spend more time than evaluators or if they have access to infrastructure or niche expertise not available during evaluations. 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.

4. OBJECTIVES AND PRACTICES TO MANAGE MISUSE RISKS

This section outlines seven objectives, as well as associated practices that can help achieve them, for organizations to map, measure, manage, and govern the risk that their foundation models will be misused to deliberately harm public safety, consistent with the NIST AI Risk Management Framework. These objectives are:

1. Anticipate potential misuse risk
2. Establish plans for managing misuse risk
3. Manage the risks of model theft
4. Measure misuse risk
5. Ensure that misuse risk is managed before deploying foundation models
6. Collect and respond to information about misuse after deployment
7. Provide appropriate transparency about misuse risk

Organizations should consider all seven objectives holistically to manage foundation models' misuse risks. This document provides practices for each objective as non-exhaustive examples of how organizations can meet the objectives. Managing risk is an iterative process, and organizations should assess which practices are most relevant at each point in the lifecycle of a foundation model. Implementing these practices also cannot guarantee that actors will not misuse a foundation model, and organizations should adopt additional risk management measures where appropriate.

The specific elements of each objective below include:

1. **Objective:** an aspirational outcome that, when achieved, will help organizations manage a foundation model's misuse risks.
2. **Practice:** a suggested practice that can help organizations achieve an objective. This document proposes a particular set of practices that can be collectively used to achieve the indicated objectives. However, an individual practice may not be appropriate for all contexts, alternative practices may also be able to 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 be effective at achieving an objective.
4. **Documentation:** information that could be shared with either the public or with select parties to help demonstrate whether and how a practice was implemented, as well as potential evidence of its effectiveness and comprehensiveness. This documentation can also help to improve collaboration with third parties and to develop shared best practices for managing misuse risks. If an organization relies on alternative practices to achieve an objective, a similar level of documentation about those practices can be used to provide transparency, along with a rationale for why the alternative practices match or exceed the efficacy of those provided here.

Objective 1: Anticipate potential misuse risk

Assess the risks that a foundation model could be misused if it became available to malicious actors, including assessing these risks before the model is developed based on an estimate of its capabilities. Identify the most significant expected risks so that they can be measured and managed, as necessary.

Practice 1.1: Identify and maintain a list of threat profiles that covers significant ways in which malicious actors might misuse the model.

Recommendations to effectively implement Practice 1.1:

1. For each threat profile, specify the relevant capabilities of concern, a threat actor or actors who might use them to cause harm, and the malicious tasks that the threat actor might accomplish using the model.
2. Consider known misuse risks, including, as relevant, the risk that a model might be used to facilitate the development of chemical, biological, radiological, or nuclear weapons, to automate offensive cyber operations, or to generate CSAM or NCII.
3. Consider consulting external experts with relevant expertise and responsibilities to help identify gaps in threat profiles, including potentially granting them model access for open-ended experimentation to identify other ways that a model could be misused to cause harm. The consultation of external experts may be particularly relevant to helping ensure that the threat profiles cover the most significant expected ways in which a model could be misused to cause harm.

The following documentation can help provide transparency about how Practice 1.1 is implemented:

a. The list of key threat profiles.[18]
b. A description of how these threat profiles were selected and judged to have adequate coverage.
c. A description of any internal and external experts involved and their role in developing profiles.

Practice 1.2: Assess the impact of each identified threat profile if the malicious actor successfully misused the model.

Recommendations to effectively implement Practice 1.2:

1. For each threat profile, assess the impact to public safety if a threat actor successfully used the model to carry out the malicious task, including estimates of how much it would help an actor increase the scale, decrease the cost, or improve the effectiveness

of their malicious activity compared to alternative resources that are available to that actor, such as other machine learning models and digital tools.[19]

2. For each threat profile, estimate the scale, frequency, or probability with which threat actors might attempt to misuse the model.
3. Consider how potential harms may be prevented or managed outside of the context of the model itself, such as mechanisms that will effectively stop harm from a given malicious activity regardless of how inexpensive or easy the model makes that malicious activity.

The following documentation can help provide transparency about how Practice 1.2 is implemented:

a. An impact assessment for each identified threat profile.

Practice 1.3: Estimate the model's capabilities of concern before it is developed by comparing it to existing models.

Recommendations to effectively implement Practice 1.3:

1. Identify similar models for which capabilities of concern have already been measured.
2. Based on comparing the characteristics of the upcoming model with existing models, estimate how its performance will compare to those models.
3. Assess the degree of uncertainty in these estimates based on how different the models are, how those differences are expected to affect their capabilities, and the reliability and completeness of the evaluations available for the similar models.
4. If capability forecasts are uncertain and include the possibility that the model may pose significant risk, consider development strategies that may reduce reliance on forecasting, such as taking steps to periodically measure capabilities throughout the development process.

The following documentation can help provide transparency about how Practice 1.3 is implemented:

a. Capability estimates that were made prior to developing the model.
b. Relevant capability measurements performed after the model was developed.
c. Estimated comparisons of the model's misuse risks compared to common benchmarks, where available.

Objective 2: Establish plans for managing misuse risk

Determine acceptable levels of misuse risk, considering legal or regulatory obligations, potential benefits that may trade off against risk, and other factors. Align development and deployment plans with the resources, time, and operational constraints that may be required to manage potential misuse risk.

Practice 2.1. Identify a level of misuse risk which the organization considers unacceptable.

Recommendations to effectively implement Practice 2.1:

1. Define a risk threshold for each identified threat profile based on the impact assessments carried out in Practice 1.2 and other relevant information.[20]
2. Account for how much the model may reduce the cost of accomplishing the malicious task, increase the number of actors who could accomplish that task, or increase the scale or quality at which a given actor can accomplish the task.
3. When determining risk thresholds, weigh potential benefits against the misuse risks.
4. Refine thresholds continuously based on adjustments to identified threat profiles or changes in factors that affect risk tolerance (such as increases in expected benefits).
5. Consider how external factors such as legal and regulatory obligations may inform what risk thresholds are acceptable.

The following documentation can help provide transparency about how Practice 2.1 is implemented:

a. The thresholds for the misuse risks of a model that the organization considers unacceptable.

Practice 2.2: Establish a roadmap to manage misuse risks for the development of planned foundation models and future versions.

Recommendations to effectively implement Practice 2.2:

1. Define how the organization expects to manage misuse risk and achieve the objectives established in this document.
2. Plan to implement security practices to protect models from model theft if necessary to manage misuse risk. Define security goals and a timeline for achieving those goals.
3. Plan to implement appropriate safeguards to protect deployed models from misuse if necessary to manage misuse risk. Define safety goals and a timeline for achieving those goals.
4. Plan to adjust deployment or development strategies if misuse risks rise to unacceptable levels before adequate security and safeguards are available to manage risk.

The following documentation can help provide transparency about how Practice 2.2 is implemented:

a. A roadmap documenting what risk mitigations are expected to be feasible over time given technical and resource constraints, and what steps will be taken if measurements suggest those mitigations are not adequate to manage misuse risk.

Objective 3: Manage the risks of model theft

Take steps to prevent the theft of information and assets that would allow a malicious actor to recreate the foundation model, where doing so is necessary to manage misuse risk. Only develop a model that relies on confidentiality to manage misuse risk when the risk of model theft is sufficiently mitigated.

Practice 3.1: Assess the risk of model theft from relevant threat actors.

Recommendations to effectively implement Practice 3.1:

1. Assess the risk of model theft by each identified actor or threat profile, based on the relevant actors' respective resources, motivations, tactics, and sophistication.[21]
2. Consider the organization's compliance with applicable cybersecurity best practices as an input into assessing the risk of model theft.
3. Consider the risk posed by an insider threat, such as an individual involved in developing or deploying the model who may behave maliciously or collaborate with an external attacker.
4. Consider using cybersecurity red teams and penetration testing to assess how difficult it would be for an actor to circumvent security measures.

The following documentation can help provide transparency about how Practice 3.1 is implemented:

a. A summary of evidence collected to evaluate the risk of model theft, including the results of any red team exercises and penetration testing.

Practice 3.2: Compare predicted misuse risk to the organization's risk tolerance prior to developing models with increased capabilities of concern.

Recommendations to effectively implement Practice 3.2:

1. Prior to developing a model, assess overall misuse risk from model theft by combining the estimated capabilities of concern with an estimate of the probability of model theft.
2. Adjust or halt further development until the risk of model theft is adequately managed and within the organization's defined risk thresholds.

3. Periodically revisit estimates of misuse risk stemming from model theft, including prior to individual development decisions that may significantly increase capabilities of concern.

The following documentation can help provide transparency about how Practice 3.2 is implemented:

a. The risk assessments that were performed, either periodically or prior to significant development and deployment decisions, related to model theft.

Practice 3.3: Maintain security practices sufficient to prevent model theft.

Recommendations to effectively implement Practice 3.3:

1. Implement security practices designed to protect against model theft.
2. Where applicable, apply security practices specifically tailored to the context of foundation models, such as protections against exfiltrating large amounts of data (e.g., model weights) and protections against vulnerabilities like extraction attacks.[22]
3. Increase the level of security proportionate to the level of misuse risk posed by the model and the importance of preventing model theft to managing that risk.
4. Apply appropriate protections against insider threats, such as limiting access to model weights within the organization.
5. Consider adopting existing security standards for sensitive data and applications, such as U.S. government cybersecurity guidance and applicable international standards.
6. Re-assess the risk of model theft prior to development with new security practices in place (if previous assessments found unacceptable risks from model theft).

The following documentation can help provide transparency about how Practice 3.3 is implemented:

a. A summary of security measures that have been implemented to reduce the risk of model theft.

Objective 4: Measure the risk of misuse

Where there is a reasonable assessment that a foundation model could be misused, measure the predicted risks to provide evidence for the model's actual misuse risk in a real-world context. Rely on methods that incorporate both technical and social factors, as well as those that provide accurate evidence and higher confidence, while avoiding any harm that could be caused by measuring dangerous activities.

Practice 4.1: Measure model capabilities relevant to assessing misuse risk.

Recommendations to effectively implement Practice 4.1:

1. Directly measure capabilities of concern or identify alternative capability measurements that allow for sufficiently reliable inferences about them, such as pursuing recommendations 2, 3, and 4 below when they are sufficiently reliable.
2. Consider comparing the performance of a new model to an existing model with known misuse risks (the "proxy model"). If the models are similar across a broad range of evaluations, results from the proxy model's capability evaluations may be suitable to assess risk from the new model. If this comparison is inconclusive or suggests that the new model may pose significant risk, then the new model's capabilities of concern should be separately evaluated.
3. Consider measuring model performance on proxy tasks that are safe and tractable while being similar enough to allow reliable inferences about the capability of concern.
4. If the most accurate capability measurement would be expensive and it is appropriate to do so, consider using cheaper tests (such as automated evaluations on simpler tasks that would be easier for the model to perform than the true capabilities of concern)[23] that may indicate that a model lacks dangerous capabilities. If these tests are inconclusive or indicate a potential risk, then rely on more costly and precise measurements.
5. Account for the possibility that the model's capabilities of concern may exceed any evaluated proxy capabilities and adjust the interpretation of evaluation results accordingly. This consideration is especially relevant when the proxy capability being evaluated differs significantly from the actual capability of concern.
6. Assess what a threat actor could achieve given access to the weights of a model and the ability to integrate it with other tools by testing AI systems which are reasonably optimized for performance on the evaluation task. If there is a gap between the effort applied to optimize system performance during testing and the effort that could be applied by a threat actor, evaluations should explicitly account for the uncertainty introduced by that gap.
7. Avoid overlap between data used to train the model and data used in capability and risk evaluations and measure the extent of any overlap.

The following documentation can help provide transparency about how Practice 4.1 is implemented:

a. A list of evaluation tasks that were used to evaluate each threat profile, and a representative subset of datasets used for each evaluation.
b. A methodological description for each evaluation in enough detail to reproduce it.
c. An analysis of the relationship between evaluation tasks and capabilities of concern, addressing the possibility that evaluation tasks are more challenging.
d. If a model was evaluated via a comparison with a proxy model: the proxy model used for comparison, evaluations used to establish comparability, the risk assessment for the

proxy model, and other ways in which the proxy may be meaningfully different from the model being assessed.

Practice 4.2: Use red teams to assess whether threat actors could bypass model and system safeguards and misuse any capabilities of concern.

Recommendations to effectively implement Practice 4.2:

1. To justify an assessment that a threat actor is unlikely to be able to misuse a model, verify that an adequately resourced red team is unable to misuse the model or accomplish a related proxy task in a realistic deployment context. Apply practices 4.1.1, 4.1.3, and 4.1.4 to the selection of proxy tasks for red team exercises as well as capability evaluations.
2. Clearly specify what goal the red team is trying to achieve in advance, provide incentives and accountability for achieving those goals, and select red teams based on their ability to achieve those goals.
3. Rely on red teams comprised of external experts that are meaningfully independent from the model developer and who do not have incentives that conflict with their red-teaming goal.
4. Compare the red team's expertise, resources, and time available to those of a relevant threat actor. To the extent that there are gaps, explicitly account for those gaps when interpreting the result of a red team exercise.
5. Consider providing additional resources to the red team or making the red team's task easier in other ways (such as providing additional access,[24] reducing task difficulty, or dividing complex tasks into constituent pieces) to compensate for any remaining gaps between the red team and threat actors.
6. Consider providing red teams with available legal protections for their tasks, such as waiving terms of service and indemnifying them for legal liability for their interactions with the model.
7. Provide the red team with at least as much information about the system as would be available to an attacker and make explicit any respects in which the red team lacks full information about the design of system safeguards.
8. 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.
9. Explicitly define the time period over which the red teaming results are intended to apply and consider the fact that a threat actor will have access to any information about how to use the model or circumvent safeguards that becomes public over that time period.

The following documentation can help provide transparency about how Practice 4.2 is implemented:

a. A description of the red team's goal.
b. A description of the red team's composition, expertise, resources, and timelines.
c. The results of the red team exercise given varying levels of access to the model.
d. A summary of the strategies pursued by the red team.

Objective 5: Ensure that misuse risk is managed before deploying foundation models

Take actions to increase access to the model (e.g., deploying a model via API or releasing its weights) only when misuse risks are adequately managed, including that they are at minimum within the organization's risk tolerance.

Practice 5.1: Assess the effect of a potential deployment on the model's misuse risk.

Recommendations to effectively implement Practice 5.1:

1. Consider potential deployments and the levels of access they would grant actors (e.g. whether they would increase the number or type of actors who can access a model or whether they would grant actors greater access to a model's features).
2. Identify the level of access that a malicious actor could obtain under each proposed deployment (e.g., would it grant them access to API inference, access to a fine-tuning API, access to model weights) and consider how the deployment may affect misuse risk. For example, allowing fine-tuning via API can significantly limit options to prevent jailbreaking and sharing the model's weights can significantly limit options to monitor for misuse (Practice 6.1) and respond to instances of misuse (Practice 6.2).

The following documentation can help provide transparency about how Practice 5.1 is implemented:

a. For each deployment: the level of access provided, the assessed misuse risks associated with the deployment, and the set of safeguards anticipated to mitigate these misuse risks.

Practice 5.2: Implement safeguards proportionate to the model's misuse risk.

Recommendations to effectively implement Practice 5.2:

1. Implement safeguards designed to protect the model from misuse. Appendix A outlines a non-exhaustive list of possible safeguards.
2. Establish reliable evidence of safeguards' effectiveness before relying on them to prevent misuse of meaningful capabilities of concern.

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 with the additional safeguards in place.

The following documentation can help provide transparency about how Practice 5.2 is implemented:

a. The list of safeguards that have been implemented.

b. A summary of the result of safeguard evaluations via red teams and other testing.

Practice 5.3: Only pursue deployments where misuse risk is adequately managed.

Recommendations to effectively implement Practice 5.3:

1. For each deployment, establish a process to determine whether the deployment should proceed based on the assessed misuse risk and a consideration of any safeguards. Otherwise, determine whether the deployment should be modified, delayed, or canceled. For instance, consider whether further safety improvements are feasible 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.
2. Consider leaving a margin of safety between the estimated level of risk at the point of deployment and the organization's risk tolerance. Consider how threat actors may continue to acquire new knowledge about how to misuse or augment the model after it is deployed[25] and how to circumvent its safeguards.[26] Consider a larger margin of safety to manage risks that are more severe or less certain.

The following documentation can help provide transparency about how Practice 5.3 is implemented:

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 thresholds.

Objective 6: Collect and respond to information about misuse after deployment

Collect information about deployed systems that improves understanding of their misuse risk to adjust deployments and improve future risk management. Engage with and encourage findings from the public, relevant civil society organizations, external researchers, and the foundation model's third-party distribution partners.

Practice 6.1: Where possible, monitor distribution channels for evidence of misuse.

Recommendations to effectively implement Practice 6.1:

1. Monitor APIs, websites, and other distribution channels for misuse while maintaining privacy of users.
2. Build or procure systems to enable automated detection of misuse.
3. Continually assess the effectiveness of systems that are intended to detect misuse to help provide ground truth for their effectiveness and evidence that misuse is not going undetected. Red teaming can be particularly helpful to assess whether malicious actors may be able to deliberately avoid detection.
4. Request that third-party distribution channels for the model monitor those channels for misuse and that they regularly share information with the developer regarding their monitoring.
5. Consider using tiered methods of detection when doing so helps prioritize limited resources, improve privacy, and increase coverage. This can include filtering for misuse first by less costly automated methods which are then, as appropriate, validated by methods requiring direct human intervention.
6. When malicious actors can avoid direct monitoring by operating the model independently, such as when its weights are widely available, consider mechanisms other than those identified in recommendations 6.1.1, 6.1.2, and 6.1.3 that could be used to monitor for patterns of malicious behavior enabled by the model.

The following documentation can help provide transparency about how Practice 6.1 is implemented:

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.[27]

Practice 6.2: Maintain a process to respond to incidents of model misuse.

Recommendations to effectively implement Practice 6.2:

1. Establish clear organizational responsibilities for responding to incidents of misuse.
2. Preemptively develop plans for plausible novel scenarios of misuse and how to respond, such as restricting access to the model or strengthening safeguards.[28]
3. When it may not be possible to effectively reduce access to a model, such as when the model's weights are widely available, plan for how identified instances of misuse will inform future development and deployment decisions.
4. Consider carrying out drills to practice responding to time-sensitive and safety-critical scenarios of misuse.

The following documentation can help provide transparency about how Practice 6.2 is implemented:

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

Practice 6.3: Establish protections for internal reporting of misuse issues.

Recommendations to effectively implement 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 raised by employees and contractors in a timely fashion.

The following documentation can help provide transparency about how Practice 6.3 is implemented:

a. A summary of the organization's policies with respect to internal safety reporting.

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

Recommendations to effectively implement Practice 6.4:

1. Publish a clear vulnerability disclosure policy for model safety issues that outlines how such vulnerabilities should be shared with the developer and the public, and how the organization will respond to reported vulnerabilities.
2. Publish a 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.[29]
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.

The following documentation can help provide transparency about how Practice 6.4 is implemented:

a. A summary of the organization's commitment to not pursue legal action against third-party researchers acting in good faith.
b. A description of the organization's process for providing vetted researchers with access to models or systems with fewer safeguards.[30]

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

Recommendations to effectively implement Practice 6.5:

1. Establish a program to incentivize researchers for finding vulnerabilities and disclosing them according to the vulnerability disclosure policy.[31]

The following documentation can help provide transparency about how Practice 6.5 is implemented:

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

Objective 7: Provide appropriate transparency about misuse risk

Provide transparency to the public and relevant entities regarding the organization's processes for developing and deploying foundation models that are related to misuse risk to facilitate understanding, accountability, collaboration, and scientific development related to model misuse.

Practice 7.1: Publish regular transparency reports that include key details regarding misuse risks and how those risks are managed.

Recommendations to effectively implement Practice 7.1:

1. Share the methodology and results of pre-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.[32]
2. Share details regarding the safeguards in place for the model, including how they are applied across different distribution channels, with as much as can be shared without rendering the safeguards ineffective.[33]
3. Share information about data used to build the model that is relevant to assessing the misuse risk, such as data sources and criteria for data filtering and selection.[34]
4. Share steps that downstream developers and deployers of AI systems that integrate the foundation model should take to manage misuse risk.
5. Share details regarding the organization's internal structure and internal governance process with respect to determining how to deploy foundation models and map, measure, and manage misuse risk.
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.[35]

Practice 7.2: Disclose information about risk management practices to promote accountability.

Recommendations to effectively implement Practice 7.2:

1. Regularly share information about risk management practices through a process structured to provide meaningful accountability.
2. 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 listed here.
3. Ensure recipients of information are effectively independent from the development and deployment process and that they are empowered to provide accountability (such as by disclosing risk management deficiencies publicly).

The following documentation can help provide transparency about how Practice 7.2 is implemented:

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 incidents and hazards related to the foundation model to AI incident databases.

Recommendations to effectively implement Practice 7.3:

1. Based on existing best practices and adequate review of the benefits and risks of disclosing certain information, define the category of misuse events to report.[36]
2. Collate verified reports of misuse in a commonly used and machine-readable format.
3. Share verified reports of misuse with relevant third parties, such as AI incident databases.[37]

The following documentation can help provide transparency about how Practice 7.3 is implemented:

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.

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.[38]

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. Artificial Intelligence 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.[39]

Capability of Concern
A capability of an AI model that poses a risk of enabling malicious behavior that would cause serious harm to public safety.

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, or access to a fine-tuned or otherwise augmented version of the model from a third-party deployer.

Dual-Use Foundation Model
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.[40]

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.

Foundation Model
An AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; and is applicable across a wide range of contexts.[41]

Jailbreaking
Methods that cause an AI model to act in a manner contrary to how its designer intended for it to behave.

Misuse Risk
A risk that an AI model will be deliberately misused to cause harm.

Model Theft
Unauthorized access to information that substantially aids an actor in recreating an AI model or its capabilities of concern, 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.

Appendix B. Example Safeguards Against the Misuse of Foundation Models

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

Table 1. Example Safeguards.

Safeguard	*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.[42] • Employ training techniques that reduce the model's propensity to produce harmful output such as by reducing its knowledge of harmful information, reducing its capabilities of concern, or causing the model to refuse harmful requests.[43] • Train models with approaches that make it more difficult for subsequent fine-tuning to remove safeguards.[44]
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.[45] • 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.
Limit access to the model's capabilities.	• Monitor the model's risk for misuse with a limited audience and gradually expand access to the model to wider audiences over time.[46] • Limit the ability to interact with the model to contexts and users where the misuse risks are lower.[47] • 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 significant misuse risk while it is in production. • Apply limitations on using the model's capabilities broadly to all interactions with the model or target limitations to particular capabilities and features at particular misuse risks, such as capabilities of concern that also have narrow beneficial purposes. • For models that pose more severe risk, consider restricting the model only to those 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.
Ensure the level of access to the model's weights is appropriate.[48]	• Consider when it is appropriate to make the model's weights widely available, such as available for download by the public. Once a model's weights are made widely available, options to roll back or prevent its further sharing and modification are severely limited.[49] • Consider when it is appropriate to allow the foundation 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.[50] • Consider limiting internal access to the models' weights within an organization.
Stop development if a model displays significant misuse risk.	• If other practically available safeguards are not sufficient to protect a model from misuse–including considering the risk of theft or internal abuse–then it may be appropriate to make significant changes to the development plan, or else not develop the model at all.

1

[1] 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.

[2] 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.

[3] AI Risk Management Framework. *National Institute of Standards and Technology*. https://www.nist.gov/itl/ai-risk-management-framework.

[4] 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/.

[5] 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/.

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

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

[8] 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.

[9] Bommasani, R. et al., (2021) On the Opportunities and Risks of Foundation Models. *arXiv*. https://arxiv.org/abs/2108.07258.

[10] 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

[11] Cen, S. et al., (2024) AI Supply Chains. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4789403; and 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.

[12] Mulani, N. et al., (2023) Proposing a Foundation Model Information-Sharing Regime for the UK. *Centre for the Governance of AI.* https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk; and Bommasani, R. et al. (2023) The Foundation Model Transparency Index. *arXiv*. https://arxiv.org/abs/2310.12941.

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

[14] 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.

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

[16] Terwilliger 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.

[17], 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.

[18] Anthropic (2023) Responsible Scaling Policy. *Anthropic*. https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf; OpenAI (2023) Preparedness Framework (Beta). *OpenAI*. https://cdn.openai.com/openai-preparedness-framework-beta.pdf; DeepMind (2024) Frontier Safety Framework. *DeepMind*. https://storage.googleapis.com/deepmind-

media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf. Magic (2024) AGI Readiness Policy. *Magic*. https://magic.dev/agi-readiness-policy.

[19] 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.

[20] Koessler, L. et al., (2024) Risk thresholds for frontier AI. *arXiv*. https://arxiv.org/pdf/2406.14713.

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

[22] Carlini, N. et al., (2024) Stealing Part of a Production Language Model. *arXiv*, https://arxiv.org/pdf/2403.06634.

[23] Barrett, A. et al, (2024) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *Center for Long-Term Cybersecurity, UC Berkeley*. https://cltc.berkeley.edu/wp-content/uploads/2024/05/Dual-Use-Benchmark-Early-Red-Team-Often.pdf; Shao, M. et al., (2024) NYU CTF Dataset: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security. *arXiv*. https://arxiv.org/pdf/2406.05590v1; and Li, N. et al., (2024) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *arXiv*. https://arxiv.org/pdf/2403.03218.

[24] Casper, S. et al., (2024) Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv*. https://arxiv.org/pdf/2401.14446.

[25] 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; and Measuring the impact of post-training enhancements. *METR*. https://metr.github.io/autonomy-evals-guide/elicitation-gap/.

[26] 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.

[27] Overview of Meta AI safety policies prepared for the UK AI Safety Summit. *Meta*. https://transparency.meta.com/en-gb/policies/ai-safety-policies-for-safety-summit; Microsoft's AI Safety Policies. *Microsoft*. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1dObQ; AI Safety Summit: An update on our approach to safety and responsibility. *Google DeepMind*. https://deepmind.google/public-policy/ai-summit-policies/; AI Safety Summit – Enhancing Frontier AI Safety. *AWS Amazon*. https://aws.amazon.com/uki/cloud-services/uk-gov-ai-safety-summit/; Our policy on frontier safety. *Inflection*. https://inflection.ai/frontier-safety; and *OpenAI*. https://openai.com/index/openai-safety-update/.

[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] Longpre, S. et al., (2024) A Safe Harbor for AI Evaluation and Red-Teaming. *arXiv*. https://arxiv.org/pdf/2403.04893.

[30] 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.

[31] 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.

[32] 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; and OpenAI (2023) GPT-4 System Card. *OpenAI*. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[33] Bommasani, R. et al., (2024) The Foundation Model Transparency Index v1.1: May 2024. *arXiv*. https://arxiv.org/pdf/2407.12929.

[34] 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.

[35] 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.
[36] McGregor, S. (2020) Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *arXiv*. https://arxiv.org/pdf/2011.08512.
[37] 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.
[38] The term "artificial intelligence" or "AI" has the meaning set forth in 15 U.S.C. 9401(3).
[39] Executive Order 14110.
[40] Executive Order 14110.
[41] Executive Order 14110.
[42] Thiel, D., et al., (2023) Identifying and Eliminating CSAM in Generative ML Training Data and Models. *Stanford Digital Repository*. https://purl.stanford.edu/kh752sm9123.
[43] Lynch, A. et al., (2024) Eight Methods to Evaluate Robust Unlearning in LLMs. *arXiv*. https://arxiv.org/pdf/2402.16835.
[44] 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.
[45] Bommasani, R. (2023) Ecosystem graphs: The social footprint of foundation models. *arXiv*, abs/2303.15772.
[46] Solaiman, I. (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv*. https://arxiv.org/pdf/2302.04844.
[47] Seger, E. et al., (2024) Open-Sourcing Highly Capable Foundation Models. https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf
[48] When a model's weights are available to a threat actor, a range of other safeguards become less effective, particularly those that are implemented at the application level, such as limiting who can use the model and detecting when it is misused. Limiting access to the model weights should be weighed against the potential benefits of access, such as for innovation and research, including research into safety. Limiting access to model weights is also only effective if organizations can prevent model theft.
[49] 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.
[50] 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.