Managing Misuse Risk for Dual-Use Foundation Models: Comments on the Initial Public Draft of NIST AI 800-1

Jonas Schuett,[1] **Sophie Williams**,[1] **Marie Buhl**,[1] **Alan Chan**,[1] **Markus Anderljung**[1,2]

[1] *Centre for the Governance of AI (GovAI)*, [2] *Center for a New American Security (CNAS)*

We welcome the opportunity to comment on the initial public draft of NIST AI 800-1. We offer the following submission for your consideration and look forward to future opportunities to provide additional input. Earlier this year, we already submitted comments on NIST's Draft Profile on Generative AI and NIST's Assignments under the Executive Order concerning AI.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK. It was founded in 2018 at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI. GovAI is part of the Artificial Intelligence Safety Institute Consortium (AISIC).

About the authors

Jonas Schuett is a Senior Research Fellow at GovAI. He is a member of AISC Working Group 5 (Task Force 5.1) and the OECD Expert Group on AI Risk & Accountability. His research focuses on the governance of dual-use foundation models, with a special focus on risk management. Before joining GovAI, he advised the UK Government on AI regulation and was part of Google DeepMind's Public Policy Team. He has a background in law.

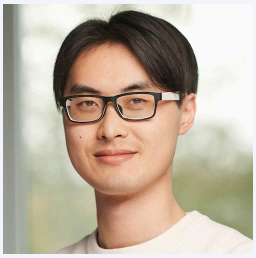
Email • LinkedIn • Twitter • Google Scholar

Sophie Williams is a Research Scholar at GovAI. Her research focuses on the development of policies and regulatory frameworks for AI. Before joining GovAI, she developed AI policy at the UK Home Office and UK Department for Science, Innovation and Technology. She has also worked on competition and economic policy at the UK's Financial Conduct Authority.

Email • LinkedIn

Marie Buhl is a Research Scholar at GovAI. Her research focuses on risk management, deployment policies, and frontier AI regulation. Her current work considers how safety cases could be used in AI governance. She holds a BA in politics, philosophy and economics from the University of Oxford.

Email • LinkedIn



Alan Chan is a Research Scholar at GovAI. His research focuses on governing AI agents. He has a background in mathematics and machine learning, and is also a final-year PhD student at Mila (Quebec AI Institute).

Email • LinkedIn • Twitter • Google Scholar

Markus Anderljung is Director of Policy and Research at GovAI, an Adjunct Fellow at the Center for a New American Security (CNAS), a member of the OECD Expert Group on AI Futures, and an AI Policy Expert Advisor to the UK Department for Science, Innovation and Technology. His research focuses on the regulation and governance of frontier AI systems and dual-use foundation models. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist.

Email • LinkedIn • Twitter • Google Scholar

***Note**: The views expressed in this submission are those of the authors and do not represent the views of GovAI.*

Executive summary

We welcome NIST's effort to provide guidance on managing misuse risk for dual-use foundation models. The initial public draft of NIST AI 800-1 is a commendable first step. In this submission, we suggest further improvements.

- **Proportionality.** NIST AI 800-1 should specify what constitutes a dual-use foundation model. Relatedly, it should specify the conditions under which it is particularly important to adhere to the guidance.

- **Risk thresholds.** The decision to deploy a dual-use foundation model should not be determined by comparing risk estimates with risk thresholds (i.e. thresholds defined in terms of the probability and severity of harm). Instead, NIST AI 800-1 should talk about thresholds more broadly, which might also include capabilities thresholds (i.e. thresholds defined in terms of model capabilities and adequate mitigations).

- **Safety frameworks.** Developers of the most capable foundation models should implement an AI safety framework as specified in the *Frontier AI Safety Commitments*. AI safety frameworks are risk management policies intended to keep the potential risks associated with developing and deploying dual-use foundation models to an acceptable level.

- **Safety case.** Developers of the most capable foundation models should prepare a safety case ahead of deployment. Safety cases are structured arguments, supported by evidence, that a system is sufficiently safe in a specific deployment context.

- **AI agents.** NIST AI 800-1 should recommend how post-deployment monitoring practices should account for AI agents. AI agents are AI systems that can pursue complex goals, with little to no explicit human instruction for how to do so

Table 1 provides a breakdown of our recommendations, which we set out in detail throughout the remainder of the document. For some of our recommendations, though not for all, we suggest sentence-level suggestions in Appendix A.

1. International harmonization	Frontier AI Safety Commitments	Developers of dual-use foundation models should implement an AI safety framework as specified in the *Frontier AI Safety Commitments*
		Recommend some additional practices included in the *Frontier AI Safety Commitments*.
	Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems	Consider recommending some additional actions set out in the G7 Code of Conduct.
	Emerging Processes for Frontier AI Safety	Consider more closely aligning terminology with the *Emerging Processes for Frontier AI Safety*.
		Recommend some additional practices set out in the *Emerging Processes for Frontier AI Safety*.
2. Proportionality of practices	Clarify what models are in scope	Though specifying this definition further is a considerable task, we expect that NIST is the best-placed institution in the US executive branch to do so.

		The guidance should leave some ambiguity in its definition.
		Specify *inclusions*, i.e. systems that are likely to fall under the definition.
		Specify more *exclusions*, i.e. systems that are unlikely to fall under the definition.
		Consider expanding the scope beyond dual-use foundation models.
	Differential treatment of different actors	Developers of the most capable models should comply with more extensive practices.
		Inversely, developers of less capable models may adopt light-touch versions of the practices.
		Developers of open foundation models should be treated differently.
	Prioritizing between different practices	Prioritize practices to inform developers' decisions about where to dedicate resources.
3. Improvements to proposed practices	Risk estimates	Add recommendations on how these estimates should be conducted.
	Validating model evaluations and risk estimates	Recommend that developers validate their model evaluations and risk estimates
	Risk thresholds	Define terms related to AI risk thresholds and use them consistently.
		Build the recommendations around "thresholds," rather than "risk thresholds".
		Adjust recommendations on how to *set* risk thresholds.
		Recommend that assessments are made as to how close thresholds are to being breached
		Adjust recommendations on how to *use* risk thresholds.
		Add recommendations on how to *use* capabilities thresholds.
		Add recommendations on how to *set* capabilities thresholds.
	Post-deployment monitoring	Add recommendations for how post-deployment monitoring practices should account for AI agents.
	Transparency requirements	Expand the recommendations for publishing regular transparency reports.
		Recommend that companies report certain information directly to the government.
		Expand the recommendations for reporting incidents and hazards.
4. Suggesting new practices	Safety frameworks	Developers of the most capable foundation models should implement a safety framework.
	Safety cases	Developers of the most capable foundation models should prepare a safety case ahead of deployment.
5. Role of other actors in the supply chain	Downstream developers	Developers of foundation models should take additional steps to manage the risk from downstream development.
		Downstream developers should consider adopting the recommendations in NIST AI 800-1 where appropriate.
	Compute providers	Developers of foundation models may benefit from choosing responsible compute providers.

Table 1: Overview of our recommendations

Contents

1. International harmonization

We think that NIST AI 800-1 should be aligned, where appropriate, with other initiatives, such as the Frontier AI Safety Commitments, the International Code of Conduct for Organizations Developing Advanced AI Systems, and the policy paper Emerging Processes for Frontier AI Safety. Such alignment could be a matter of ensuring compatibility with best practice and existing commitments, which may reduce compliance burdens and increase adherence to the guidelines. Alignment could also help identify measures already undertaken or committed to by companies that would support the goals of NIST AI 800-1.

Frontier AI Safety Commitments

At the AI Seoul Summit 2024, leading AI companies—including those that are most likely to develop dual-use foundation models—signed the *Frontier AI Safety Commitments* (DSIT, 2024), in which they commit to take a number of safety measures ahead of the 2025 AI Action Summit in France. We think that NIST AI 800-1 should be compatible with these commitments, to the extent that implementing the recommendations would support adherence with the commitments that are relevant to misuse risks.

- **Developers of the most capable foundation models should implement a safety framework as specified in the *Frontier AI Safety Commitments*.** Safety frameworks are risk management policies intended to keep the potential risks associated with developing and deploying dual-use foundation models to an acceptable level. NIST AI 800-1 should include specific reference to safety frameworks. For more detail on this recommendation, see AI safety frameworks below.

- **Recommend some additional practices included in the *Frontier AI Safety Commitments*.** One example would be to recommend that developers explain how external actors, including the public, "are involved in the process of assessing the risks of their AI models and systems, the adequacy of their safety framework [...], and their adherence to that framework" (DSIT, 2024). Doing so would help external stakeholders more accurately judge the trustworthiness of companies' statements about their models and safety practices.

Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems

At the Hiroshima Summit 2023, the G7 agreed to an *International Code of Conduct for Organizations Developing Advanced AI Systems* (G7, 2023b), which is based on the *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems* (G7, 2023a) and the *OECD AI Principles* (OECD, 2024a). We think that NIST AI 800-1 is already broadly aligned with the Code of Conduct, but including some additional actions could be beneficial.

- **Consider recommending some additional actions set out in the G7 Code of Conduct.** For example, NIST AI 800-1 could specifically recommend that developers implement "physical" security controls, as part of the recommended practices to manage security risks, in addition to cybersecurity and insider threat safeguards. Doing so will help ensure that developers consider the full range of security controls available, as well as how these different types of controls complement one another.

Emerging Processes for Frontier AI Safety

Ahead of the 2023 AI Safety Summit in Bletchley Park, the UK's Department for Science, Innovation and Technology (DSIT) published a policy paper on *Emerging Processes for Frontier AI Safety* (DSIT, 2023) which set out nine practices for the safe development and deployment of frontier AI models. NIST AI 800-1 would benefit from drawing on aspects of this paper, as specified below.

- **Consider more closely aligning terminology with the *Emerging Processes for Frontier AI Safety*.** For example, both documents talk about certain model capabilities. While DSIT uses the term "dangerous capabilities", NIST AI 800-1 mostly refers to them as "capabilities of concern", but it occasionally also uses the term "dangerous capabilities". "Dangerous capabilities" seems to be a more commonly used term in AI safety frameworks (Anthropic, 2023; OpenAI, 2023) and in the literature (Shevlane et al., 2023; Phuong et al., 2024), so it may be appropriate for NIST AI 800-1 to instead use this term consistently.

- **Recommend some additional practices set out in the *Emerging Processes for Frontier AI Safety*.** For example, NIST AI 800-1 could recommend that developers establish policies that set out what information they will share externally and with whom (e.g. governments, the public etc.), as well as what information they will not share. As DSIT explains, these policies could also specify when information-sharing would be subject to a risk assessment, along with guidelines for carrying out an assessment. Furthermore, there are some additional practices that NIST AI 800-1 may want to include in Table 1 (Example Safeguards Against the Misuse of Foundation Models), such as "data auditing", given its potential to help developers identify potentially harmful data in training datasets.

2. Proportionality of practices

It is important that NIST AI 800-1 mitigates misuse risk without unnecessarily stifling innovation. The practices should therefore be proportionate to the risks involved. To this end, we suggest clarifying what models are in scope and to make changes that allow for differential treatment of different actors and prioritization between different practices.

Clarify what models are in scope

Developers may find it difficult to determine whether their models fall under NIST's definition of "dual-use foundation models", which is drawn from *Executive Order 14110* (White House, 2023).[1] While it is clear what constitutes a model trained using self-supervised learning containing tens of billions of parameters, we estimate that at least 180 such models have been publicly released to date, with many more to come.[2] Most importantly, it may not be clear what constitutes "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". Although we think that few, if any, present serious risks today, others may disagree and companies may end up taking an overly cautious approach.

---

[1] *Executive Order 14110* defines "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."

[2] We estimate that ~120 models if narrowly applicable models, such as those trained solely on biological sequence data, are excluded. Estimates using data from (Epoch AI, 2024).

- **Though specifying this definition further is a considerable task, we expect that NIST is the best-placed institution in the US executive branch to do so.** They could make some progress on this point in this iteration of the guidance, though we expect future iterations and dedicated workstreams will be required.

- **The guidance should leave some ambiguity in its definition.** Such ambiguity is necessary as it reflects our uncertainty of what models pose serious risks. Nonetheless, the guidance can specify inclusion and exclusion criteria, thereby providing more certainty for at least some actors.

- **Specify *inclusions*, i.e. systems that are likely to fall under the definition.** For example, this might include any model that is trained using more than 10^{26} operations. It could also include any model that exceeds certain predefined thresholds of misuse-relevant capabilities. Though developing such thresholds may have to be left for future iterations of the guidance.

- **Specify more *exclusions*, i.e. systems that are unlikely to fall under the definition.** For example, NIST AI 800-1 could specify that image-only models are not "applicable across a wide range of contexts".

- **Consider expanding the scope beyond dual-use foundation models.** There are certain AI models that present significant dual-use potential and where the guidance may be applicable, but may not be appropriately termed foundation models or otherwise fit the definition from *Executive Order 14110*. Notably, biological design tools (BDTs) may fit into this category (Sandbrink, 2023). We recommend that the NIST AI 800-1 specifies that it (or parts thereof) might be appropriate for certain AI models that do not fit the dual-use foundation model definition, such as BDTs.

Differential treatment of different actors

To increase proportionality, NIST AI 800-1 should clarify that not all developers of dual-use foundation models should necessarily comply with all practices. Instead, different actors should be treated differently. This differentiation should be based on the amount of risks and benefits they produce, as well as their ability to comply with more costly practices. We suggest that the guidance has at least a paragraph describing the conditions under which more effort to reduce misuse risk is warranted, while also highlighting particular practices that are suitable for actors whose models pose the greatest misuse risk. In more detail:

- **Developers of the most capable models should comply with more extensive practices.** Generally speaking, more capable models pose higher risks. Hence, developers of the most capable models should comply with the most extensive practices. A good, but imperfect, proxy for a model's capabilities are the computational resources used to train it ("training compute") (Heim & Koessler, 2024; Koessler et al., 2024). We think it makes sense to use the same threshold as *Executive Order 14110* (White House, 2023), i.e. models trained with more than 10^{26} operations. Developers of dual-use foundation models trained with more than 10^{26} operations should be expected to comply with more extensive practices.

- **Inversely, developers of less capable models may adopt light-touch versions of the practices.** For example, developers of models trained with less than 10^{26} operations should not be expected to prepare a safety case (see below). They may also follow a simplified process for anticipating potential misuse risk (e.g. by including fewer and less detailed threat profiles).

- **Developers of open foundation models should be treated differently.** By "open foundation models", we mean models with broadly available model weights (Kapoor et al., 2024). Open foundation models have significant benefits (Seger et al., 2023; Kapoor et al., 2024; Eiras et al., 2024; NTIA, 2024; Bateman et al., 2024; Seger & O'Dell, 2024). For example, they can be used for new research, enable external oversight, and decentralize control. At the same time, they might increase misuse risk. This is because safety measures can be removed (Zhan et al., 2023; Lermen et al., 2023; Gade et al., 2023), deployment of open models is irreversible, and it is not possible to restrict access to the model for certain people (e.g. terrorists or authoritarian governments). Against this background, NIST AI 800-1 should modify some of the practices (perhaps in the Appendix). In particular, developers of open foundation models should consider alternatives to open-source, conduct more extensive risk assessments before open-sourcing a highly capable foundation model, and elect not to open-source foundation models that are more capable than all existing models.

  We wish to emphasize that, in our view, the ideological conflict between "pro-open" and "anti-open" camps has not been particularly helpful. We therefore welcome the emerging consensus (Bateman et al., 2024) and more nuanced discussion of openness in AI (Seger & O'Dell, 2024).

Prioritizing between different practices

Another way of increasing proportionality could be to prioritize between different practices. NIST AI 800-1 should clarify that some practices are more urgent or warrant more investment than others. This would support developers in identifying what are likely to be the most important practices.

- **Prioritize practices to inform developers' decisions about where to dedicate resources.** Developers will likely be at different stages when it comes to adopting practices to manage misuse risk. While some developers may have already adopted many of the practices in NIST AI 800-1, others may not have done so. Furthermore, developers that have already adopted practices covered in NIST AI 800-1 may need to review and update them in light of the guidelines. Developers will therefore need to make decisions about which practices they choose to implement first and how much they invest in each of the practices. Against this background, NIST AI 800-1 could indicate which practices are likely to be of higher or lower priority for the majority of developers.

3. Improvements to proposed practices

We recommend further improvements to the following practices: risk estimates, risk thresholds, post-deployment monitoring, and transparency.

Risk estimates

- **Add recommendations on *how* these estimates should be conducted.** NIST AI 800-1 does not provide any guidance on how developers should conduct different types of risk estimates. Since best practices do not yet exist, we think more guidance in the form of suggestions would be useful. For example, NIST AI 800-1 might specify whether estimates should be qualitative, quantitative, or semi-quantitative; who should conduct them; how estimates should be aggregated; how aggregated estimates should be combined, etc. (Schuett et al., forthcoming).

Validating model evaluations and risk estimates

- **Recommend that developers validate their model evaluations and risk estimates.** The science of AI model evaluation and risk estimation is still in its infancy. It is currently not clear what the results from misuse-relevant benchmarks entail about a model's potential for misuse. Sometimes, this is due to problems of internal validity. For example, a model might only perform well on the Weapons of Mass Destruction Proxy (WMDP) benchmark (Li & Pan et al., 2024) because it has memorized the test set, rather than having generally useful WMD-relevant capabilities.[3] Other times, there are problems of external validity. Even if the model is capable of supporting certain WMD-relevant tasks, users may fail to elicit or benefit from those capabilities. We recommend that NIST AI 800-1 (particularly under Objective 1) clearly states the problem of designing valid model evaluations and risk estimates, while highlighting methodologies that could be used to increase validity. These include post-deployment monitoring (mentioned in Practice 6.1), human uplift studies, and threat modeling.

Risk thresholds

Risk thresholds play a key role in NIST AI 800-1. For example, Practice 2.1 recommends identifying a level of misuse risk which the organization considers unacceptable, and Objective 5 recommends that organizations should only deploy a model if misuse risks are within their risk tolerance. Although we think these and other recommendations on risk thresholds are appropriate, we suggest the following changes.

- **Define terms related to AI risk thresholds and use them consistently.** NIST AI 800-1 does not define the terms "threshold", "risk threshold", and "risk tolerance". This is problematic because different people might interpret the terms differently.[4] We therefore suggest defining and using the following terms:

Term	Definition
Thresholds	Predefined values above which additional mitigations are deemed necessary.[5]
Risk thresholds	Thresholds defined in terms of the probability and severity of harm.
Capability thresholds	Thresholds defined in terms of model capabilities and adequate mitigations.
Red lines	Thresholds defined in terms of unacceptable model capabilities regardless of mitigations.
Training compute thresholds	Thresholds defined in terms of the computational resources used to train a model.

  These terms and definitions are consistent with our recent papers *Risk Thresholds for Frontier AI* (Koessler, et al., 2024) and *Training Compute Thresholds: Features and Functions in AI Regulation* (Heim & Koessler, 2024). They will also be used in an upcoming *OECD Survey on Thresholds for Advanced AI Systems* (see OECD, 2024b). We recommend adding these terms and definitions to the Glossary in Appendix A. Once defined, the terms should be used consistently throughout the document. This is currently not the case. For example, the terms "risk thresholds" and "risk tolerance" seem to be used interchangeably (the difference is at least not explained). NIST AI 800-1 seems to mostly use the term "risk threshold" to refer to the risk of harm. However, in Practice 3.2,

[3] See e.g. Alzahrani et al., 2024 model performance on popular benchmarks were shown to shift substantially by simply changing the ordering of the question responses.
[4] In particular, the term "risk threshold" is sometimes used both to refer to what we define as "risk threshold" and "threshold", which leads to confusion.
[5] Note that DSIT (2024) uses a slightly different definition, namely predefined values "at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable".

Recommendation 2, the term refers to the risk of model theft, which can be conceptualized as a risk factor.

- **Build the recommendations around "thresholds," rather than "risk thresholds" and "risk estimates".** These thresholds could be risk thresholds (i.e. thresholds defined in terms of the probability and severity of harm), but they could also be capabilities thresholds (i.e. thresholds defined in terms of model capabilities and adequate mitigations). In fact, we think that developers of dual-use foundation models should use a combination of risk thresholds and capabilities thresholds. Capabilities thresholds are useful because they can be evaluated more reliably, whereas risk thresholds are useful because they are more principled than capability thresholds (Koessler, et al., 2024).

- **Adjust recommendations on how to *set* risk thresholds.** The recommendations to effectively implement Practice 2.1 provide fairly detailed guidance on how to set risk thresholds. However, although they talk about *risk thresholds* (as defined above), they may also be appropriate for setting *capabilities thresholds*. Since there are no best practices on how to set risk thresholds, the recommendations should be toned down somewhat. They should acknowledge multiple ways of setting thresholds, including (1) by adapting existing risk thresholds from other industries, (2) by weighing potential harms and benefits, (3) by surveying people, or (4) by reviewing revealed preferences and ambient risks (McCaslin et al., forthcoming).

- **Recommend that assessments are made as to how close thresholds are to being breached**, in particular for the most capable models. This is consistent with Commitment II in the *Frontier AI Safety Commitments* (DSIT, 2024). It is important to make this recommendation explicit as some thresholds could be qualitative or otherwise difficult to assess.

- **Adjust recommendations on how to *use* risk thresholds.** NIST AI 800-1 recommends comparing risk estimates to risk thresholds to determine if a dual-use foundation model should be deployed. However, since the underlying estimates of the level of risk are still unreliable, we think risk thresholds should not be used to determine such decisions (Koessler, et al., 2024). Instead, they should primarily be used to set capabilities thresholds (see below). Besides that, they may be used as additional input to holistic, risk-informed decision making.

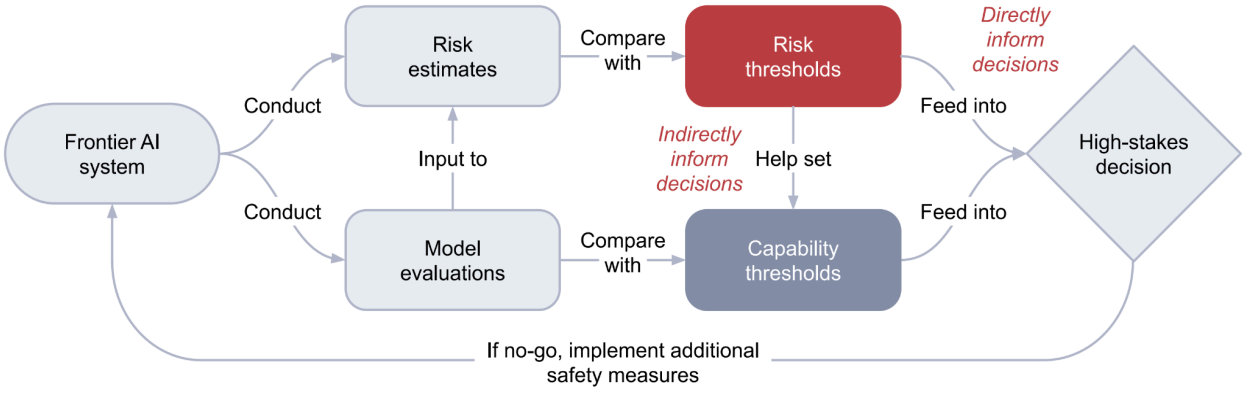


Figure 1: Risk thresholds can directly and indirectly inform high-stakes AI development and deployment decisions (Koessler, et al., 2024).

- **Add recommendations on how to *use* capabilities thresholds.** Capabilities thresholds are thresholds that are defined in terms of model capabilities and adequate mitigations. The basic idea is that developers conduct model evaluations (Shevlane et al., 2023; Phuong et al., 2024) and compare the results with pre-defined capabilities thresholds. These could be defined in terms of, for example, results on benchmarks or human uplift studies. They may only proceed (e.g. deploy a model) if either the model does not have certain capabilities of concern or they have put in place adequate mitigations. Several developers currently use this approach to make high-stakes development and deployment decisions (Anthropic, 2023; OpenAI, 2023; Google DeepMind, 2024). NIST AI 800-1 should recommend that developers should set capabilities thresholds and only deploy a model if either the model does not have certain capabilities of concern or the developer has put in place adequate mitigations.

- **Add recommendations on how to *set* capabilities thresholds.** There are not yet any best practices on how to set capabilities thresholds. In practice, most developers seem to use a somewhat unstructured process. First, they create risk models (also known as "threat models" or "threat profiles"). Then, they use human uplift studies and risk thresholds to specify different tiers of model capabilities. NIST AI 800-1 should recommend using a similar approach.

Post-deployment monitoring

Objective 6 recommends that developers of dual-use foundation models collect and respond to information about misuse after deployment. We think that the recommendations concerning information collection should make reference to AI agents.

- **Add recommendations for how post-deployment monitoring practices should account for AI agents.** AI agents are AI systems that can pursue complex goals, with little to no explicit human instruction for how to do so (Chan et al., 2023, 2024a; Shavit et al., 2023; Kolt, 2024a). Many developers are interested in building agents to automate a variety of real world tasks (e.g. OpenAI Assistants, Microsoft AutoGen, MultiOn, Auto-GPT). In the future, AI agents could facilitate misuse by reducing expertise bottlenecks. For example, they could potentially automate laboratory procedures (Bran et al., 2023). Since AI agents may be able to autonomously change their environments, evidence of misuse may not be obvious just from looking at the agents' outputs. Furthermore, since AI agents could interact with each other, evidence of misuse may also not be obvious just from looking at the outputs or the impacts of a single agent. One promising direction, therefore, is to increase visibility into AI agents and their actions (Chan et al., 2024a). Recommendations to support this could include (i) monitoring the impacts of an AI system's actions on its environment, in addition to monitoring the AI system's outputs, and (2) implementing methods for AI systems to identify themselves to human or software entities with which they interact.

Transparency

According to Practice 7.1, developers of dual-use foundation models should publish regular transparency reports. According to Practice 7.3, they should also report incidents and hazards to AI incident databases. While we think these recommendations are sensible, we suggest the following changes.

- **Expand the recommendations for publishing regular transparency reports.** We think that transparency reports should also include current and anticipated applications of the developers' model. This could include a description of: (1) the domains in which a model is deployed or is anticipated to be deployed, (2) the range of tasks the model performs or is anticipated to perform, and

(3) usage trends (Kolt et al., 2024b). This information could help improve transparency about potential misuse risks.

- **Recommend that companies report certain information directly to the government.** Many misuse risks will benefit from government visibility and potential intervention, even where publicly revealing such information could be harmful or undermine a companies' interest. For example, ahead of the release of an AI model that could significantly increase non-state actors' cyber capabilities, the government may benefit from putting more resources into cybersecurity. We expect that responsible practice from developers would, at a minimum, involve adhering to Executive Order 14110's requirements to notify the Department of Commerce about models trained using more than 10^{26} operations, along with information about results from red teaming. NIST AI 800-1 could further specify the types of information it may be important to report to the government ahead of a models' release, including estimated misuse risk and possible actions the government could take to reduce it.

- **Expand the recommendations for reporting incidents and hazards.** We think that developers should be encouraged to distinguish between "safety" and "security" incidents in their reports. This distinction may be helpful as there are usually different risk profiles and different treatments associated with safety and security incidents (Khlaaf 2023; Qi et al. 2024). In addition, we think that incident reports should include "near miss" incidents. These are events where a system's interaction with its environment had the potential to cause harm but did not (Wei & Heim, forthcoming). Near miss events are important because they can provide valuable lessons for risk management (Wei & Heim, forthcoming).

4. Suggesting new practices

We think that additional practices should be added to NIST AI 800-1, namely with regards to AI safety frameworks and AI safety cases.

AI safety frameworks

AI safety frameworks are risk management policies intended to keep the potential risks associated with developing and deploying dual-use foundation models to an acceptable level (METR, 2023, 2024). Some developers have already published their frameworks (Anthropic, 2023; OpenAI, 2023; Google DeepMind, 2024), while another 13 developers have signaled their intent to release similar frameworks by February 2025 (DSIT, 2024). Safety frameworks are important because they play a key role in companies' efforts to manage catastrophic risks from AI, including misuse risks. They are also mentioned prominently in several policy papers (e.g. DSIT, 2023).

- **Developers of the most capable foundation models should implement a safety framework.** NIST AI 800-1 should recommend that developers of the most capable models (e.g. models trained with more than 10^{26} operations) should implement an AI safety framework as specified in the *Frontier AI Safety Commitments* (DSIT, 2024).

AI safety cases

Safety cases are structured arguments, supported by evidence, that a system is sufficiently safe in a specific deployment context. They are common in other safety-critical industries like nuclear energy (Bounds, 2020), aviation (Denney et al., 2019), and autonomous vehicles (Myklebust et al., 2020).

Several scholars (Clymer et al., 2024; Bengio et al., 2024; Wasil et al., 2024; Buhl et al., forthcoming), companies (Anthropic, 2023; Google DeepMind, 2024), and governments (Irving, 2024) have recently suggested that safety cases should also be applied to AI systems.

Safety cases have a number of benefits relative to other risk management practices: (1) They ensure that deployment risk is analyzed in a systematic, comprehensive, and coherent way. They can help with noticing gaps or unjustified assumptions. (2) They ensure that deployment decisions are explicitly assessed in terms of overall risk, not just in terms of whether a process has been carried out successfully. Developers must explain not just what they have done at each stage, but also why they think it is sufficient to prevent unacceptable risk. (3) They facilitate transparency with respect to deployment decisions. They present information about the model, risks, and safeguards in a way that makes it easier to understand the overall reasoning behind a deployment decision. This enables both internal and external scrutiny. It would also help downstream developers to create safe products by making it clear under what conditions the systems would no longer be considered safe from misuse.

- **Developers of the most capable foundation models should prepare a safety case ahead of deployment.** NIST AI 800-1 already includes many elements of safety cases (e.g. risk thresholds, risk estimates, proportional safeguards). Safety cases would tie these elements together in a single comprehensive analysis of a given deployment decision. We think that NIST AI 800-1 should recommend that developers of the most capable models (e.g. models trained with more than 10^{26} operations) should prepare a safety case ahead of deployment. Initially, these safety cases may simply argue that (1) a dual-use foundation model would not pose unacceptable risk if it remains within predefined capability thresholds such as those set out by the developer in a safety framework, and that (2) a specific model is in fact within these thresholds. Over time, safety cases may come to be more bespoke for each individual release.

5. Role of other actors in the supply chain

Although NIST AI 800-1 focuses on the initial developers of dual-use foundation models, it acknowledges the role of other actors in the supply chain. We think NIST AI 800-1 should put more emphasis on how the initial developers should interact with downstream developers and compute providers.

Downstream developers

By “downstream developers”, we mean developers that apply foundation models to end user-facing applications (Küspert et al., 2023). Among other things, downstream developers could increase misuse risks by enhancing foundation models (Davidson et al., 2023) or by removing safeguards put in place by upstream developers (Zhan et al., 2023; Lermen et al., 2023; Gade et al., 2023). NIST AI 800-1 already recommends some actions that developers should take with respect to downstream developers, but we think the recommendations should be expanded further.

- **Developers of foundation models should take additional steps to manage risk from downstream development.** We think that NIST AI 800-1 should (1) highlight that incident-response processes should also consider the potential misuse of downstream applications, (2) recommend that “protections for reporting of misuse issues” apply to downstream developers, (3) recommend that developers consider what restrictions may be needed on how downstream developers can adjust the model, and (4) recommend that developers establish two-way communication channels with downstream developers to facilitate effective sharing of misuse risks and incidents.

- **Downstream developers should consider adopting the recommendations in NIST AI 800-1 where appropriate.** Although we agree that the focus should remain on the developers of foundation models, NIST AI 800-1 could highlight that downstream developers may want to consider how the recommendations apply to them. One way to do this would be to make a general recommendation (for example, in Section 2 on "Scope") that downstream developers also consider the guidelines where appropriate. An alternative approach would be to set out more specifically which recommendations are likely to apply to downstream developers (for example, in an Appendix to the guidelines).

Compute providers

Computing power ("compute") is crucial for the development and deployment of foundation models (Buchanan, 2020; Sastry et al., 2024). Since compute providers (also referred to as "cloud providers") can play a key role in AI governance and infrastructure security (Heim et al., 2024), it has been suggested that they should have responsibilities associated with AI development and deployment (Heim et al., 2024; Egan & Heim, 2023). Among other things, we think this should likely include taking steps to manage misuse risks.

- **Developers of foundation models may benefit from choosing responsible compute providers.** It will be in developers' interest to ensure that their supply chains (including their compute providers) are secure. For example, compute providers that offer their services to non-state actors may be endangering a developers' security due to the potential risk of "side-channel attacks". Furthermore, encouraging developers to choose responsible compute providers will likely be important for promoting responsible practices and reducing misuse risks more broadly. Examples of steps that developers may want compute providers to take include (1) implementing "Know Your Customer" requirements for large training runs, (2) securing IP and model weights to prevent theft, and (3) establishing processes for record keeping to allow for forensics and post-incident attribution (Heim et al., 2024; Egan & Heim, 2023).

Appendix A: List of recommended changes

Below, we suggest specific changes to NIST AI 800-1. Note that this does not cover all of our recommendations above, but only those where we have specific sentence-level suggestions.

	Current version	Suggested changes	Rationale
Section 2 (Scope)	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. These developers contribute most to determining how their models are made available, the models' capabilities, and safeguards against their misuse. […]	The practices in this document focus primarily on developers of foundation models ~~are principally focused on the central role that foundation models' initial developers have in the supply chain for their models~~. These developers contribute most to determining how their models are made available, the models' capabilities, and safeguards against their misuse. However, not all developers of foundation models should be treated equally. Developers of the most capable foundation models, i.e. models trained with less than 10^{26} operations, are expected to take more extensive measures than developers of less capable models. Inversely, developers of less capable models may adopt light-touch versions of the practices. Open foundation models should also be treated differently. Since open foundation models are particularly beneficial to society, while also exacerbating misuse risk, several practices need to be modified for developers of open foundation models. 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.	[Learn more] Treating all developers of foundation models the same will likely be disproportionate. The need for treating the most capable models differently is generally accepted in the literature (e.g. Bommasani et al., 2021; Anderljung et al., 2023; Bengio et al., 2024) and has been acknowledged in several policy documents (e.g. G7, 2023a, 2023b; DSIT, 2023, 2024). There is also emerging consensus that open foundation models require a more nuanced treatment (Seger et al., 2023; Bateman et al., 2024; Seger & O'Dell, 2024). Since open foundation models might also exacerbate misuse risk, they should not be exempt from NIST AI 800-1, even if they greatly benefit society. [Learn more] Suggesting that relevant stakeholders (e.g. downstream developers) may wish to consider how these practices apply to them increases the potential impact of the guidelines, whilst still maintaining the focus on initial developers.

	Current version	Suggested changes	Rationale
		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, compute providers, third-party evaluators and auditors, civil society organizations, and government agencies. 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. Relevant stakeholders are further encouraged to consider how these practices might form part of their own risk mitigation processes, where it is appropriate and proportionate to do so.	
Practice 1.2, Recommendation 4	N/A	Validate threat profiles and related estimates using methodologies such as post-deployment monitoring, human uplift studies, and threat modeling.	[Learn more] Employing relevant validation techniques will be important for improving the reliability of estimates.
Practice 1.3, Recommendation 5	N/A	Validate estimates of the model's capabilities using methodologies such as post-deployment monitoring, human uplift studies, and threat modeling.	[Learn more] Employing relevant validation techniques will be important for improving the reliability of estimates.
Practice 2.2, Recommendation 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.	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. Consider the use of physical, cybersecurity and insider threat safeguards, as well as how they work to complement each other to mitigate misuse risk.	[Learn more] Making reference to physical security controls, as well as cybersecurity and insider threat safeguards, will help to reduce misuse risk by ensuring that developers consider the full range of security controls available and how they complement one another.
Practice 5.3, Recommendation 1	For each deployment, establish a process to determine whether the deployment should proceed	For each deployment, establish a process to determine whether the deployment should proceed	[Learn more] For the most capable models, it is necessary to have a high

	Current version	Suggested changes	Rationale
	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.	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. For the most capable models, only deploy after a safety case has been produced and has passed internal review, with external input where appropriate.	level of assurance of safety before deploying a model. A safety case ensures that a comprehensive assessment of whether misuse risk has been adequately managed. An internal review process for the safety case decreases the risks of errors or flaws in reasoning that might mean that misuse risk is greater than estimated, helping to prevent the deployment that exceeds the organization's risk tolerance.
Practice 5.3, Recommendation 3	N/A	Ahead of each deployment, consider whether there should be any restrictions on how downstream developers can adjust the model.	[Learn more] Restrictions on how downstream developers can adjust the model could help to keep misuse risks at an acceptable level.
Practice 6.1, Recommendation 7	N/A	Consider monitoring the impacts of an AI system's actions on its environment, in addition to monitoring the AI system's outputs. To perform such monitoring, developers may have to work with actors that provide digital services for AI systems (e.g. financial transactions or access to a robotic laboratory). Developers may also have to collate information across different monitoring systems.	[Learn more] Such monitoring could be particularly helpful for AI agents (Chan et al, 2024a). This is consistent with the goal of Objective 6 to "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, Recommendation 8	N/A	Consider implementing methods for AI systems to identify themselves to human or software entities with which they interact. For example, an AI system could present an ID whenever it sends a JSON request to an external service.	[Learn more] As above. Entities that interact with AI agents could have information pertinent to detecting and addressing misuse, if they can identify with which agent they are interacting (Chan et al. 2024a, 2024b).

	Current version	Suggested changes	Rationale
Practice 6.2, Recommendation 5	N/A	Consider how processes for responding to incidents of model misuse will apply to the potential misuse of downstream applications.	[Learn more] Ensures that incident response processes are broad enough to cover the range of downstream applications that may result in misuse.
Practice 6.3, Recommendation 1	Adopt policies that protect and reward individuals who report model issues related to misuse risk.	Adopt policies that protect and reward individuals who report model issues related to misuse risk. Consider how these policies may apply to downstream developers.	[Learn more] Ensures that relevant policies are extended to downstream developers, where appropriate, to encourage reporting of model issues.
Practice 7.1	Publish regular transparency reports that include key details regarding misuse risks and how those risks are managed.	Publish regular transparency reports, as well as safety cases for the most capable models, that include key details regarding misuse risks and how those risks are managed.	[Learn more] See below.
Practice 7.1, Recommendation 7	N/A	For the most capable models, also produce a safety case that demonstrates that deployment does not pose misuse risk that exceeds the organization's risk tolerance, and explain the conditions under which the safety case would be invalidated. Publish a version of the safety case that includes as much information as possible without disclosing proprietary information or introducing risks to public safety. Share more detailed versions with the most relevant external stakeholders, such as downstream developers, third party evaluators, or regulatory bodies.	[Learn more] To facilitate transparency into deployment decisions for the most capable models, organizations should not only publish relevant information about the model, as outlined in Recommendation 1-5, but explain why and how the information demonstrates that misuse risk does not exceed the organization's risk tolerance. They should also explain the conditions under which the safety case holds. This will enable downstream developers and other stakeholders to take appropriate actions to mitigate the risks of potential future products that use or are based on the model. Complete safety cases will likely include proprietary information, for example copyrighted training algorithms, as well as information that might pose risks to public safety, for example detailed threat profiles. Organizations should

	Current version	Suggested changes	Rationale
			therefore not publish complete safety cases, but should instead produce several versions to publish to different audiences.
Practice 7.1, 7.2 and 7.3, Documentation	N/A	A summary of the organization's policies with respect to what information the organization will share externally, who it will share it with (e.g. government, the public), and what information it will **not** share. Consider including details on when information-sharing would be subject to a risk assessment and how to conduct such an assessment.	[Learn more] This would help to provide greater transparency around organizations' information-sharing practices.
Practice 7.1, Recommendation 7	N/A	Share information about the current and anticipated applications of the model, including the domains in which it is or will likely be deployed, the tasks it performs or is expected to perform, and any relevant information on usage trends.	[Learn more] This would help to provide greater transparency around potential misuse risks.
Practice 7.3, Recommendation 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.	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, clearly distinguishing between safety and security incidents.	[Learn more] Safety and security incidents usually have different risk profiles and require different treatment. Therefore, it is sensible to distinguish between them when recording and reporting incidents.
Practice 7.3, Recommendation 4	N/A	Consider including "near miss" incidents in reports to inform future management of potential misuse risks.	[Learn more] Including near-miss events improves the accuracy of reporting and is valuable for safety learning.
Practice 7.3, Recommendation 5	N/A	Establish two-way communication channels with downstream developers to facilitate effective sharing of misuse risks and incident reports.	[Learn more] Ensures that initial developers and downstream developers are effectively communicating about misuse risks and incidents which will support the development of effective risk mitigation strategies.

Appendix B: Overview of objectives and practices

The following table gives an overview of all objectives and practices in NIST AI 800-1.

Objectives	Practices
Objective 1: Anticipate potential misuse risk	**Practice 1.1**: Identify and maintain a list of threat profiles that covers significant ways in which malicious actors might misuse the model.
	Practice 1.2: Assess the impact of each identified threat profile if the malicious actor successfully misused the model.
	Practice 1.3: Estimate the model's capabilities of concern before it is developed by comparing it to existing models.
Objective 2: Establish plans for managing misuse risk	**Practice 2.1**: Identify a level of misuse risk which the organization considers unacceptable.
	Practice 2.2: Establish a roadmap to manage misuse risks for the development of planned foundation models and future versions.
Objective 3: Manage the risks of model theft	**Practice 3.1**: Assess the risk of model theft from relevant threat actors.
	Practice 3.2: Compare predicted misuse risk to the organization's risk tolerance prior to developing models with increased capabilities of concern.
	Practice 3.3: Maintain security practices sufficient to prevent model theft.
Objective 4: Measure misuse risk	**Practice 4.1**: Measure model capabilities relevant to assessing misuse risk.
	Practice 4.2: Use red teams to assess whether threat actors could bypass model and system 5 safeguards and misuse any capabilities of concern.
Objective 5: Ensure that misuse risk is managed before deploying foundation models	**Practice 5.1**: Assess the effect of a potential deployment on the model's misuse risk.
	Practice 5.2: Implement safeguards proportionate to the model's misuse risk.
	Practice 5.3: Only pursue deployments where misuse risk is adequately managed.
Objective 6: Collect and respond to information about misuse after deployment	**Practice 6.1**: Where possible, monitor distribution channels for evidence of misuse.
	Practice 6.2: Maintain a process to respond to incidents of model misuse.
	Practice 6.3: Establish protections for internal reporting of misuse issues.
	Practice 6.4: Provide safe harbors for third-party safety research.
	Practice 6.5: Create bounties for issues related to the misuse risk.
Objective 7: Provide appropriate transparency about misuse risk	**Practice 7.1**: Publish regular transparency reports that include key details regarding misuse risks and how those risks are managed.
	Practice 7.2: Disclose information about risk management practices to promote accountability.
	Practice 7.3: Report incidents and hazards related to the foundation model to AI incident databases.

References

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., … Khan, H. (2024). *When benchmarks are targets: Revealing the sensitivity of large language model leaderboards*. arXiv. https://arxiv.org/abs/2402.01781

Anthropic. (2023, September 19). *Anthropic's Responsible Scaling Policy* (Version 1.0). https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Bateman et al., 2024). *Beyond open vs. closed: Emerging consensus and key questions for foundation AI model governance*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance

Bengio, Y., et al., (2024, May 17). *International scientific report on the safety of advanced AI: Interim report*. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

Bounds, A. (2020). Implementation of nuclear safety cases. *Safety and Reliability, 39*(3–4), 203–214. https://doi.org/10.1080/09617353.2020.1800977

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2023). *ChemCrow: Augmenting large-language models with chemistry tools*. arXiv. http://arxiv.org/abs/2304.05376

Buchanan, B. (2020). *The AI triad and what it means for national security strategy*. Center for Security and Emerging Technology. https://doi.org/10.51593/20200021

Buhl, M., Schuett, J., & Anderljung, M. (forthcoming). *Safety cases for frontier AI*.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., … Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. *ACM Conference on Fairness, Accountability, and Transparency*, 651–666. https://doi.org/10.1145/3593013.3594033

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024a). Visibility into AI agents. *ACM Conference on Fairness, Accountability, and Transparency*, 958–973. https://doi.org/10.1145/3630106.3658948

Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). *Safety cases: Justifying the safety of advanced AI systems*. arXiv. http://arxiv.org/abs/2403.10462

Davidson, T., Denain, J.-S., Villalobos, P., & Bas, G. (2023). *AI capabilities can be significantly improved without expensive retraining*. arXiv. http://arxiv.org/abs/2312.07413

Denney, E., Pai, G., & Whiteside, I. (2019). The role of safety architectures in aviation safety cases. *Reliability Engineering & System Safety*, 191(106502), https://doi.org/10.1016/j.ress.2019.106502

DSIT. (2023). *Emerging processes for frontier AI safety*. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety

DSIT. (2024). *Frontier AI Safety Commitments, AI Seoul Summit 2024*. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

Egan, J., & Heim, L. (2023). Oversight for frontier AI through a Know-Your-Customer scheme for compute providers. arXiv. https://arxiv.org/abs/2310.13625

Eiras, F., Petrov, A., Vidgen, B., de Witt, C. S., Pizzati, F., Elkins, K., ... & Foerster, J. (2024). *Near to mid-term risks and opportunities of open source generative AI*. arXiv. https://arxiv.org/abs/2404.17047

G7 (2023a). Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System. https://www.mofa.go.jp/files/100573471.pdf

G7 (2023b). Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. https://www.mofa.go.jp/files/100573473.pdf

Gade, P., Lermen, S., Rogers-Smith, C., & Ladish, J. (2023). *BadLlama: Cheaply removing safety fine-tuning from Llama 2-Chat 13B*. arXiv. http://arxiv.org/abs/2311.00117

Heim, L., & Koessler, L. (2024). *Training compute thresholds: Features and functions in AI regulation*. arXiv. http://arxiv.org/abs/2405.10799

Irving, G. (2024). *Safety cases at AISI*. UK AI Safety Institute. https://www.aisi.gov.uk/work/safety-cases-at-aisi

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... & Narayanan, A. (2024). *On the societal impact of open foundation models*. arXiv. https://arxiv.org/abs/2403.07918

Khlaaf, H. (2023). *Toward comprehensive risk assessments and assurance of AI-based systems*. Trail of Bits. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf

Kolt, N. (2024a). *Governing AI agents*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772956

Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., Heim, L., Rodriguez, M., Sandbrink, J. B., & Woodside, T. (2024b). *Responsible reporting for frontier AI development*. arXiv. https://arxiv.org/abs/2404.02675

Küspert, S., Moës, N. & Dunlop, C. (2023). *The value chain of general-purpose AI*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai

Lermen, S., Rogers-Smith, C., & Ladish, J. (2023). *LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B*. arXiv. http://arxiv.org/abs/2310.20624

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., … Hendrycks, D. (2024). *The WMDP benchmark: Measuring and reducing malicious use with unlearning*. arXiv. http://arxiv.org/abs/2403.03218

McCaslin, T., Koessler, L., & Schuett, J. (forthcoming). *Setting risk thresholds for frontier AI*.

METR (2023). *Responsible scaling policies (RSPs)*. https://metr.org/blog/2023-09-26-rsp

METR (2024). *Common elements of frontier AI safety policies*. https://metr.org/blog/2024-08-29-common-elements-of-frontier-ai-safety-policies

Myklebust, T., Stalhane, T., Jenssen, G. D., & Waro, I. (2020). Autonomous cars, trust and safety case for the public. *Annual Reliability and Maintainability Symposium*. https://doi.org/10.1109/RAMS48030.2020.9153618

NTIA (2024). *Dual-use foundation models with widely available model weights report*. https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report

OECD (2024a). *OECD AI principles*. https://oecd.ai/en/ai-principles

OECD (2024b). *Public consultation on risk thresholds for advanced AI systems*. https://oecd.ai/en/wonk/seeking-your-views-public-consultation-on-risk-thresholds-for-advanced-ai-systems-deadline-10-september

OpenAI. (2023, December 18). *Preparedness Framework (Beta)*. https://openai.com/safety/preparedness

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., ... & Shevlane, T. (2024). *Evaluating frontier models for dangerous capabilities*. arXiv. https://arxiv.org/abs/2403.13793

Qi, X., Huang, Y., Zeng, Y., Debenedetti, E., Geiping, J., He, L., … Mittal, P. (2024). *AI risk management should incorporate both safety and security*. arXiv. http://arxiv.org/abs/2405.19524

Sandbrink, J. B. (2023). *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*. arXiv. https://arxiv.org/abs/2306.13952

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). *Computing power and the governance of artificial intelligence*. arXiv. https://arxiv.org/abs/2402.08797

Schuett, J., et al. (forthcoming). *How to estimate the likelihood and impact of risks from AI*.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Ó hÉigeartaigh, S., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., … Gupta, A. (2023b). *Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. arXiv. https://arxiv.org/abs/2311.09227

Seger, E., & O'Dell, B. (2024). *Open Horizons: Exploring nuanced technical and policy approaches to openness in AI*. Demos. https://demos.co.uk/research/open-horizons-exploring-nuanced-technical-and-policy-approaches-to-openness-in-ai

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). *Model evaluation for extreme risks*. arXiv. https://arxiv.org/abs/2305.15324

Wasil, A., Clymer, J., Krueger, D., Dardaman, E., Campos, S., & Murphy, E. (2024). *Affirmative safety: An approach to risk management for advanced AI*. SSRN. https://dx.doi.org/10.2139/ssrn.4806274

Wei, K., & Heim, L. (forthcoming). *Designing incident reporting systems for harms from AI*.

White House. (2023). *Safe, secure, and trustworthy development and use of artificial intelligence* (Executive Order 14110). https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., & Kang, D. (2023). *Removing RLHF protections in GPT-4 via fine-tuning*. arXiv. https://arxiv.org/abs/2311.05553