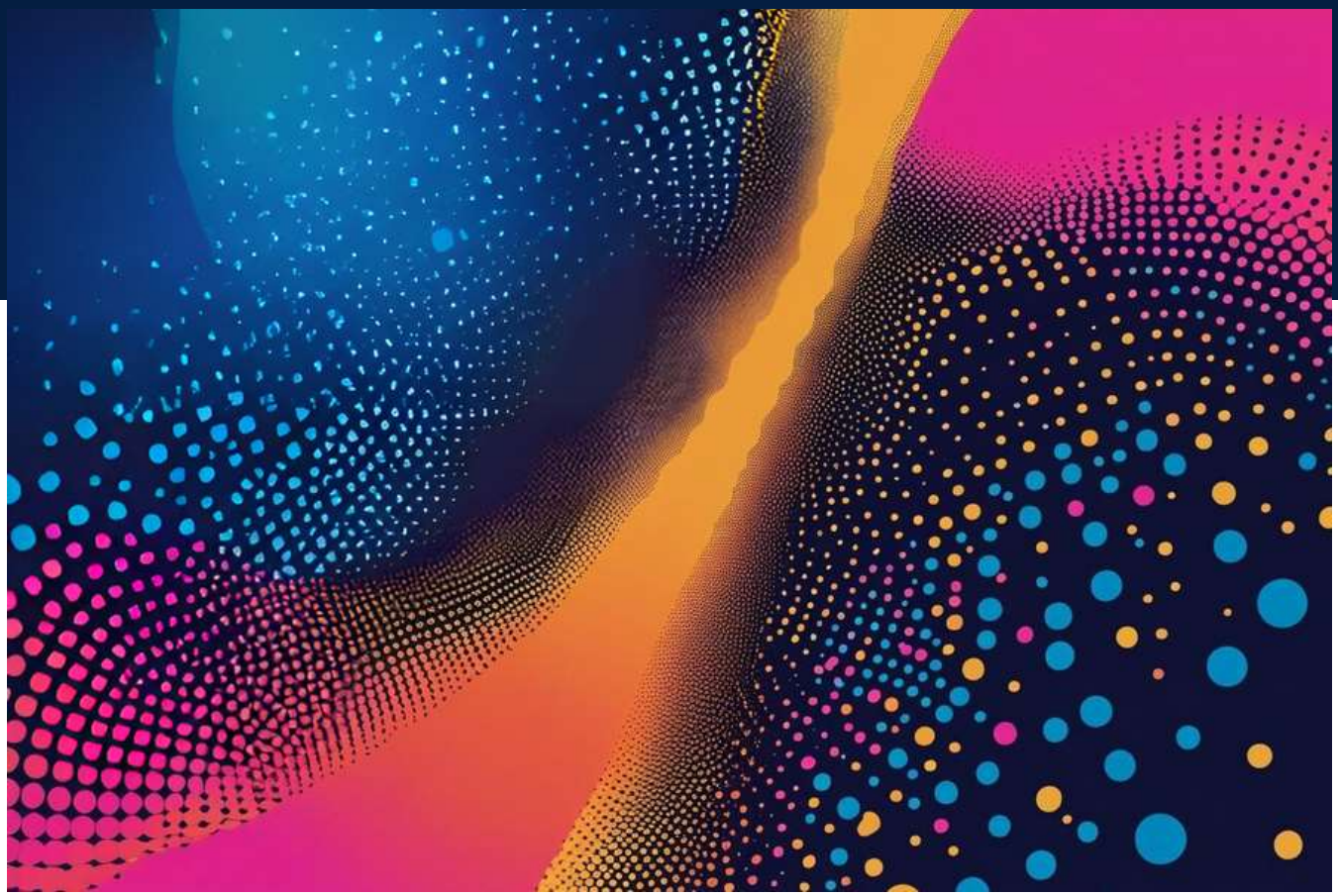
RESEARCH MEMO
2025 JUNE

Risk Tiers: Towards a Gold Standard for Advanced AI



Authors: Nicholas A. Caputo, Siméon Campos, Stephen Casper, James Gealy, Bosco Hung, Julian Jacobs, Daniel Kossack, Toni Lorente, Malcolm Murray, Seán Ó hÉigeartaigh, Amin Oueslati, Henry Papadatos, Jonas Schuett, Anna Katariina Wisakanto, Robert Trager

Risk Tiers: Towards a Gold Standard for Advanced AI

Nicholas A. Caputo,*[1] Siméon Campos,[2] Stephen Casper,[3] James Gealy,[2] Bosco Hung,[4] Julian Jacobs,[5] Daniel Kossack,[2] Toni Lorente,[6] Malcolm Murray,[2] Seán Ó hÉigeartaigh,[7] Amin Oueslati,[6] Henry Papadatos,[2] Jonas Schuett,[8] Anna Katariina Wisakanto,[9] Robert Trager**[1]

* Corresponding author: Nicholas A. Caputo nick.caputo@oxfordmartin.ox.ac.uk
** Senior author
[1] Oxford Martin AI Governance Initiative
[2] SaferAI
[3] MIT
[4] Department of Politics and International Relations, University of Oxford
[5] Google DeepMind
[6] The Future Society
[7] Leverhulme Centre for the Future of Intelligence
[8] Centre for the Governance of AI
[9] Center for AI Risk Management & Alignment

Given this document's scope, inclusion as an author does not necessarily entail endorsements of all aspects of the report.

Executive Summary

Increasing risks from advanced AI demand effective risk management systems tailored to this rapidly changing technology. One key part of risk management is establishing risk tiers. Risk tiers are categories based on expected harm that specify in advance which mitigations and responses will be applied to systems of different risk levels. Risk tiers force AI companies to identify potential risks from their systems and plan appropriate responses. They also provide public transparency regarding the risk level society is accepting from AI and how those risks are being managed.

Risk management, including risk tiering, has received attention from both policymakers and industry, but different organizations have taken divergent and sometimes incompatible approaches. This diversity has facilitated innovation and experimentation in adapting risk management to the challenges of advanced AI. However, it has also made it difficult to understand the overall risk picture and how each system and developer contributes to it, as well as to compare the effectiveness of different risk estimation and mitigation practices. As such, a more standardized approach to risk tiering–one that achieves the benefits of effective aggregation, comparison, and consistent scientific grounding while preserving space for innovation–is needed.

To explore such standardization, the Oxford Martin AI Governance Initiative (AIGI) convened experts from government, industry, academia, and civil society to lay the foundation for a gold standard for advanced AI risk tiers. A complete gold standard will require further work. However, the convening provided insights for how risk tiers might be adapted to advanced AI while also establishing a framework for broader standardization efforts.

Insights from the convening included the following:

1. **Quantitative risk tiers clarify the relationship between hazardous capabilities and expected harm; systematic qualitative modeling should apply where quantitative approaches fail.** Quantitative risk modeling provides a basis for risk-informed decisionmaking and represents best practice in safety-critical industries like nuclear safety and aviation. Such modeling helps risk managers and the public understand what risks a system actually poses, instead of simply whether a harmful capability exists. Quantitative modeling also facilitates mitigations by providing clarity on how they reduce risk, for example whether they reduce the likelihood or severity of harm. For those AI risks where modeling is possible, risk managers should apply quantitative estimates. However, for some risks, quantitative estimates come with unacceptably large error bars. There, systematic scenario- or source-based

modeling should be used to identify possible harms transparently while clearly conveying the risk level.

2. **AI systems should be classified into risk tiers at defined points throughout their lifecycle.** Risk assessment, tier classification based on that assessment, and risk treatment should occur repeatedly from before pretraining (based on capability estimates) to after deployment. Mitigations should map onto each risk tier at each step.
3. **Benefits from AI releases should be considered alongside the risks, but more measurement work is needed before risks and benefits can be compared effectively.** In some cases, advanced AI's benefits may outweigh certain risk increases. However, benefits remain difficult to estimate, and more work must occur to enable reasonable benefit-risk comparisons. Quantitative risk tiers could facilitate these comparisons by providing a shared comparative framework.
4. **Standardized risk management practices likely enable better oversight of risks, their interactions, and responses from risk managers, auditors, and regulators.** Current scholarly discussions of frontier AI labs' risk management often focus on individual company processes, but society is accepting risk from all companies collectively. Determining the overall risk level from advanced AI, which systems contribute specific parts of that risk, and how different systems might interact to cause unforeseen harms is challenging. Innovation in risk management should be balanced with standardization efforts that allow governments to understand the overall risk landscape. Frameworks for integrating risk assessments, analyzing organizational risk contributions, and verifying best practice adoption would aid this understanding.
5. **Risk tier modeling should go beyond capability assessments to include how they might become threats when deployed.** User capabilities and the characteristics of the overall risk environment determine the threat levels presented by key AI risk sources. To accurately underpin risk tiers, evaluations should account for these factors. AI companies may need to work with other actors better situated to provide certain informational inputs. For example, organizations like AISIs should provide risk landscape inputs for risk management processes, supplementing capabilities evaluations performed by AI companies.

Risk tiers clarify the harms AI might present and identify the measures being taken to prevent them. Establishing a gold standard for risk tiers will help create consensus on existing best practices and where more work is needed.

Table of Contents

I. Introduction

As AI capabilities advance, the technology's risks become more significant. Frontier systems could enable significant misuse in cybersecurity, biological, chemical, radiological, and nuclear (CBRN) domains and potentially create loss of control risks. To address these emerging threats, those working to govern advanced AI must develop robust risk management practices to identify, evaluate, mitigate, and respond to potential harms.

Risk tiers are a key element in such risk management practices. Their clear structure can guide AI developers, regulators, and the public in understanding specific risks from advanced systems and facilitate deliberation on acceptable risk levels. Defining what risk level is "unacceptable" sharpens the debate and forces people to decide how much risk they will accept. Risk tiers also provide definite thresholds to trigger the use of higher precautions, simplifying the risk management chain. In advanced AI, risk tiers could classify systems based on the expected harm that they might cause, with each tier mapped to specific mitigations required for systems in that tier.

Risk management has become an increased focus of attention in advanced AI governance, manifesting in both government frameworks and self-regulatory initiatives by leading companies. The European Union AI Act General Purpose AI Code of Practice (CoP) draft explicitly calls for risk tiering among its risk management practices, and many leading AI developers have released risk management frameworks pursuant to their commitments at the Seoul AI Summit.[1] These frameworks often include risk tiers that distinguish between safe and unsafe systems and specify what companies will do when a system exceeds safety thresholds. Researchers have also drawn on insights from other safety critical domains like nuclear and aviation to propose risk management practices suited to advanced AI,[2] while the underlying AI risk identification, measurement, and mitigation sciences continue advancing.

However, AI risk management remains immature, and the various efforts just enumerated are diffuse and with sometimes contradictory approaches to addressing risk. As this domain evolves and technology advances, determining and establishing best practices in each risk management component becomes essential so that regulators and companies can better understand and adopt them. Creating a gold standard for risk tiering in advanced AI would provide a model for broader standardization, create a focal point for deliberation, coordination, and adoption, and bring clarity to an increasingly complex field. While the underlying science has not reached a point of sufficient maturity such that fully optimal approaches can be determined, working toward a gold standard provides a framework for identifying where more research is needed and helps guide overall efforts.

To address this need, the Oxford Martin AI Governance Initiative convened a group of experts from government, industry, academia, and civil society to lay groundwork for a gold standard for advanced AI risk tiers. The convening explored key considerations and alternatives that could inform policymakers and practitioners. The discussion positioned risk tiers within the broader risk management landscape and considered how risk tiers could operate, directly and in their interactions with the broader risk cycle, modeling, and mitigations. Participants also debated whether risk tiering should

---

[1] *See Third Draft of the General-Purpose AI Code of Practice, Commitments By Providers Of General-Purpose AI Models With Systemic Risk, Safety And Security Section*, EUROPEAN COMMISSION (Mar. 11, 2025), https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts; *Frontier AI Safety Commitments, AI Seoul Summit 2024*, UK DEPARTMENT FOR SCIENCE, INNOVATION AND TECHNOLOGY (Feb. 7, 2025), https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

[2] *See, e.g.*, Siméon Campos, Henry Papadatos, et al., *A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management*, ARXIV 3 (Feb. 19, 2025), https://arxiv.org/abs/2502.06656; Leonie Koessler, Jonas Schuett, & Markus Anderljung, *Risk thresholds for frontier AI*, ARXIV (Jun. 20, 2024), https://arxiv.org/abs/2406.14713.

consider the potential benefits of advanced AI and harms from not deploying a safe system. Finally, they discussed questions of the relationship between risk governance and risk tiers and how to ensure that tiers work effectively to mitigate possible harms once established.

While substantial work remains to achieve a gold standard in risk tiering and risk management overall, beginning now offers significant benefits. Risks from advanced AI systems are emerging, and putting into place even imperfect versions of risk management could prevent serious harms. Furthermore, formulating gold standards identifies where improvements are needed and enables future work.

II. Risk tiers and the broader risk management landscape

Risk management is a complex process in which risk tiering plays just one important part, and it is important to situate risk tiers within that process to understand their organizing role. Fundamentally, risk management identifies and seeks to understand risk sources, then determines if and how to respond. The set of possible harms from advanced AI remains open, though some consensus has emerged. Regulators and AI companies agree that misuse risks around AI systems enabling cyber and CBRN hazards present pressing threats.[3] Loss-of-control risks, where an AI pursues goals or takes actions contrary to designer or user intent, have also become an increasing focus and are included in some company risk management frameworks.[4] However, rapidly-improving advanced AI will likely present novel "unknown unknown" risks beyond those currently anticipated. As such, risk management frameworks must orient toward future risk identification and classification while remaining responsive to existing risks.

Once risks are identified, risk managers must determine acceptable risk levels and decide measures to take when a risk exceeds thresholds. Risk tiering facilitates classification by dividing risk distributions into clear sections defined by thresholds.

---

[3] *See, e.g.*, Yoshua Bengio et al., *International AI Safety Report 2025*, UK AI Security Institute (Jan. 2025), https://arxiv.org/abs/2501.17805; Ben Nimmo et al., Disrupting malicious uses of our models: an update February 2025, OpenAI (Feb. 21, 2025), https://openai.com/global-affairs/disrupting-malicious-uses-of-ai/; Google Threat Intelligence Group, *Adversarial Misuse of Generative AI*, Google (Jan. 29, 2025), https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai.

[4] OpenAI's Preparedness Framework, for example, now includes a risk category for "Autonomous Replication and Adaptation" of systems. *Preparedness Framework v2*, OpenAI (Apr. 15, 2025), https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Each risk tier can then be associated with specific responses, from halting development or deployment to post-deployment mitigations like content guardrails. Risk tiers force risk managers to define in advance which risk levels are serious or unacceptable and commit to responding when they have been reached. Who sets these levels is an important political question currently being left to frontier AI companies in the absence of regulatory intervention.

Risk tiering requires risk managers to measure a system's levels of risks with sufficient precision for the system to be classified into a tier. In advanced AI, risk measurement has so far mostly taken the form of capability evaluations, in which AIs are tested on particular tasks and benchmarked according to their success. The more successful an AI is at facilitating or completing risk-related tasks, the more risk it may present. Some capability evaluations aim to prove model capabilities within areas that might generalize to dangerous domains. For example, mathematics-based capabilities estimates that might demonstrate generalizable intelligence or facility with computer programming operate this way.[5] Other evaluations focus on directly determining how much systems present or increase a defined risk. A biological synthesis capability evaluation that measures how systems improve human abilities to synthesize dangerous biological agents over some baseline like simple internet search exemplifies this approach.[6]

Most risk management has relied on capability evaluations to assess risks and set thresholds. Capability evaluations are relatively simple to carry out and avoid uncertainty associated with more complicated risk thresholding approaches that estimate the likelihood and severity of risk manifestation. However, they have drawbacks, including potentially missing risks from "jagged" capability profiles, not capturing the harm potential if mitigations fail, and acting only as lower bounds rather than full capabilities estimates. As risk management matures and evaluation, classification, and risk modeling improve, more complex risk thresholding and estimation approaches may supplement or replace capability-based approaches. Meanwhile, scenario-based approaches can bridge the gap. The convening explored these new techniques and their integration into the risk management process.

---

[5] *See, e.g.*, Tamay Besiroglu, Elliot Glazer, & Caroline Falkman Olsson, FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, Epoch AI (Nov. 8, 2024), https://epoch.ai/frontiermath/the-benchmark (arguing that mathematics provides a relatively generalizable context for evaluating reasoning capabilities).

[6] *See Issue Brief: Preliminary Taxonomy of AI-Bio Safety Evaluations*, FRONTIER MODEL FORUM (Dec. 20, 2024), https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-ai-bio-safety-evaluations/.

III. Defining risk tiers

The convening discussed several ways risk tiers for advanced AI could be defined given existing practices, likely future risks, and the pace of development of risk management. How best to define risk tiers given different potential harms from advanced AI is challenging. However, best practices can be clarified as advanced AI risk management matures. Engaging in risk tiering now can help identify remaining uncertainty around gold standards and address them effectively.

A. What might risk tiers look like?

A risk tiering approach should produce comprehensible divisions of the AI's risk space into different parts, each associated with mitigations that reduce or manage risk appropriately. For example, the risk space might be divided into something like "acceptable," "risky," "dangerous," and "unacceptable" tiers for system classification based on predicted or measured risk. The number of tiers and placement of cutoffs should depend on society's risk tolerance and each category's risk sources. What that societal risk choice process should look like remains undetermined, though other risk-sensitive domains suggest techniques like revealed preference estimation or surveying of affected populations as starts.[7]

Any risk tiering process requires risk estimation allowing risk managers to understand the risk space and tier placement. Leading practice in other safety-critical industries like nuclear and aviation is to seek quantitative estimates of the likelihood and severity of identified harms to create risk matrices or graphs to sort expected harms by acceptability. For example, an 85% chance of relatively insignificant harm and a 1% chance of serious harm might both be deemed acceptable while a 95% chance of relatively insignificant harm and a 5% chance of serious harm might both be deemed to be unacceptable. Such a matrix might appear as follows, though each category should map to quantitative measures:

[7] See Reuben C. Arslan et al., *How people know their risk preference*, 10 NATURE SCIENTIFIC REPORTS 15365, https://doi.org/10.1038/s41598-020-72077-5 for a recent discussion and method of risk acceptance elicitation.

Likelihood \ Severity	Negligible	Minor	Serious	Severe	Catastrophic
Near-certain	Medium	High	Extreme	Critical	Critical
Likely	Medium	Medium	High	Extreme	Critical
Moderate	Low	Medium	Medium	High	Extreme
Unlikely	Minimal	Low	Medium	Medium	High
Rare	Minimal	Minimal	Low	Medium	Medium

Subsequently, mitigations can map to risk tiers. Some mitigations could apply to any system reaching a given tier, while others could apply to systems in a given tier to push them into a lower tier. If mitigations cannot reduce risks to acceptable levels, the system at issue should not be deployed.

B. What underlies risk tiers?

All risk tiers should use robust risk modeling that determines criticality thresholds in the risk spectrum. This modeling should begin from model characteristics, hazards, hazardous situations, and possible harms the risk management process is seeking to prevent. It should then trace the causal pathways where changes in model characteristics and other risk sources might contribute to actualizing harms. Modeling both AI system capabilities and complex deployment environments is necessary. Risk managers must understand not just what their system can do but also malicious actors' abilities to estimate the likelihood and severity of misuse events. Quantitative measures of severity and likelihood are best practices in many safety-critical industries and allow risk managers to create risk budgets to balance safety and cost across development and deployment. Achieving such an estimate where possible in advanced AI would represent a significant advance in AI risk management.

However, many harms confronting AI risk management are difficult to fit into quantitative frameworks, at least given current evaluation and measurement tools. For example, the severity of loss of control risks seems difficult to estimate given the variance of potential outcomes. Developing a quantitative science for such harms is an important research direction. But other significant risks like damage to the information environment or to fundamental rights seem inherently qualitative and difficult to quantitatively model. Determining how to set risk tiers across qualitative spectrums presents significant challenges requiring both scientific progress and public

deliberation. Identifying criteria for when to use quantitative or qualitative risk tiering methods would enable managers to address a broader set of risks.

C. How might risk tiers be developed?

Creating a set of preferred risk tiering approaches that adapt as AI and its risk sciences improve should help risk managers implement effective tiering. Where risks are amenable to quantitative analysis and resemble traditional risk management domains, using the estimated hazard approach is best practice.

Cybersecurity risks are one domain where this approach may fit. Cybersecurity is a relatively well-developed risk domain with management methodologies battle-tested over time. This maturity provides a risk baseline against which increased risk from advancing AI capabilities can be distinguished and measured. Furthermore, relatively minor harms from AI cyberoffensive capabilities will likely emerge before major ones do because the technology will advance such that easier vulnerabilities in weak systems will be exploited before more difficult ones in hardened systems. Existing cyber incident reporting infrastructure means that many attacks will be identified and reported. These "practical evaluations" of the technology will provide useful information about how AI is changing cybersecurity and allow modeling of its future effects.

Given these favorable risk management characteristics, the gold standard for risk tiers in this context would likely involve traditional severity and likelihood estimation approaches to create risk matrices or distributions with clear tiering cutoffs. As discussed above, something akin to the X% chance of Y harm (whether measured in lives lost, economic damage, or other measures) would fit well in this context. Similar approaches could be deployed in other domains as advances in measurement make them more susceptible to quantification. Adopting quantitative estimates across domains would allow for better overall understanding of AI systems' risk profiles. Many inputs developed for early efforts could be used to provide a basis for that future work (because, for example, a bioterrorist group might be willing to use chemical weapons for similar purposes, though their chemistry capabilities might be different).

Where quantitative measurement is impossible because of risk's nature, risk managers should establish scenario-based risk tiering. While mapping every possible harm scenario from general-purpose AI systems like those that we are concerned with will be difficult or impossible, identifying a core set of representative scenarios to estimate qualitative risks would be a good first step. In this framework, full actualization of selected harmful scenarios would represent the final risk tier, while partial appearance of harms could represent lower risk tiers. Systematic methods for the development of risk scenarios could ensure transparent and effective scenario development. Scenarios should look beyond misuse include mitigations failures, systemic effects, and system or

capability interactions presenting novel risks. Where scenario-based risk tiering is impossible because of the risk's nature, risk managers should establish capability-based tiering.

Even as evaluation and mitigation science advances, measurement uncertainty will persist, making exact risk determination impossible. As such, risk tiers should incorporate safety buffers providing margins of error ensuring borderline risk cases avoid miscategorization into lower risk tiers. These safety buffers should be calibrated to possible measurement error within risk domains, such that less measurement error requires a smaller safety buffer. Furthermore, as the system climbs risk tiers, more substantial safety buffers should be required with higher burdens of proof for shrinking buffers, ensuring more caution is used where risks are more severe. Given potential mitigation failures and increased harms to society, extending safety buffers in these contexts would provide protection against possible harms.

IV. How should the risk tiering process work?

A. When does the risk tiering process need to attach?

Once risk tiers are defined and a classification system established, they must be integrated into an overall risk management framework providing classification inputs and guiding decisionmaking around mitigations, development, and deployment. Choosing who establishes risk tiers and what mechanisms ensure effective risk tiering and classification processes is critical. However, practical questions closer to the technology and its effects must also be answered.

AI systems should be evaluated and sorted into tiers at different points throughout development, training, post-training, and deployment to ensure coverage across the entire AI lifecycle. Initial risk forecasts for a new system can be based on production inputs. Most high-risk models covered by advanced AI risk management follow model scaling laws that allow for reasonable predictions about ultimate capabilities based on compute, data, and architectures used even before pretraining has occurred (though models specialized for dual use domains like biological research and the recent inference or reasoning paradigm make predicting final capability ceilings of a model more difficult).

These initial capability forecasts should sort systems into preliminary risk tiers with associated mitigations. Some mitigations should be enacted immediately determining which preliminary tier the model fits into. For example, cybersecurity mitigations aimed at preventing theft of model weights during the course of pretraining should be

implemented as soon as a covered model begins development. Risk managers should also use capability forecasting to plan other mitigations at later development stages, like post-training mitigations around model refusals. Given a system's initial projected risk and its associated tier, different mitigations can be identified and prepared to reduce risk levels to an acceptable point.

During pretraining, the system should undergo capability evaluations at various checkpoints. This practice is common among leading frontier AI companies and provides a useful input for potential classification or reclassification into risk tiers. Capabilities at training checkpoints will not map directly onto final system capabilities, especially given how capabilities can be enhanced after pretraining using scaffolding and inference. But they provide a check for whether the pre-pretraining risk predictions were reasonable, and whether additional precautions are needed based on demonstrated capabilities exceeding predicted levels.

After pretraining is complete, managers should perform another round of evaluations to determine final base model capabilities in general and in sensitive domains. This capability baseline can then be combined with information about post-training architectures and techniques used to increase performance in different domains and the risk environment to provide a complete risk picture. Evaluations during post-training, capturing gains from inference scaling, for example, will also likely be necessary.

If, during any stage of evaluations, the system presents some risk sufficient for classification into a high risk tier, the AI company should perform mitigations to bring model risks down to acceptable levels. These mitigations should be prepared in advance based on the initial capabilities predictions but may need supplementation if deemed to be insufficient. Many mitigations (like refusal-based mitigations) occur during post-training, and after post-training the system should be re-evaluated to determine whether more mitigations are necessary or whether release can be considered.

Finally, regulators and AI companies should consider when and how to reevaluate released systems when there are substantial changes to the threat environment or post-release capability increases. If system capabilities remain relatively constant over time, reevaluation based on changes in the threat environment can combine existing capability evaluations with new threat information to estimate new risk levels presented by existing models. If that new risk level results in classification into a higher risk tier, then the model should receive new mitigations bringing it back down below the risk threshold it crossed. Changes in the threat environment might include the emergence of a new dangerous and skilled malicious actor or the discovery of a new model jailbreaking technique making it easier for adverse parties to elicit dangerous assistance from existing systems.

Post-release capability increases might also change the risk profiles of existing deployed systems, pushing them into a new tier. Scaffolding and other forms of capability extension, as well as breakthroughs like inference scaling, could mean incapable early models could present later dangers. Eliciting the maximum capabilities from a newly-empowered system quickly enough to prevent those capabilities from being used for harm will challenge risk managers. Regulators and AI companies may fail to predict and evaluate all potential future capability uplift sources, allowing dangerous actors to unlock new dangerous capabilities in deployed models. If so, incident detection or other measures of new dangerous capabilities might be necessary to prevent catastrophic harms. Evaluators should identify methods to group older models to efficiently test how capabilities increases uplift them. For example, models trained on similar amounts of compute and data likely respond to new architectures similarly, but determining whether and how this relationship exists and then selecting a representative model for evaluation after supplementation with the best new post-training techniques would help mitigate this problem.

B. What should risk modeling for tiers look like?

Risk tiering relies on risk modeling to provide inputs used to establish tiers and then sort AI systems into them. Risk modeling will improve as the science of measurement and evaluation advances and as experience facilitates adaptation to the reality of AI development and deployment. However, early observations suggest how risk modeling could improve and to new research directions.

First, capabilities-focused evaluations and risk tiers should be supplemented by more complete modeling of the entire risk ecosystem where possible. Capability evaluations will likely remain the baseline for determining risk, as they identify what the model can do beyond existing systems. However, broader risk ecosystem evaluations could determine which sources present the greatest threats and their actual risk levels. For example, risk managers should supplement system capabilities evaluations with estimations of the likelihood that malicious actors overcome cybersecurity protections against model weight theft or post-deployment misuse mitigations. Adding predictions about what these actors might do with that unlocked capability would fill out a system's true risk profile.

Gathering some inputs for these ecosystem evaluations likely exceeds the capabilities of AI companies and traditional risk practitioners. For example, surveying malicious groups and estimating their AI capabilities may be better tasks for governments than AI labs. Developing information-sharing methods between organizations like the AI Security Institute and frontier companies might be necessary to complete risk

evaluation and modeling processes. However, such methods should be structured to avoid regulatory capture or undue influence.

C. How should mitigations map onto tiers?

Identifying risks and mapping them to tiers based on their levels is only a first step. Once risks are identified, mitigations should be used to reduce them to acceptable levels (e.g., into a lower tier). Companies have developed a substantial set of mitigations to respond to advanced AI's risks. These mitigations can generally be sorted into security mitigations and deployment mitigations, where security mitigations have to do with physical and cybersecurity measures against theft and deployment mitigations directly prevent misuse by intervening in use. Other risks, including loss of control, might require new mitigations used at different places along the development process. Measuring existing mitigations' effectiveness and developing new ones are both important steps to increase risk management robustness and prevent harms.

How should mitigations relate to risk tiers? What mitigations might be necessary given different risks and risk levels? Mitigations are often costly and reduce people's ability to use AIs how they want, so mapping mitigations to risk tiers is an important part of the process of limiting risk while maximizing the benefits of AI.

Two broad approaches to mapping mitigations to risk tiers suggest themselves. The first involves using an abstract risk budget to guide where and how mitigations apply. In this risk budget approach, evaluations establish the overall risks a new system likely poses, quantified as X% chance of Y harm. That risk level guides classification into a tier. Proposed mitigations have been evaluated for how much they reduce risk levels, quantified in the same way as risks. For example, some new refusals mitigation technique might reduce elicitation 10% but at a cost of $500,000 (picking two numbers arbitrarily). This mitigation would be applied to the system's risk profile and evaluated to see if it pushes the system into a lower risk tier. AI companies would decide from their menu of mitigation options which to use so long as they reduce risks to a socially acceptable level. Reporting or outside auditing should be employed to verify this process and that risks were reduced to acceptable levels. Safety cases could also show risks had been reduced below the relevant threshold.[8] This approach would allow AI companies the flexibility to balance the costs of mitigations with risk reduction and apply their specialized knowledge about what would work best with their system. It might also drive innovation in mitigations because reducing the cost (or increasing the effectiveness) of mitigations would allow companies to produce and deploy systems more easily.

[8] *See* Joshua Clymer et al., *Safety Cases: How to Justify the Safety of Advanced AI Systems*, ARXIV (Mar. 18, 2024), https://arxiv.org/abs/2403.10462.

An alternative approach might involve a regulator mandating certain mitigations for systems in particular risk tiers regardless of cost. Such an approach would lack the risk budget's flexibility but might be better while the science of mitigation is less developed and if regulators do not trust AI companies to appropriately balance cost and risk. Some mitigations might be worth mandating regardless of cost, such as hardening physical and cybersecurity to prevent the theft of model weights. If regulators know risk management best practices and can mandate them, as well as update requirements as the state of the art improves, this more definite approach might be preferable to allowing companies to select their preferred mitigations *à la carte*. Which of these two approaches is preferable is partly a political question, but also one where a gold standard could provide guidance.

Finally, if mitigations can be removed or circumvented by attackers (as is currently the case), then extremely high capability systems may fall into unacceptable risk tiers regardless of the mitigations applied. If inherent risk potentials exceed societal tolerances, then these systems should not be deployed unless mitigations can be hardened to render threats inaccessible to adversaries or by mistake.

V. Risk managers should consider benefits once potential harms are understood

Risk tiering should consider benefits from advanced AI systems which would be lost by not deploying them, instead of only considering harms. However, envisioning how to incorporate benefits into the risk management process remains difficult because of lack of evidence in key domains and problems of comparison. Research measuring benefits and determining how to compare them with risks is necessary. A gold standard process for creating risk tiers could facilitate this work by providing a clear framework into which benefits could be translated for comparison with risks. However, even with a gold standard, real problems would remain. The convening discussed these problems and potential paths forward.

Including benefits in risk calculations might entail determining which risk tier a system fits before release and then considering whether its benefits outweigh the need for mitigations or qualify even a high-risk system for release. Considering possible benefits in risk management calculations allows for the inclusion of the risk of losing out on AI's benefits by stopping deployment into the risk analysis, where they would otherwise be neglected.

However, the science of AI risk measurement is immature and measuring possible benefits seems even more challenging. Many benefits of advanced AI are likely to be highly diffuse and difficult to predict. For example, while economic benefits of a system might be measured in productivity or growth increases, the effects of AI on those different measures may be indirect and the uplift provided by a particular system hard to ascertain. Similarly, advanced AI will likely help with lifesaving medical breakthroughs, but predicting how likely a specific system is to help cure cancer seems difficult. Other benefits are relatively incommensurable with risks. Increasing GDP probably improves the quality and extent of human life in most cases, but it is unclear how to compare that relationship to the expected lives lost from a terrorist attack from AI misuse. More abstract benefits from advanced AI, including things like the possible reduction in loneliness from AI conversation partners will be difficult to quantify and compare with risks.

In short, advanced AI will create significant benefits that should be considered in risk management but more work must be done to create a good basis for comparison with risks. A gold standard would likely incorporate benefit considerations as measurement and evaluation improves. However, until risk management develops enough to allow the clear demarcation of risk thresholds and the comparison of risks and benefits, the gold standard should focus on the questions of risk and harm.

VI. Risk governance and risk tiers

Finally, complex questions of governance and comparative advantage shape questions of who should run the risk tiering process and what structures are needed to ensure risk management is done right. At present, AI companies mostly lead risk management, though aided by a growing ecosystem of governmental and third-party evaluators. These AI companies are well-suited to AI risk management because they have significant concentrations of technical expertise and are closest to the technology. However, the key questions of social risk acceptance that underlie risk tiering should be answered by public deliberation and governments, not private companies. AI companies should not be able to simply dictate to society how much risk they are exposed to by the development of new technologies without oversight and public intervention. More directly, companies might avoid proper risk management if it undermines their ability to compete in the market. This pressure to cut corners will be especially severe around the top systems which also present the most risk. Third party auditors and government mandates might make corner-cutting less likely, but will not solve it. Furthermore, regulatory capture of auditors and governmental authorities must be guarded against.

Developing independent institutions with expertise in AI risk assessment and management and ensuring that auditors can evaluate internal company risk management practices will become increasingly essential. The exact mix of corporate, governmental, and third-party cooperation that would best address the growing risks from advanced AI remains uncertain, but there was consensus in the convening that the current mix is likely suboptimal.

Risk governance in advanced AI could learn from existing governance approaches in other safety-critical sectors, which have substantive and well-defined governance roles and functions that could inform AI risk management. Further research and collaboration with risk managers from other sectors could flesh out how better risk governance within and outside advanced AI companies. Some degree of governmental or industry-level standardization or regulation might be helpful in ensuring that these governance practices are adopted by the developers of advanced AI around the world.

Conclusion

The field of risk management for advanced AI remains immature, and many key questions about structuring and implementing necessary practices remain. However, creating a gold standard for risk tiering would provide a focal point for governments, companies, and publics to discuss what kind of risk management is best and who should do it. A gold standardization process would also help practitioners, regulators, and researchers identify where the science of risk management is falling short and what improvements are needed.

Risk tiers provide a useful framework for breaking down the risks that are presented by advanced AI into comprehensible parts, allowing regulators, risk managers, and the broader public to understand emerging AI risks that might soon affect their lives. As these risks become more significant, this framework will provide a tool for ensuring the mitigation of potential harms from advanced AI systems, allowing society to enjoy the benefits that they create.