Microsoft

Frontier Governance Framework

Contents

1. Evidence-based management of frontier risk

Microsoft's Frontier Governance Framework manages potential national security and at-scale public safety risks that could emerge as AI models increase in capability. The framework has its genesis in the voluntary Frontier AI Safety Commitments that Microsoft and fifteen other AI labs made in May 2024 with the support of governments from around the world.

The framework serves as a monitoring function, tracking the emergence of new and advanced AI model capabilities that could be misused to threaten national security or pose at-scale public safety risks. It also sets out a process for assessing and mitigating these risks so that AI models can be deployed in a secure and trustworthy way. In crafting Microsoft's framework, we have sought to advance an evidence-based approach to managing frontier risk that is:

- **Integrated with Microsoft's broader AI governance program.** Effective management of AI risks requires a comprehensive approach. While model-level assessments and interventions to stay ahead of frontier risks are necessary, they alone are not sufficient. This framework is integrated with Microsoft's broader AI governance program, which sets out a comprehensive risk management program that applies to all AI models and systems developed and deployed by Microsoft.

- **Focused on high-risk capabilities.** The framework operationalizes state-of-the-art risk management practices for a set of risks that are connected to high-impact model capabilities. We are focused on capabilities that could emerge in the short-to-medium term. Longer-term or more speculative capabilities are the subject of ongoing research that we and many others across industry and academia are invested in. Evaluations and mitigations under the framework target capability-related risks, complementing Microsoft's broader AI governance program that manages a broader set of risks, including more culturally contextual risks that are heavily shaped by use case and deployment environments, as well as laws and norms that vary across regions.

- **Targeted and proportional.** The framework monitors Microsoft's most capable AI models for leading indicators of high-risk capabilities and triggers deeper assessment if leading indicators are observed. As and when risks are identified, proportional mitigations are applied so that risks are kept at an appropriate level. This approach provides confidence that highly capable models are identified before relevant risks emerge, without imposing requirements on less capable models that are governed by Microsoft's broader AI governance program. The framework is built on a foundation of full-stack security, advancing comprehensive protections for key assets.

- **Flexible and durable.** This is the first version of a framework that we expect will be revised significantly over time to reflect ongoing advances in the capabilities of AI technologies and in the science and practice of AI risk management. We have crafted it with an eye

toward flexibility and durability, outlining processes to which best practice evaluations and risk management techniques can be applied as they mature. We look forward to further collaboration across industry, government, and civil society to spur continued progress.

2. Capabilities and risks

Tracked high-risk capabilities

This framework tracks the following capabilities that we believe could emerge over the short-to-medium term and threaten national security or pose at-scale public safety risks if not appropriately mitigated. In formulating this list, we have benefited from the advice of both internal and external experts.

- **Chemical, biological, radiological, and nuclear (CBRN) weapons.** A model's ability to provide significant capability uplift to an actor seeking to develop and deploy a chemical, biological, radiological, or nuclear weapon.

- **Offensive cyberoperations.** A model's ability to provide significant capability uplift to an actor seeking to carry out highly disruptive or destructive cyberattacks, including on critical infrastructure.

- **Advanced autonomy.** A model's ability to complete expert-level tasks autonomously, including AI research and development.

This framework assesses Microsoft's most advanced AI models for signs that they may have these capabilities and, if so, whether the capability poses a low, medium, high, or critical risk to national security or public safety (more detail in Appendix I). This classification then guides the application of appropriate and proportionate mitigations so that a model's risks remain at an acceptable level.

AI technology continues to develop rapidly, and there remains uncertainty over which capabilities may emerge and when. We continue to study a range of potential capability-related risks that could emerge, conducting ongoing assessment of the severity and likelihood of these risks. We then operationalize the highest-priority risks through this framework. We will revisit our list of tracked capabilities frequently, ensuring it remains up to date in light of technological developments and improved understanding of model capabilities, risks, and mitigations.

Integrated governance

In addition to high-risk capabilities, a broader set of risks are governed when Microsoft develops and deploys AI technologies. Under Microsoft's comprehensive AI governance program, frontier models—as well as other models and AI systems—are subject to relevant evaluation, with mitigations then applied to bring overall risk to an appropriate level. Information on model or system performance, responsible use, and suggested system-level evaluations is shared with downstream actors integrating models into systems, including external system developers and deployers and teams at Microsoft building models. Appropriate information sharing is important to facilitate mitigation of a broader set of risks, many of which are heavily shaped by use case and deployment context as well as laws and norms that vary across jurisdictions. While different risk profiles may thus inform different mitigation strategies, Microsoft's overall approach of mapping, measuring, and mitigating risks, including through robust evaluation and measurement, applies consistently across our AI technologies. Our efforts to assess and mitigate risks related to this framework's tracked capabilities benefit from this broadly applied governance program, which is continuously improved. The remainder of this framework addresses more specifically the assessment and mitigation of risks relating to the framework's tracked capabilities.

3. Evaluation and assessment

Monitoring for leading Indicators of high-risk capabilities

Through the processes described in this framework, Microsoft's most advanced models are assessed for leading indicators of the framework's high-risk capabilities. This is done using state-of-the-art benchmarks[1] for the following advanced general-purpose capabilities, identified as precursors to high-risk capabilities:

- General reasoning
- Scientific and mathematical reasoning
- Long-context reasoning
- Spatial understanding and awareness
- Autonomy, planning, and tool use
- Advanced software engineering

A leading indicator assessment is run on any model that teams at Microsoft are optimizing for frontier capabilities or that Microsoft otherwise expects may have frontier capabilities.[2] In addition, any model pre-trained using more than 10^26 FLOPs is subject to leading indicator assessment, given the (imperfect) correlation between pre-training compute and performance. This pre-training compute trigger will be revisited over time given improvements in training efficiency and as new approaches to enhancing model capabilities outside of pre-training are further developed, including techniques leveraging test-time compute.

The leading indicator assessment is run during pre-training, after pre-training is complete, and prior to deployment to ensure a comprehensive assessment as to whether a model warrants deeper inspection. This also allows for pause, review, and the application of mitigations as appropriate if a model shows signs of significant capability improvements. Models in scope of this framework will undergo leading indicator assessment at least every six months to assess progress in post-training capability enhancements, including fine-tuning and tooling. Any model demonstrating frontier capabilities is then subject to a deeper capability assessment to provide strong confidence about whether it has a tracked capability and to what level, informing mitigations.

The leading indicator assessment helps provide early warning that a model may have a tracked capability. It also ensures the framework is targeted at only those models that warrant deeper inspection and related mitigations. This is important given the wide range of models being developed, used, and deployed by Microsoft, many of which are optimized for characteristics such as speed and efficiency, rather than advanced capabilities. All models developed by Microsoft are subject to Microsoft's broader AI governance program.

Assessing performance based on advanced general-purpose benchmarks provides the most useful leading indicator of high-risk capabilities currently. As evaluations mature further, we intend to move this leading indicator screening toward a more direct assessment of tracked capabilities such as CBRN weapon development, offensive cyberoperations, and advanced autonomy.

[1] For a benchmark to be included in our suite of leading indicator assessments it must: 1) have low saturation (i.e., the best performing models typically score lower than 70%); 2) measure an advanced capability, for example, mathematical reasoning, rather than an application-oriented capability like financial market prediction; and 3) have a sufficient number of prompts to account for non-determinism in model output.

[2] Frontier capabilities are defined as a significant jump in performance beyond the existing capability frontier in one advanced general-purpose capability or beyond frontier performance across the majority of these advanced general-purpose capabilities.

Deeper capability assessment

Deeper capability assessment provides a robust indication of whether a model possesses a tracked capability and, if so, whether this capability is at a low, medium, high, or critical risk level, informing decisions about appropriate mitigations and deployment. We use qualitative capability thresholds to guide this classification process as they offer important flexibility across different models and contexts at a time of nascent and evolving understanding of frontier AI risk assessment and management practice. We lay out further detail on these capability thresholds in Appendix I. Deeper capability assessment involves the following:

- **Capability evaluation:** This involves robust evaluation of whether a model possesses tracked capabilities at high or critical levels, including through adversarial testing and systematic measurement using state-of-the-art methods. Evaluations are documented in a consistent fashion setting out the capability being evaluated, the method used, and evaluation results. This evaluation also includes a statement on the robustness of the evaluation

method used and any concerns about the effectiveness or validity of the evaluation. As appropriate, evaluations involve qualified and expert external actors that meet relevant security standards, including those with domain-specific expertise.

- **Capability elicitation:** Evaluations include concerted efforts at capability elicitation, i.e., applying capability enhancing techniques to advance understanding of a model's full capabilities. This includes fine-tuning the model to improve performance on the capability being evaluated or ensuring the model is prompted and scaffolded to enhance the tracked capability—for example, by using a multi-agent setup, leveraging prompt optimization, or connecting the model to whichever tools and plugins will maximize its performance. Resources applied to elicitation should be extrapolated out to those available to actors in threat models relevant to each tracked capability.

- **Holistic risk assessment:** The results of capability evaluation and an assessment of risk factors external to the model then inform a determination as to whether a model has a tracked capability and to what level. This includes assessing the impact of potential system-level mitigations and societal and institutional factors that can impact whether and how a hazard materializes. This holistic risk assessment also considers the marginal capability uplift a model may provide over and above currently available tools and information, including currently available open-weights models.

- **Timing of deeper capability assessment:** After the first deeper capability assessment, we will conduct subsequent deeper capability assessments on a periodic basis, and at least once every six months.

Pre-mitigation capability assessment: The results of the deeper capability assessment are used to assign a model a pre-mitigation score of low, medium, high, or critical for each tracked capability on the basis of the framework's capability thresholds (see Appendix I). Models assessed as posing low or medium risk may be deployed with appropriate safeguards as outlined below. Models assessed as having a high or critical risk are subject to further review and safety and security mitigations prior to deployment.

4. Mitigations

Security measures

Securing frontier models is an essential precursor to safe and trustworthy use and the first priority of this framework. Any model that triggers leading indicator assessment is subject to robust baseline security protection. Security safeguards are then scaled up depending on the model's pre-mitigation scores, with more robust measures applied to models with High and Critical risk levels.

As Microsoft operates the infrastructure on which its models will be trained and deployed, we adopt an integrated full-stack approach to the security of frontier models, implementing safeguards at the infrastructure, model, and system layers. Security measures will be tailored to the specifics of each model, including its capabilities and the method by which it is made available and integrated into a system, so that the marginal risks a model may pose are appropriately addressed.

We expect scientific understanding of how to best secure the AI lifecycle will advance significantly in the coming months and years and will continue to contribute to, and apply, security best practices as relevant and appropriate. This includes existing best practice defined in leading standards and frameworks, such as NIST SP 800-53, NIST 800-218, SOC 2, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, and Deploying AI Systems Securely, as well as industry practices, including from the Frontier Model Forum. Security safeguards are scaled up depending on the model's pre-mitigation scores, with more robust measures applied to models with high and critical risk levels.

Models posing high-risk on one or more tracked capability will be subject to security measures protective against most cybercrime groups and insider threats. Examples of requirements for models having a high-risk score include:

- **Restricted access,** including access control list hygiene and limiting access to weights of the most capable models other than for core research and for safety and security teams. Strong perimeter and access control are applied as part of preventing unauthorized access.

- **Defense in depth across the lifecycle,** applying multiple layers of security controls that provide redundancy in case some controls fail. Model weights are encrypted.

- **Advanced security red teaming,** using third parties where appropriate, to reasonably simulate relevant threat actors seeking to steal the model weights so that security safeguards are robust.

Models posing critical risk on one or more tracked capability are subject to the highest level of security safeguards. Further work and investment are needed to mature security practices so that they can be effective in securing highly advanced models with critical risk levels that may emerge in the future. Appropriate requirements for critical risk level models will likely include the use of high-trust developer environments, such as hardened tamper-resistant workstations with enhanced logging, and physical bandwidth limitations between devices or networks containing weights and the outside world.

Safety mitigations

We apply state-of-the-art safety mitigations tailored to observed risks so that the model's risk level remains at low or medium once mitigations have been applied. We will continue

to contribute to research and best-practice development, including through organizations such as the Frontier Model Forum, and to share and leverage best practice mitigations as part of this framework. Examples of safety mitigations we utilize include:

- **Harm refusal,** applying state-of-the-art harm refusal techniques so that a model does not return harmful information relating to a tracked capability at a high or critical level to a user. This is an area of active research, and we continue to invest in and apply robust harm refusal techniques as they are developed.

- **Deployment guidance,** with clear documentation setting out the capabilities and limitations of the model, including factors affecting safe and secure use and details of prohibited uses. This documentation will also include a summary of evaluation results, the deeper capability assessment, and safety and security mitigations. For example, the documentation could outline specific capabilities and tasks that the model robustly fails to complete which would be essential for a high or critical risk rating. This includes identification of residual risks, for example relating to bias or discrimination, that are sensitive to deployment context, subject to laws and norms that vary across jurisdictions, and are likely to require further evaluation and mitigation on the part of those integrating the model into a system and deploying it.

- **Monitoring and remediation,** including abuse monitoring in line with Microsoft's Product Terms and provide channels for employees, customers, and external parties to report concerns about model performance, including serious incidents that may pose public safety and national security risks. We apply mitigations and remediation as appropriate to address identified concerns and adjust customer documentation as needed. Other forms of monitoring, including for example, automated monitoring in chain-of-thought outputs, are also utilized as appropriate. We continue to assess the tradeoffs between safety and security goals and legal and privacy considerations, optimizing for measures that can achieve specific safety and security goals in compliance with existing law and contractual agreements.

- **Phased release, trusted users, and usage studies,** as appropriate for models demonstrating novel or advanced capabilities. This can involve sharing the model initially with defined groups of trusted users with a view to better understanding model performance while in use before general availability. We are also progressing work to further study models when in use and assess the real-world effectiveness of mitigations, while upholding stringent levels of privacy and confidentiality and benefiting from external expertise where appropriate.

Post-mitigation capability assessment and safety buffer: Following application of safety and security mitigations, the model will be re-evaluated to ensure capabilities are rated low or medium and, if not, to guide further mitigation efforts. If, during the implementation of this framework, we identify a risk we cannot sufficiently mitigate, we will pause development and deployment until the point at which mitigation practices evolve to meet the risk.

5. Governance

Risk-informed deployment decision

Documentation regarding the pre-mitigation and post-mitigation capability assessment will be provided to Executive Officers responsible for Microsoft's AI governance program (or their delegates) along with a recommendation for secure and trustworthy deployment setting out the case that: 1) the model has been adequately mitigated to low or medium risk level, 2) the marginal benefits of a model outweigh any residual risk and 3) the mitigations and documentation will allow the model to be deployed in a secure and trustworthy manner.

The Executive Officers (or their delegates) will make the final decision on whether to approve the recommendation for secure and trustworthy deployment. The Executive Officers (or their delegates) are also responsible for assessing that the recommendation for secure and trustworthy deployment and its constituent parts have been developed in a good faith attempt to determine the ultimate capabilities of the model and mitigate risks.

Information about the capabilities and limitations of the model, relevant evaluations, and the model's risk classification will be shared publicly, with care taken to minimize information hazards that could give rise to safety and security risks and to protect commercially sensitive information.

Oversight and reporting channels

This framework is subject to Microsoft's broader corporate governance procedures, including independent internal audit and board oversight. Microsoft employees have the ability to report concerns relating to this framework and its implementation, as well as AI governance at Microsoft more broadly, using our existing concern reporting channels, with protection from retaliation and the option for anonymity.

Updates to this framework

We will update our framework to keep pace with new developments. Every six months, we will have an explicit discussion on how this framework may need to be improved. We acknowledge that advances in the science of evaluation and risk mitigation may lead to additional requirements in this framework or remove the need for existing requirements. Any updates to our practices will be reviewed by Microsoft's Chief Responsible AI Officer prior to their adoption. Where appropriate, updates will be made public at the same time as we adopt them.

6. Collaboration and continuous learning

This framework adopts an outcome-oriented approach to facilitate flexibility and innovation in AI risk management, which remains a nascent area of rapidly evolving research and practice. As with the history and evolution of risk management for other complex technologies, we anticipate AI risk management practice to develop markedly over time as we advance a deeper scientific understanding of AI and its risks.

Microsoft will prioritize ongoing contributions to this work and expand its collaboration with government, industry, and civil society, including through organizations like the Frontier Model Forum, to solve the most pressing challenges in AI risk management. We have identified a need for accelerated best practice development for risk assessment and evaluation methodologies. This includes more holistic risk assessment techniques that appropriately factor in considerations outside of model capabilities, such as system-level mitigations, as well as societal and institutional factors external to the AI system that can impact whether and how a hazard materializes and the scale of its impact. This type of holistic assessment will be needed if countries are to meaningfully calibrate risk thresholds and related governance requirements.

We also highlight the value of learning from experts outside of AI, including those with expertise in measurement science and in scientific domains like chemistry and biology, as well as those with knowledge of managing the risks of other complex technologies. We will continue to share information and lessons from our governance efforts, taking care to do so in a way that minimizes the chance of inadvertently increasing risk, including by being judicious about sharing sensitive safety- and security-relevant information.

Appendix I – Capability thresholds

Capability thresholds and corresponding risk levels for the framework's tracked capabilities are set out below.

CBRN weapons

Risk level	Capability threshold	Deployment requirements
Low	The model either does not respond to queries related to developing and delivering CBRN attacks or responds with information that is already publicly available.	Deployment allowed in line with Responsible AI Program requirements.
Medium	The model provides some assistance increasing the ability of an existing expert (PhD level education in related fields) to develop a known CBRN threat.	Deployment allowed in line with Responsible AI Program requirements.
High	The model provides a meaningful uplift to an expert's (PhD level education in related fields) ability to develop and deliver a novel CBRN threat. The model provides a meaningful uplift to a medium-skilled actor's (e.g., STEM education) ability to develop and deliver a known CBRN threat.	Further review and mitigations required.
Critical	The model provides a meaningful uplift to an expert's ability to develop a highly dangerous novel threat or significantly lowers the barriers to a low-skilled actor developing and delivering a known CBRN threat.	Further review and mitigations required.

Offensive cyberoperations

Risk level	Capability threshold	Deployment requirements
Low	The model supports gathering of publicly available information related to cyber security threats, including through question and answer or content generation, but cannot meaningfully plan or execute a cyberattack.	Deployment allowed in line with Responsible AI Program requirements.
Medium	The model increases the ability of a low-skilled actor to conduct low-level spoofing, phishing, or social engineering attacks by partially automating discrete tasks related to vulnerability exploitation like generic target identification and planning (but not execution).	Deployment allowed in line with Responsible AI Program requirements.
High	The model provides meaningful uplift to a low-to-medium skilled actor's ability to create and conduct highly disruptive or destructive cyber-attacks, including on critical infrastructure, for example, through discovering novel zero-day exploit chains or developing complex malware or other tactics, techniques, and procedures.	Further review and mitigations required.
Critical	The model provides a meaningful uplift to a low-skilled actor's ability to identify and exploit major vulnerabilities or enables a well-resourced and expert actor to develop and execute novel and effective strategies against hardened targets.	Further review and mitigations required.

Advanced autonomy

Risk level	Capability threshold	Deployment requirements
Low	The model can complete a small number of basic tasks, including software engineering tasks that take a human less than one hour.	Deployment allowed in line with Responsible AI Program requirements
Medium	The model can autonomously complete more complex tasks, including software engineering tasks equivalent to a few hours of human labor, but requires human intervention to correct for complex error conditions or changes to the operating environment.	Deployment allowed in line with Responsible AI Program requirements
High	The model can autonomously complete a range of generalist tasks equivalent to multiple days' worth of generalist human labor and appropriately correct for complex error conditions, or autonomously complete the vast majority of coding tasks at the level of expert humans.	Further review and mitigations required
Critical	The model can fully automate the AI R&D pipeline at a fraction of human labor costs, majorly accelerating AI R&D.	Further review and mitigations required

Appendix II – Change log

8 February 2025 – First version