OpenAI

Frontier Governance Framework

Contents

1 Introduction

OpenAI's mission is to ensure that artificial general intelligence (AGI) benefits all of humanity. To pursue that mission, we are committed to safely developing and deploying highly capable AI models, which create significant benefits and also bring new risks. We build for safety at every step and share our learnings so that society can make well-informed choices to manage new risks from frontier AI.

We are committed to transparency about how our safety practices align with emerging global legal and governance frameworks. To that end, we have developed this Frontier Governance Framework (FGF) to explain how we fulfill our regulatory obligations and to document our current technical and organizational processes for systemic risk assessment and mitigation across key risk categories, including cybersecurity, CBRN (chemical, biological, radiological, and nuclear), harmful manipulation, and loss of control risks.

Our FGF is designed to meet the baseline legal requirements of various frontier AI laws, including:

- Under California's Transparency in Frontier AI Act (TFAIA), this FGF is our Frontier AI Framework, documenting OpenAI's technical and organizational protocols to manage, assess, and mitigate catastrophic risks, as defined under the TFAIA.
- Under the European Union's General-Purpose AI Code of Practice (the EU CoP), this FGF serves as our publicly available summary of OpenAI's Safety & Security Framework, describing how we assess and mitigate systemic risks and ensure adequate cybersecurity protection for models covered under Regulation (EU) 2024/1689 (the EU AI Act).

The FGF overlaps in some areas with our existing Preparedness Framework (PF). The PF and FGF together describe OpenAI's practices, and we will continue to use and evolve the PF to define and operationalize OpenAI's own approach to managing the most serious risks from advanced AI systems, including in situations where our internal practices go beyond current legal requirements. For example, because the PF is intended both to advance the science of managing severe risks of advanced AI systems and to reflect OpenAI's evolving practices, it may use different definitions of catastrophic risk and does not depend on specific legal compute thresholds like the FGF.

The FGF covers frontier models as defined under the TFAIA and "general-purpose models with systemic risk" under the EU AI Act. In this document, when we refer to "systemic" risks, we mean both catastrophic risks under the TFAIA and systemic risks under the EU AI Act. The systemic risk assessment and mitigation processes described here apply to covered models that OpenAI has deployed externally, and in some cases internally with respect to risks resulting from circumventing oversight mechanisms. We expect our approach will continue to evolve, and we will update this FGF as our processes and regulatory requirements change.

Our approach to mitigating frontier AI safety risks is informed by emerging national and international AI risk management standards, including ISO 42001, the NIST AI Risk Management Framework, and frontier safety laws in relevant jurisdictions. Our approach also draws on the proposal for Responsible Scaling Policies first introduced by METR as well as industry best practices.

2 Systemic risk assessment & mitigation

2.1 Systemic risk identification

OpenAI has implemented a variety of structured processes to identify systemic risks stemming from our frontier models and to develop risk scenarios and threat models through which these systemic risks may develop or materialize.

This FGF's definition of systemic risk includes foreseeable and material risks of severe harm

from the development, storage, use, or deployment of our most advanced frontier models, including risks that a model will materially contribute to greater than 50 fatalities or $1 billion of property damages or losses arising from a single incident.

We evaluate whether frontier capabilities create a risk of severe harm through a holistic risk assessment process. This process draws on our own internal research and signals, and, where appropriate, incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government, the European Commission and other government agencies, as well as relevant legal and policy guidance and requirements.

Based on this analysis, this FGF definition currently addresses the following systemic risk categories:

Cyber offense	Risks of enabling large-scale sophisticated cyber-attacks, including on critical systems, such as by significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations, e.g. through automated vulnerability discovery, exploit generation, operational use, and attack scaling.
Chemical, biological, radiological & nuclear (CBRN)	Risks of enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents, such as by significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.
Harmful manipulation	Risks stemming from the strategic distortion of human behavior, including the use of model capabilities to conduct influence operations, election interference, or other coordinated campaigns to manipulate public opinion or undermine democratic processes.
Loss of control	Risks stemming from the inability to reliably direct, modify, or shut down a model, including evading the controls of a model developer or user, or autonomous conduct that, if conducted by a human, would constitute a crime of murder, assault, extortion, or theft.

2.2 Systemic risk analysis

Our risk assessment process may take place throughout a model's lifecycle, including during development and after deployment. Risk assessments draw on various sources, including a range of model evaluations, model-independent information such as external research and data, literature reviews, market analyses, consultation with internal and external experts, insights from other deployed models, post-release monitoring, and, if applicable, investigations of serious incidents or critical safety incidents.

We conduct systemic risk modelling that informs how we evaluate systemic risks before deployment. Currently, we estimate the severity and probability of harm of risks related to CBRN, cyber offense, and loss of control under this FGF. We are still in the early stages of developing an approach for assessing risks from harmful manipulation. Many of the risks stemming from harmful manipulation, such as the use of model capabilities to conduct influence operations, are best addressed through system level mitigations, such as post-deployment monitoring, rather than model evaluations before deployment.

We employ a risk assessment process that includes state-of-the-art evaluations to test the specific threats and risk scenarios identified through our threat modeling, assess a model's capabilities against systemic risk thresholds, and understand and assess the effectiveness of our safeguards and mitigations. We estimate systemic risks where appropriate and practicable using risk tiers that concretely describe things an AI system might be able to help someone do or might be able to do on its own that could meaningfully increase risk of severe harm.

2.3 Systemic risk acceptance determination

We employ science-backed evaluations that inform whether a covered model presents systemic risks that remain within acceptable levels and help us assess residual risk. Our assessment recognizes that progress in model scaffolding and elicitation techniques can add uncertainty, and we accordingly treat one-time capability elicitations in a frontier model as a lower-bound, rather than a ceiling, on capabilities that may emerge in real world use and misuse. Our consideration of residual risk takes account of the scale and probability of harm, and is evaluated in light of the sufficiency of implemented safeguards proportionate to the level of risk. In assessing whether to accept residual risk, we review the risk tiers for each systemic risk category and incorporate appropriate safety margins. For example, out of an abundance of caution we have treated models as crossing a capability threshold in circumstances where we are unable to rule out that a new threshold had been reached, even in the absence of direct evidence that it has occurred.

Before deployment, models in scope of this Framework undergo a suite of scalable evaluations designed to measure proxies that approximate whether a model's systemic risks reach a new risk tier based on levels of performance that have been pre-determined to indicate that a deployment may have reached a capability threshold. Some evaluations can help rule out that a model has reached a new capability threshold, but passing them does not establish that the threshold has actually been reached.. When analyzing whether a threshold has been reached, we also take into account additional evidence to validate the findings of evaluations, such as human expert red-teaming, expert consultations, and resource-intensive third party evaluations.

The determination that a threshold has or has not been reached is informed by these results, and also reflects a holistic judgment based on the totality of available evidence, such as information about the methodological robustness of evaluation results.

2.4 Risk tiers

Cyber offense

For cyber offense systemic risks, we have established a tier system that seeks to quantify model capabilities against cybersecurity threats and incorporates risk estimates and other suitable metrics. We use these measurable thresholds for decision-making related to offensive cyber capabilities. The system consists of the following tiers:

Tier	Description	Examples
Tier 1	The model provides assistance equivalent to publicly available resources and/or may provide material technical assistance to active cyber operations using known attack methodologies.	Requests for publicly documented exploit information, generic attack methodology questions, coordinated campaigns affecting multiple organizations, operational security guidance extending campaign longevity
Tier 2	The model provides substantial capability uplift to small organizations by removing existing bottlenecks to scaling cyber operations, including by automating end-to-end cyber operations against reasonably hardened targets or by automating the discovery and exploitation of operationally relevant vulnerabilities.	Demonstrated abilities to automate and scale the volume of existing cyberattacks, and is capable of exhibiting technical novelty in the discovery and exploitation of operationally relevant vulnerabilities, conducting long-running stealthy or short-running aggressive, goal oriented campaigns, and demonstrates consistency in operations to scale damage or avoid discovery.
Tier 3	A tool-augmented model can identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems without human intervention OR the model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high-level desired goal.	Autonomous discovery and exploitation of unknown vulnerabilities, self-directed campaign orchestration, resilience and adaptation to hardened security defenses, sustained end-to-end operations without human intervention.

CBRN

For CBRN risks, we have established a tier system that quantifies model capabilities against weapons and threat development uplift metrics. We principally build safeguards against biological and chemical threats, and we prioritize biological capability evaluations due to the higher potential severity of biological threats. With respect to nuclear and radiological threats, we note that information and expertise needed to actually create a working nuclear weapon is significant, and classified. Further, there are significant physical barriers to success, like access to fissile material, specialized equipment, and ballistics. Because of the significant resources required and the legal controls around information and equipment, nuclear weapons development cannot be fully studied outside of a classified context. We will continue to research and develop threat models for nuclear and radiological risks and engage with national security stakeholders on how best to assess these risks.

Tier	Description	Examples
Tier 1	The model's capabilities provide minimal advantage compared to publicly available information or search engines, or provides minor uplift or lowers barriers to experts.	Aggregates publicly available information within distinct domains, identifies relevant scientific literature, provides general information that requires expert interpretation to be actionable, troubleshoots common failure modes without prompting, provides step-by-step procedural guidance with specific parameters.
Tier 2	The model can provide meaningful counterfactual assistance (relative to unlimited access to baseline of tools available in 2021) to novice actors (anyone with a basic relevant technical background) that enables them to create known biological or chemical threats.	Provides accurate, detailed, and contextually relevant information that is similar to direct interactions with subject matter experts, that would be difficult for a non-specialist to obtain via open internet or manuals, and that enables accurate and successful troubleshooting.

Tier	Description	Examples
Tier 3	The model can enable an expert to develop a highly dangerous novel threat vector (e.g., comparable to a novel CDC Class A biological agent) OR the model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel biological threat without human intervention.	Discovery of novel agents, pathways, autonomously executes full design-acquisition-production-deployment pipeline.

Harmful manipulation

OpenAI's approach to harmful manipulation remains exploratory due to its nascency as a systemic risk area. This risk tier is subject to further research and may be substantially changed over time. As described above, OpenAI believes that these risks may be best addressed through system level mitigations, such as post-deployment monitoring, rather than model evaluations before deployment.

Loss of control

Loss of control refers to risks arising from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-improvement, model deception, or autonomous self-improvement. Loss of control could result in a model evading the controls of its developer or a user. We have established a tier system that describes model capabilities in light of a model's level of autonomy, ability to deceive, and the potential to take unsanctioned or unauthorized action. Outside of risks related to AI self-improvement, these risk tiers remain exploratory and may evolve substantially.

Tier	Description	Examples
Tier 1	The model can reliably complete complex, multi-step tasks but lacks the execution of abilities approaching expert humans. The model demonstrates basic situational awareness when elicited and executes deceptive behaviors in narrow, highly prompted scenarios.	Reliably completes well-defined coding tasks can understand it is being evaluated if prompted and may subtly underperform when instructed.
Tier 2	The model's impact is broadly equivalent to an expert human executing complex, sustained tasks, but may still have relative strengths and weaknesses. The model demonstrates the capability to reliably evade detection across a variety of evaluation methods, including the capability to evade chain of thought monitoring.	Reliably completes complex end-to-end software projects, can insert subtle flaws or hidden vulnerabilities that survive expert review, can identify and exploit gaps in monitoring systems if incentivized to do so.
Tier 3	The model is superior to the most expert humans in executing most complex projects and can operate autonomously for extended, sustained periods of time. Demonstrates nuanced situational awareness and stealth such that monitoring the model and its chain of thought cannot reliably detect or rule out evasion of human control.	Exceeds the performance of the world's leading experts across diverse, open-ended, complex projects, avoids detection while operating detailed, multi-step plans, can acquire resources and expand capacity against active countermeasures.

2.5 Safety mitigations

OpenAI implements appropriate safety mitigations to address the systemic risks stemming from our models, taking into account the model's release and distribution strategy. OpenAI tailors these measures to the capability of a model and its deployment, as appropriate, to mitigate systemic risks to acceptable levels. If residual risks associated with the model exceed acceptable risk levels, the model is not deployed unless additional mitigation measures are implemented that sufficiently minimize risk. We rely on a variety of techniques to identify whether additional measures are needed, including post-deployment threat intelligence monitoring, post-launch monitoring tools (classifiers, automated detection, human review, and investigations), consultation with experts, and other techniques.

Where the residual risk falls within acceptable levels, taking into account appropriate safety margins, the model may be approved for continued development and internal use (where applicable), and may be deployed under this Frontier Governance Framework. We document a justification for why the systemic risks stemming from the model are acceptable, including reasonably foreseeable conditions under which the justification may no longer hold, as informed by recommendations from OpenAI's Safety Advisory Group and external experts, as available and appropriate.

2.6 Critical safety incident identification and response

OpenAI maintains an AI Safety Incident Response Plan (AIRP) that outlines OpenAI's plan for identifying and responding to AI safety incidents. The AIRP has a broad scope and uses definitions optimized for operational decision-making, covering a range of different types of safety incidents that may arise, including those that are reportable under applicable frontier safety laws and regulations, including critical safety incidents under the TFAIA. We also maintain a Cybersecurity Incident Response Plan that may cover certain cybersecurity incidents that are reportable.

We have measures in place to monitor for and report on potential AI safety incidents discovered by both internal and external actors. Potential AI safety incidents are triaged, investigated, escalated, and remediated as appropriate according to the procedures described in the AIRP. These incidents are in turn analyzed to determine whether they meet the criteria for external reporting under applicable laws and regulations.

Detection and triage

A potential AI safety incident may be detected through various channels, including automated monitoring, employee escalation, end-user feedback, (including support tickets and external reporting forms), notification from regulators or the press, and review of on- or off-platform activity. Once a potential AI safety incident is identified, the AIRP dictates procedures for assessing whether the event qualifies as an AI safety incident and for determining severity and notifying appropriate internal teams.

Investigation, mitigation, and response

OpenAI maintains response teams that investigate and, if necessary, take action to mitigate and contain incidents. The investigation includes determining the root cause, scope, and impact of the incident.

Once the investigation has been completed, OpenAI takes steps to implement longer-term solutions to address root causes. As part of our commitment to preventing similar future incidents, we may also conduct a retrospective if needed to document key learnings and address open action items.

External reporting

As part of our response, we analyze whether we have reporting obligations relating to the incident under applicable laws and regulations or if any other type of external outreach is advisable, including under any voluntary commitments. If an incident is determined to be reportable, we will gather relevant information from the investigation and remediation phase for reporting to appropriate authorities within the required deadlines.

3 Security risk management

OpenAI maintains an Information Security and Privacy Program aligned with ISO 27001, 27017, 27018, and 27701, and supported by SOC 2 Type II evaluations. The program is structured to protect critical assets, including model weights, training data, and customer data, from unauthorized access, disclosure, or compromise. It applies a risk-based approach to allocate resources effectively, and promotes security awareness across the organization. Continuous monitoring and improvement ensure that controls adapt to changing risks and operational needs.

Security for our frontier models is structured around layered controls addressing network, device, and personnel risks, along with protections for sensitive model parameters. Access is tightly managed, environments are continuously monitored, and independent assessments validate effectiveness. This approach ensures that OpenAI can advance AI development while maintaining the confidentiality, integrity, and availability of its critical assets.

We implement appropriate security mitigations to meet the goals described above. A non-exhaustive list of such mitigations includes:

Protection of unreleased model weights	Model weights are protected through encryption (at rest and in transit), continuous monitoring, and access controls (e.g., multi-factor authentication, multi-party approval, and detailed logging). Physical security of hosting infrastructure is protected by access controls and inspections.
Hardening interface-access to unreleased model parameters	Interface access to model parameters is limited to authorized personnel, with access controls that are subject to regular review. Access is rate limited and subject to monitoring, and provisioning of access may be monitored and logged.
Insider threats	Personnel (employees and contractors) are screened and subject to regular training. Internal monitoring for anomalous activity is used for early identification of potential risks. Model execution is sandboxed, with restricted egress by default.
Security assurance	Security controls are validated through internal and external assessments, including red teaming, penetration testing, vulnerability scanning, SOC 2 Type II audits, and ISO 27001 certification. OpenAI also maintains vulnerability disclosure programs and 27/7/365 incident response capabilities.

4 Model reporting

For models in scope of this Framework, results of our systemic risk assessment and mitigation processes and measures are documented in a Safety and Security Model Report (referred to as “Transparency Reports” under the TFAIA). We also publish details of these assessments in system cards or in other reporting when a model launches.

Model report updates

For models in scope of this Framework that are subject to the EU AI Act, if we have reasonable grounds to believe that the basis for considering the model’s systemic risks acceptable has been materially undermined, we will update our Model Report as appropriate after completing a systemic risk assessment. For example, this could be the case if the model’s capabilities materially change through further post-training, if the model’s use or integrations into OpenAI’s systems materially increase risk, or in the event of a serious incident.

We will in any event determine whether to update the Model Report for our most capable frontier models once every six months. We will not consider an update necessary if (1) the model’s capabilities have not materially changed since the last update, (2) we plan to release a more capable model in less than a month, or (3) if the model is considered similarly safe or safer (under Appendix 2.2 of the Code of Practice).

Light touch evaluations

Separate from any full systemic risk assessment, we may from time to time conduct lighter-touch model evaluations at appropriate trigger points to determine whether (1) additional systemic risk mitigations may be required or (2) if a Model Report update is required. These trigger points may include (1) the release of an updated model or (2) where we have reason to believe a model’s risk profile may have materially changed.

5 Input from external experts

We may solicit and obtain input from external experts in relevant domains, and other stakeholders, to assist in systemic risk assessment or in determining the sufficiency of safety mitigations. This may include independent third-party evaluators, stress testing of safeguards for models approaching or reaching a new risk tier, or to provide independent expert opinions to assist the Safety Advisory Group in assessing the safety of a proposed deployment. We may also rely on commissioned research reports, public research, and other discussions with internal and external domain experts.

6 Allocation of responsibility for risk management

OpenAI OpCo LLC and OpenAI Ireland Limited maintain internal governance structures and practices designed to meet the requirements of applicable laws and ensure implementation of the processes in this Framework. OpenAI's internal governance practices include managing risks across the lifecycle of our models and ongoing legal and compliance reviews to ensure that risk management functions adhere to this Framework.

OpenAI OpCo LLC is responsible for compliance with the TFAIA for covered models in the United States.

OpenAI Ireland Limited is the provider of OpenAI's GPAI-SR models in the EU and responsible for compliance with the EU Code of Practice. The board of directors of OpenAI Ireland Limited exercise systemic risk oversight under this Framework for EU purposes.

Additional detail about our deployment governance can be found in Appendix B of the PF.

7 Framework change management

OpenAI commits to ensuring that this FGF is state-of-the-art and kept up to date with OpenAI's policies and procedures with respect to the TFAIA and the EU Code of Practice.

7.1 Update and approval process

Updates to this FGF may be proposed by Open AI's Safety Advisory Group, Head of Safety Systems, Head of Preparedness, Chief Information Security Officer, Chief Compliance Officer, General Counsel, or executive leadership. OpenAI's Legal function, in collaboration with relevant internal stakeholders, will oversee the process for FGF updates, including determining which updates are required to ensure the FGF remains state-of-the-art and adequate for its purpose. Updates will also occur as needed after a Framework assessment, as described in Section 7.2 below. Material updates will be presented to the Safety and Security Committee of the board of directors of the OpenAI Foundation and to the board of directors of OpenAI Ireland Limited for oversight, with changes and justifications for material updates documented in a changelog and published within 30 days of the update.

7.2 Framework assessment

OpenAI will complete a Framework Assessment at least once every 12 months from the Effective Dates of the TFAIA and the EU Code of Practice, and based on factors considered by OpenAI's legal function, such as changes in law or regulatory guidance, changes in frontier model capabilities and related technologies, new approaches to mitigations and safeguards, other incidents affecting the industry, and new industry best practices and standards.

Framework Assessments will consider the adequacy of our FGF and our factors for determining whether updates are required. For the EU Code of Practice, if we identify any instances of non-adherence or any measures that are required to be implemented to ensure continued adherence, we will draft and implement a remediation plan, and update the FGF as appropriate.