Preparedness Framework

Version 2. Last updated: 15th April, 2025

OpenAI

OpenAI's mission is to ensure that *AGI* (artificial general intelligence) benefits all of humanity. To pursue that mission, we are committed to safely developing and deploying highly capable AI systems, 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.

The Preparedness Framework is OpenAI's approach to tracking and preparing for frontier capabilities that create new risks of severe harm.[1] We currently focus this work on three areas of frontier capability, which we call *Tracked Categories*:

- **Biological and Chemical capabilities** that, in addition to unlocking discoveries and cures, can also reduce barriers to creating and using biological or chemical weapons.
- **Cybersecurity capabilities** that, in addition to helping protect vulnerable systems, can also create new risks of scaled cyberattacks and vulnerability exploitation.
- **AI Self-improvement capabilities** that, in addition to unlocking helpful capabilities faster, could also create new challenges for human control of AI systems.

In each area, we develop and maintain a threat model that identifies the risks of severe harm and sets thresholds we can measure to tell us when the models get capable enough to meaningfully pose these risks. We won't deploy these very capable models until we've built safeguards to sufficiently minimize the associated risks of severe harm. This Framework lays out the kinds of safeguards we expect to need, and how we'll confirm internally and show externally that the safeguards are sufficient.

In this updated version of the Framework we also introduce a set of *Research Categories*. These are areas of capability that could pose risks of severe harm, that do not yet meet our criteria to be Tracked Categories, and where we are investing now to further develop our threat models and capability elicitation techniques.

We are constantly refining our practices and advancing the science, to unlock the benefits of these technologies while addressing their risks. This revision of the Preparedness Framework focuses on the safeguards we expect will be needed for future models more capable than those we have today.

[1] By "severe harm" in this document, we mean the death or grave injury of thousands of people or hundreds of billions of dollars of economic damage. Our safety stack addresses a broad spectrum of risks, including many with harms below this severity. In choosing to set a high bar here, we aim to ensure that the most severe risks receive attention commensurate with their magnitude.

Contents

1 Introduction

We believe there are a limited number of AI capabilities that could pose new risks of severe harm. In order to safely unlock the beneficial uses of frontier AI capabilities, we exercise particular caution and implement safeguards that sufficiently minimize the risk of severe harm in these areas. To do this, we:

- **Decide where to focus** – we use a holistic risk assessment to decide which frontier capability categories to track or research further, and to define threshold levels of those capabilities that are associated with meaningful increases in risk of severe harm.
- **Measure capabilities associated with risks of severe harms** – we run in–scope models through frontier capability evaluations to measure the full extent of model capabilities before we deploy our models and during development. Our capability elicitation efforts are designed to detect the threshold levels of capability that we have identified as enabling meaningful increases in risk of severe harms.
- **Safeguard against severe harms** – we evaluate the likelihood that severe harms could actually occur in the context of deployment, using threat models that take our safeguards into account. We do not deploy models that reach a High capability threshold until the associated risks that they pose are sufficiently minimized. If a model under development reaches a Critical capability threshold, we also require safeguards to sufficiently minimize the associated risks during development, irrespective of deployment plans.
- **Build trust** – we engage with subject–matter experts across and beyond OpenAI to inform these efforts and to build confidence that we are meeting our commitments and effectively managing risk.

An internal, cross-functional group of OpenAI leaders called the *Safety Advisory Group* (SAG) oversees the Preparedness Framework and makes expert recommendations on the level and type of safeguards required for deploying frontier capabilities safely and securely. OpenAI Leadership can approve or reject these recommendations, and our Board's Safety and Security Committee provides oversight of these decisions.

1.1 Why we're updating the Preparedness Framework

Our environment is changing in four key ways:

- **Safeguarding stronger models will require more planning and coordination.** Until now, our models' own limitations have given us confidence that, in the areas tracked under the Preparedness Framework, they were safe to deploy. Our evaluations have so far shown that even our most advanced models, even without safeguards in place, aren't yet capable enough to pose severe risks in areas like bio- or cybersecurity. We are on the cusp of systems that can do new science, and that are increasingly agentic - systems that will soon have the capability to create meaningful risk of severe harm. This means we will need to design and deploy safeguards we can rely on for safety and security – which requires a new level of planning and coordination across the company, beyond what's needed to measure capabilities.
- **More frequent deployments require scalable evaluations.** Today, we're able to develop and deploy better models more often than ever thanks to reasoning advances that can unlock new capabilities without as much training as the previous paradigm. As a result, it's important to embrace methods that scale, including scalable capability evaluations that work well for a faster cadence of model deployment, as well as periodic deeper dives that validate those scalable evaluations.
- **A highly dynamic development landscape for frontier AI** makes it important for us to share our latest thinking. With a growing number of labs producing frontier AI models, it is more important than ever for us and other labs to contribute to community efforts on frontier safety and security, including by adapting safety work for the reasoning paradigm and for increasingly agentic systems, and advocating for and contributing to advanced protection measures in a world of continual frontier AI proliferation.
- **We and the broader field have gained more experience and built conviction on how to do this work.** The past year of research and deployment on safety frameworks both at OpenAI and across the AI field have given us greater clarity on how to prioritize and approach categories of risk. This

includes our threat modelling, development work on new capability evaluations, and external consultations, as well as relevant publications from external researchers and updated frameworks from industry peers.[2]

2 Deciding where to focus

2.1 Holistic risk assessment and categorization

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 and its partners, as well as relevant legal and policy mandates.

Where we determine that a capability presents a real risk of severe harm, we may decide to monitor it as a **Tracked Category** or a **Research Category**.

Tracked Categories are those capabilities which we track most closely, measuring them during each covered deployment and preparing safeguards for when a threshold level is crossed. We treat a frontier capability as a Tracked Category if the capability creates a risk that meets five criteria:[3]

1. **Plausible**: It must be possible to identify a causal pathway for a severe harm in the capability area, enabled by frontier AI.
2. **Measurable**: We can construct or adopt capability evaluations that measure capabilities that closely track the potential for the severe harm.
3. **Severe**: There is a plausible threat model within the capability area that would create severe harm.[1]
4. **Net new**: The outcome cannot currently be realized as described (including at that scale, by that threat actor, or for that cost) with existing tools and resources (e.g., available as of 2021) but without access to frontier AI.
5. **Instantaneous or irremediable**: The outcome is such that once realized, its severe harms are immediately felt, or are inevitable due to a lack of feasible measures to remediate.

We review and update Tracked Categories periodically or when we learn significant new information.

Research Categories are capabilities that, while they do not meet the above criteria, nonetheless have the potential to cause or contribute to severe harm, and where we are working now in order to prepare to address risks in the future (including potentially by maturing them to Tracked Categories).

2.2 Tracked Categories

For each Tracked Category, we develop and maintain a threat model identifying specific risks of severe harms that could arise from the frontier capabilities in that domain and sets corresponding capability thresholds that would lead to a meaningful increase in risk of severe harm.[4] SAG reviews and approves these threat models. Capability thresholds 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.

High capability thresholds mean capabilities that significantly increase existing risk vectors for severe harm. Covered systems that cross this capability threshold are required to have robust and effective safeguards that sufficiently minimize the associated risk of severe harm before they are deployed and appropriate security controls as they are developed. **Critical** capability thresholds mean capabilities that present a meaningful risk of a qualitatively new threat vector for severe harm with no ready precedent. Critical capabilities require safeguards even during the development of the covered system, irrespective of deployment plans.

[2]For example, our adoption of Capability Reports and Safeguards Reports parallels Anthropic's updated RSP.

[3]These criteria were informed in part by Meta's recent Frontier AI Framework.

[4]Threat models are informed both by our broader risk assessment process, and by more specific information that we gather across OpenAI teams and external experts. While past deployment experiences are an important source of information about existing risks, we also recognize that in the case of net-new risks of severe harm, significant safeguards may be needed to reduce the risk of harms that have never been realized.

Our team develops or adopts evaluations designed to track progress on each of these capabilities and to indicate when the identified capability thresholds may have been reached.

We further break down the specific capability thresholds, associated risks, and safeguards for those thresholds in Table 1.

Tracked Category	Capability threshold that could lead to the risk	Associated risk of severe harm	Risk-specific safeguard guidelines
Biological and Chemical[5]: The ability of an AI model to accelerate and expand access to biological and chemical research, development, and skill-building, including access to expert knowledge and assistance with laboratory work.	**[High]** 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.	Significantly increased likelihood and frequency of biological or chemical terror events by non-state actors using known reference-class threats.	• Require **security controls** meeting High standard (Appendix C.3) • Require **safeguards against misuse** meeting High standard (Appendix C.1) before external deployment
	[Critical] 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.	Proliferating the ability to create a novel threat vector of the severity of a CDC Class A biological agent (i.e., high mortality, ease of transmission) could cause millions of deaths and significantly disrupt public life, with few available societal safeguards.	• Until we have specified safeguards and security controls that would meet a Critical standard, halt further development • Contribute towards improved public policy and pandemic preparedness
Cybersecurity: The ability of an AI model to assist in the development of tools and executing operations for cyberdefense and cyberoffense.	**[High]** The model removes 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.	Removing bottlenecks limiting malicious cyber activity may upset the current cyberoffense-cyberdefense balance by significantly automating and scaling the volume of existing cyberattacks. In conjunction with a Long-range Autonomy capability (Section 2.3), models that could bypass OpenAI's technical safeguards constraining model activity, such as sandboxing or monitoring services, could compromise OpenAI's ability to track and mitigate all other risks.	• Require **security controls** meeting High standard (Appendix C.3) • Require **safeguards against misuse** meeting High standard (Appendix C.1) for external deployment • Require **safeguards** meeting High standard against **misalignment** (Appendix C.2) for large-scale internal deployment • Contribute towards improved cyberdefense policies and tools for cyberdefense

[5] We will build safeguards against both biological and chemical threats. Given the higher potential severity of biological threats relative to chemical ones, we will prioritize Biological capability evaluations and these will be used as indicators for High and Critical capabilities for the category.

	[Critical] 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 model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal.	Finding and executing end-to-end exploits for all software could lead to catastrophe from unilateral actors, hacking military or industrial systems, or OpenAI infrastructure. Novel cyber operations, e.g., those involving novel zero-days or novel methods of command-and-control, generally pose the most serious threat, as they are unpredictable and scarce.	• Until we have specified safeguards and security controls standards that would meet a Critical standard, halt further development
AI Self-improvement: The ability of an AI system to accelerate AI research, including to increase the system's own capability.	**[High]** The model's impact is equivalent to giving every OpenAI researcher a highly performant mid-career research engineer assistant, relative to those researchers' 2024 baseline.	This milestone suggests AI self-improvement may be beginning to accelerate. To meet the large-scale safety research, operations, and security investments that need to be in place when we reach Critical self-improvement, we will by then need to have made significant progress on those investments at this earlier milestone to avoid later risks.	• Require **security controls** meeting High standard (Appendix C.3)
	[Critical] The model is capable of recursively self improving (i.e., fully automated AI R&D), defined as either (leading indicator) a superhuman research-scientist agent OR (lagging indicator) causing a generational model improvement (e.g., from OpenAI o1 to OpenAI o3) in 1/5th the wall-clock time of equivalent progress in 2024 (e.g., sped up to just 4 weeks) sustainably for several months.	A major acceleration in the rate of AI R&D could rapidly increase the rate at which new capabilities and risks emerge, to the point where our current oversight practices are insufficient to identify and mitigate new risks, including risks to maintaining human control of the AI system itself.	• Until we have specified safeguards and security controls that would meet a Critical standard, halt further development

Table 1: Tracked Categories

2.3 Research Categories

There are also some areas of frontier capability that do not meet the criteria to be Tracked Categories, but where we believe work is required now in order to prepare to effectively address risks of severe harms in the future. These capabilities either need more research and threat modeling before they can be rigorously measured, or do not cause direct risks themselves but may need to be monitored because further advancement in this capability could undermine the safeguards we rely on to mitigate existing Tracked Category risks. We call these Research Categories, and in these areas we will take the following steps, both internally and in collaboration with external experts:

- Further developing the threat models for the area,

- Advancing the science of capability measurement in the area and investing towards the development of rigorous evaluations (which could be achieved internally or via partnerships), and
- Sharing summaries of our findings with the public where feasible.

We will periodically review the latest research and findings for each Research Category. SAG will receive the results of such reviews and evaluate whether there is sufficient evidence to recommend updates to our internal practices or to the Preparedness Framework.

Research Category	**Potential response**
Long-range Autonomy: ability for a model to execute a long-horizon sequence of actions sufficient to realize a "High" threat model (e.g., a cyberattack) without being directed by a human (including successful social engineering attacks when needed)	If a model has High or Critical capabilities in any of the Tracked Categories, require a "misalignment" safeguards report (see Section 4.2). As this category matures, we will make decisions about how this should influence our governance, including setting internal and external deployment safeguards milestones.
Sandbagging: ability and propensity to respond to safety or capability evaluations in a way that significantly diverges from performance under real conditions, undermining the validity of such evaluations.	Adopt elicitation approach that overcomes sandbagging, or use a conservative upper bound of the model's non-sandbagged evaluation results
Autonomous Replication and Adaptation: ability to survive, replicate, resist shutdown, acquire resources to maintain and scale its own operations, and commit illegal activities that collectively constitute causing severe harm (whether when explicitly instructed, or at its own initiative), without also utilizing capabilities tracked in other Tracked Categories.	Convert Autonomous Replication and Adaptation to a Tracked Category
Undermining Safeguards: ability and propensity for the model to act to undermine safeguards placed on it, including e.g., deception, colluding with oversight models, sabotaging safeguards over time such as by embedding vulnerabilities in safeguards code, etc.	If a model has High or Critical capabilities in any of the Tracked Categories, require the Safeguards case to be robust to the discovered capability and/or propensity
Nuclear and Radiological: ability to meaningfully counterfactually enable the creation of a radiological threat or enable or significantly accelerate the development of or access to a nuclear threat while remaining undetected.	Heighten safeguards (and consider further actions) in consultation with appropriate US government actors, accounting for the complexity of classified information handling.

Table 2: Research Categories

Changes to Tracked Categories in version 2

AI Self-improvement (now a Tracked Category), **Long-range Autonomy** and **Autonomous Replication and Adaptation** (now Research Categories) are distinct aspects of what we formerly termed Model Autonomy. We have separated self-improvement because it presents a distinct plausible, net new, and potentially irremediable risk, namely that of a hard-to-track rapid acceleration in AI capabilities which could have hard-to-predict severely harmful consequences. In addition, the evaluations we use to measure this capability are distinct from those applicable to Long-range Autonomy and Autonomous Replication and Adaptation. Meanwhile, while these latter risks' threat models are not yet sufficiently mature to receive the scrutiny of Tracked Categories, we believe they justify additional research investment and could qualify in the future, so we are investing in them now as Research Categories.

Nuclear and Radiological capabilities are now a Research Category. While basic information related to nuclear weapons design is available in public sources, the 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 a classified context. Our work on nuclear risks also informs our efforts on the related but distinct risks posed by radiological weapons. We build safeguards to prevent our models from assisting with high-risk queries related to building weapons, and evaluate performance on those refusal policies as part of our safety process. Our analysis suggests that nuclear risks are likely to be of substantially greater severity and therefore we will prioritize research on nuclear-related risks. We will also engage with US national security stakeholders on how best to assess these risks.

Persuasion: OpenAI prohibits the use of our products to manipulate political views as part of our Model Spec, and we build in safeguards to back this policy. We also continue to study the persuasive and relational capabilities of models (including on emotional well-being and preventing bias in our products) and monitor and investigate misuse of our products (including for influence operations). We believe many of the challenges around AI persuasion risks require solutions at a systemic or societal level, and we actively contribute to these efforts through our participation as a steering committee member of C2PA and working with lawmaker and industry peers to support state legislation on AI content provenance in Florida and California. Within our wider safety stack, our Preparedness Framework is specifically focused on frontier AI risks meeting a specific definition of severe harms[1], and Persuasion category risks do not fit the criteria for inclusion.

3 Measuring capabilities

3.1 Evaluation approach

We invest deeply in developing or adopting new science-backed evaluations that provide high precision and high recall indications of whether a covered system has reached a capability threshold in one of our Tracked Categories.

Our evaluations are intended to approximate the full capability that the adversary contemplated by our threat model could extract from the deployment candidate model, including by using the highest-capability tier of system settings, using a version of the model that has a negligible rate of safety-based refusals on our Tracked Category capability evaluations (which may require a separate model variant), and with the best presently-available scaffolds. These measures are taken to approximate the high end of expected elicitation by threat actors attempting to misuse the model, and should be tailored depending on the level of expected access (e.g., doing finetuning if the weights will be released).

Nonetheless, given the continuous progress in model scaffolding and elicitation techniques, we regard any one-time capability elicitation in a frontier model as a lower bound, rather than a ceiling, on capabilities that may emerge in real world use and misuse. We incorporate this uncertainty into our assessments. We monitor the technical landscape for changes to the elicitation techniques and best practices, and reassess our evaluations as needed.

Capability evaluations come in two different forms:

- **Scalable Evaluations:** automated evaluations designed to measure proxies that approximate whether a capability threshold has been crossed. Scalable evaluations have associated "indicative thresholds," which are levels of performance that we have pre-determined to indicate that a deployment may have reached a capability threshold.
- **Deep Dives:** designed to provide additional evidence validating the scalable evaluations' findings on whether a capability threshold has been crossed. These may include a wide range of evidence gathering activities, such as human expert red-teaming, expert consultations, resource-intensive third party evaluations (e.g., bio wet lab studies, assessments by independent third party evaluators), and any other activity requested by SAG.

An example of a Tracked Category capability evaluation

To assess the degree to which a covered system can reduce the barriers to creating a biological weapon, our current evaluations test both how capable the system is at providing useful information to someone creating a weapon and how capable it is of directly integrating with relevant tools, such as ordering precursor materials via the Internet.

Our evaluations test acquiring critical and sensitive information across the five stages of the biological threat creation process: Ideation, Acquisition, Magnification, Formulation, and Release. These evaluations, developed by domain experts, cover things like how to troubleshoot the laboratory processes involved.

You can learn more about our Tracked Category capability evaluations in past system cards, such as those for OpenAI o1 and Operator.

3.2 Testing scope

The Preparedness Framework applies to any new or updated deployment that has a plausible chance of reaching a capability threshold whose corresponding risks are not addressed by an existing Safeguards Report. Examples of such covered deployments are:

- every frontier model (e.g., OpenAI o1 or OpenAI o3) that we plan to deploy externally
- any agentic system (including significant agents deployed only internally) that represents a substantial increase in the capability frontier
- any significant change in the deployment conditions of an existing model (e.g., enabling finetuning, releasing weights, or significant new features) that makes the existing Capabilities Report or Safeguards Report no longer reasonably applicable
- incremental updates or distilled models with unexpectedly significant increases in capability. [6]

If justified by our forecasts and threat models as potentially posing a severe risk during development and prior to external deployment, we will select an appropriate checkpoint during development to be covered by the Preparedness Framework.

In cases of ambiguity about whether a model is covered by the Preparedness Framework, the SAG is responsible for making the final determination.

3.3 Capability threshold determinations

Prior to deployment, every covered model undergoes the suite of Scalable Evaluations. The results of these evaluations, and any noteworthy observations that should affect interpretation of the results, are compiled into a Capabilities Report that is submitted to the SAG.

The determination that a threshold has been reached is informed by these indicative results from capability evaluations, and also reflects holistic judgment based on the totality of available evidence – for example, information about the methodological robustness of evaluation results.

The SAG reviews the Capabilities Report and decides on next steps. These can include:

- **Determine that the capability threshold has been crossed, and therefore recommend implementing and assessing corresponding safeguards**, if they have not already been implemented and assessed.
- **Determine that a threshold has not been crossed**: If the scalable evaluations did not cross their indicative thresholds, the SAG may conclude that the model does not have High or Critical capability, and recommend no further action.

[6] In general, models that we distill, fine-tune, or quantize from a model that was previously determined not to cross a High capability threshold will ordinarily not require additional safety measures barring reason to believe a significant increase in capability has occurred.

- **Recommend deep dive research**: This is appropriate if SAG needs additional evidence in order to make a recommendation.

4 Safeguarding against severe harm

Safeguards that mitigate the risk of severe harm are at the core of our safe deployment approach for covered systems that reach a capability threshold. This section describes the process through which we select which safeguards to apply and how we evaluate whether those safeguards sufficiently minimize the risk of severe harm.

If a covered system appears likely to cross a capability threshold, we will start to work on safeguards to sufficiently reduce the risks of severe harm associated with that threshold even if a formal capability determination has not yet been made.

4.1 Safeguard selection

Each capability threshold has a corresponding class of risk-specific safeguard guidelines under the Preparedness Framework. We use the following process to select safeguards for a deployment:

- We first identify the plausible ways in which the associated risk of severe harm can come to fruition in the proposed deployment.
- For each of those, we then identify specific safeguards that either exist or should be implemented that would address the risk.
- For each identified safeguard, we identify methods to measure their efficacy and an efficacy threshold.

We consider separate safeguards for two of the main ways in which risks can be realized: a **malicious user**, who can leverage the model to cause the severe harm, and a **misaligned model**, which autonomously causes the harm. We illustrate the types of safeguards for each in Table 3.

The methodology for security controls is similar, although focused on the relevant security threat models.

Appendix C provides illustrative examples of potential safeguards and safeguard efficacy assessments that could be used to establish that we have sufficiently minimized the risk of severe harm. The examples aim to provide insight on our thinking but should not be construed as a definitive checklist of the safeguards we will apply to a given launch.

4.2 Safeguard sufficiency

We compile the information on the planned safeguards needed to minimize the risk of severe harm into a Safeguards Report. The Safeguards Report should include the following information:

- Identified ways a risk of severe harm can be realized for the given deployment, each mapped to the associated security controls and safeguards
- Details about the efficacy of those safeguards
- An assessment on the residual risk of severe harm based on the deployment
- Any notable limitations with the information provided

Many safeguards can be reused across different deployments, and so in practice we may find that the majority of deployments need few new safeguards and associated stress-testing beyond those already in place for prior deployments.

SAG is responsible for assessing whether the safeguards associated with a given deployment sufficiently minimize the risk of severe harm associated with the proposed deployment. The SAG will make this determination based on:

- The **level of capability** in the Tracked Category based on the Capabilities Report.
- The **associated risks of severe harm**, as described in the threat model and where needed, advice of internal or external experts.

Safeguards Against Malicious Users	Safeguards Against a Misaligned Model
Potential claims: • **Robustness:** Malicious users cannot use the model to cause the severe harm because they cannot elicit the necessary capability, such as because the model is modified to refuse to provide assistance to harmful tasks and is robust to jailbreaks that would circumvent those refusals. • **Usage Monitoring:** If a model does not refuse and provides assistance to harmful tasks, monitors can stop or catch malicious users before they have achieved an unacceptable scale of harm, through a combination of automated and human detection and enforcement within an acceptable time frame. • **Trust-based Access:** The actors who gain access to the model are not going to use it in a way that presents an associated risk of severe harm under our threat model.	Potential claims: • **Lack of Autonomous Capability**: The model is not capable of carrying out tasks autonomously, including the risk of severe harm. • **Value Alignment**: The model consistently applies human values in novel settings (without any instructions) to avoid taking actions that cause harm, and has shown sufficiently minimal indications of misaligned behaviors like deception or scheming. • **Instruction Alignment**: The model consistently understands and follows user or system instructions, even when vague, and those instructions rule out pathways to causing severe harm. • **Reliable and Robust System Oversight**: Effective AI and human oversight of model actions detects and prevents execution of harm or subversion of safeguards. • **System Architecture**: The model can't take actions that cause harm because it lacks access to output channels or mechanisms to persist sufficiently to execute the harm, due to system design and restricted permissions.

Table 3: Types of safeguards. See Appendix C.1 and C.2 for additional details.

- The **safeguards in place and their effectiveness** based on the Safeguards Report.
- The **baseline risk from other deployments**, based on a review of any non-OpenAI deployments of models which have crossed the capability thresholds and any public evidence of the safeguards applied for those models.

Covered systems that reach High capability must have safeguards that sufficiently minimize the associated risk of severe harm before they are deployed. Systems that reach Critical capability also require safeguards that sufficiently minimize associated risks during development.

Based on this evidence, SAG then has the following decision points:

1. SAG can find that it is confident that the safeguards sufficiently minimize the associated risk of severe harm for the proposed deployment, and **recommend deployment**.
2. SAG can **request further evaluation of the effectiveness of the safeguards** to evaluate if the associated risk of severe harm is sufficiently minimized
3. SAG can **find the safeguards do not sufficiently minimize the risk of severe harm** and recommend potential alternative deployment conditions or additional or more effective safeguards that would sufficiently minimize the risk.

The SAG will strive to recommend further actions that are as targeted and non-disruptive as possible while still mitigating risks of severe harm. All of SAG's recommendations will go to OpenAI Leadership for final decision-making in accordance with the decision-making practices outlined in Appendix B.

We expect to continuously improve our safeguards over time. If we find reasonable evidence that our safeguards are not working as expected, we will validate the information being received and review the

sufficiency of our safeguards.

4.3 Marginal risk

We recognize that another frontier AI model developer might develop or release a system with High or Critical capability in one of this Framework's Tracked Categories and may do so without instituting comparable safeguards to the ones we have committed to. Such an action could significantly increase the baseline risk of severe harm being realized in the world, and limit the degree to which we can reduce risk using our safeguards. If we are able to rigorously confirm that such a scenario has occurred, then we could adjust accordingly the level of safeguards that we require in that capability area, but only if:

- we assess that doing so does not meaningfully increase the overall risk of severe harm,
- we publicly acknowledge that we are making the adjustment,
- and, in order to avoid a race to the bottom on safety, we keep our safeguards at a level more protective than the other AI developer, and share information to validate this claim.

4.4 Increasing safeguards before internal use and further development

Models that have reached or are forecasted to reach Critical capability in a Tracked Category present severe dangers and should be treated with extreme caution.

Such models require additional safeguards (safety and security controls) during development, regardless of whether or when they are externally deployed. We do not currently possess any models that have Critical levels of capability, and we expect to further update this Preparedness Framework before reaching such a level with any model.

Our approach to Critical capabilities will need to be robust to both malicious actors (either internal or external) and model misalignment risks. The SAG retains discretion over when to request deep dive evaluations of models whose scalable evaluations indicate that they may possess or may be nearing critical capability thresholds.

5 Building trust

Effective implementation of the Preparedness Framework requires internal and external accountability, so that the public, governments, and our industry peers can trust in our adherence to this policy.

5.1 Internal governance

- **Clear internal decision-making practices**. We have clear roles and responsibilities and decision-making practices as described in Appendix B.
- **Internal Transparency**. We will document relevant reports made to the SAG and of SAG's decision and reasoning. Employees may also request and receive a summary of the testing results and SAG recommendation on capability levels and safeguards (subject to certain limits for highly sensitive information).
- **Noncompliance**. Any employee can raise concerns about potential violations of this policy, or about its implementation, via our Raising Concerns Policy. We will track and appropriately investigate any reported or otherwise identified potential instances of noncompliance with this policy, and where reports are substantiated, will take appropriate and proportional corrective action.

5.2 Transparency and external participation

- **Public disclosures**: We will release information about our Preparedness Framework results in order to facilitate public awareness of the state of frontier AI capabilities for major deployments. This published information will include the scope of testing performed, capability evaluations for each Tracked Category, our reasoning for the deployment decision, and any other context about a model's development or capabilities that was decisive in the decision to deploy. Additionally, if the model is beyond a High threshold, we will include information about safeguards we have implemented to

sufficiently minimize the associated risks. Such disclosures about results and safeguards may be redacted or summarized where necessary, such as to protect intellectual property or safety.

- **Third-party evaluation of tracked model capabilities**: If we deem that a deployment warrants deeper testing of Tracked Categories of capability (as described in Section 3.1), for example based on results of Capabilities Report presented to them, then when available and feasible, OpenAI will work with third-parties to independently evaluate models.
- **Third-party stress testing of safeguards**: If we deem that a deployment warrants third party stress testing of safeguards and if high quality third-party testing is available, we will work with third parties to evaluate safeguards. We may seek this out in particular for models that are over a High capability threshold.
- **Independent expert opinions for evidence produced to SAG**: The SAG may opt to get independent expert opinion on the evidence being produced to SAG. The purpose of this input is to add independent analysis from individuals or organizations with deep expertise in domains of relevant risks (e.g., biological risk). If provided, these opinions will form part of the analysis presented to SAG in making its decision on the safety of a deployment. These domain experts may not necessarily be AI experts and their input will form one part of the holistic evidence that SAG reviews.

A Change log

In this version of the Preparedness Framework, we make a number of updates, designed to reflect what we've learned and update our safety and governance process for the next generations of highly capable models. Key changes include that we:

1. **Clarify the relationship among capabilities, risks and safeguards**. In our updated framework, we make clear that we use a holistic process to decide which areas of frontier AI capability to track, and to define threshold levels of those capabilities that are associated with meaningful increases in risk of severe harm. We describe how we develop and maintain threat models that identify these severe risks, and how we evaluate model capabilities and build and test safeguards that sufficiently minimize the associated risks. We make clear that safeguards can take a variety of forms, and that reducing risk generally does not require reducing capability.
2. **Define how High and Critical capability thresholds relate to underlying risks**. High capability thresholds mean capabilities that significantly increase existing risk vectors for severe harm under the relevant threat model. Critical capability thresholds mean capabilities that present a meaningful risk of a qualitatively new threat vector for severe harm with no ready precedent under the relevant threat model. Also, we are removing terms "low" and "medium" from the Framework, because those levels were not operationally involved in the execution of our Preparedness work.
3. **Give specific criteria for which capabilities we track**. We track capabilities that create risks meeting five criteria – they are plausible, measurable, severe, net new, and instantaneous or irremediable.
4. **Update the Tracked Categories** of frontier capability accordingly, focusing on biological and chemical capability, cybersecurity, and AI self-improvement. Going forward we will handle risks related to persuasion outside the Preparedness Framework, including via our Model Spec and policy prohibitions on the use of our tools for political campaigning or lobbying, and our ongoing investigations of misuse of our products (including detecting and disrupting influence operations). We are moving Nuclear and Radiological capabilities into Research Categories.
5. **Introduce Research Categories**, areas of capability that do not meet the criteria to be Tracked Categories, but where we believe additional work is needed now. For these areas, in collaboration with external experts, we commit to further developing the associated threat models and advancing the science of capability measurement for the area, including by investing in the development of rigorous capability evaluations. These include Long-range Autonomy, Sandbagging, Autonomous Replication and Adaptation, Undermining Safeguards, and Nuclear and Radiological.
6. **Provide more detail on our capability elicitation approach**, making clear that we will consider a range of techniques in order to test a model version that approximates the high end of expected elicitation by threat actors attempting to misuse the model. We also define "scalable evaluations," which are automated, and distinguish these from "deep dive" evaluations that may include consultation with human experts and are designed in part to validate the scalable evaluations.
7. **Provide risk-specific safeguard guidelines**. This information gives more detail on how we expect to safely develop and deploy models advanced enough to pose severe risks in tracked capability areas. As we move toward increasingly capable models, we are planning for safeguards that will be tailored to the specific risks they are intended to address.
8. **Establish Capabilities Reports and Safeguards Reports**, the key artifacts we use to support informed decision-making under the Preparedness Framework in the context of systems that are capable enough to pose severe risks.
9. **Clarify approach to establishing safeguard efficacy**, moving beyond the flawed approach of re-running capability evaluations on the safeguarded model and towards a more thorough assessment of each safeguard and its efficacy (Section 4.1).
10. **Deprioritize safety drills**, as we are shifting our attention to a more durable approach of continuously red-teaming and assessing the effectiveness of our safeguards.
11. **Clarify our focus on marginal risk**, including the context of other systems available on the market, and outline our approach for maintaining responsible safeguards and reinforcing responsible practices across the industry if another actor releases a system we would assess as having High or Critical capability.

12. **Clarify the governance process**. Our Safety Advisory Group oversees the effective design, implementation, and adherence to the Preparedness Framework, in partnership with safety leaders in the company. For covered launches, SAG assesses residual risk in tracked areas, net of safeguards, and makes expert recommendations on safeguard adequacy and deployment decision-making to OpenAI leadership.

B Decision-making practices

We establish an operational structure to oversee our procedural commitments within the Preparedness Framework.

The Safety Advisory Group (SAG) is responsible for:

- Overseeing the effective design, implementation, and adherence to the Preparedness Framework in partnership with the safety organization leader
- For each deployment in scope under the Preparedness Framework, reviewing relevant reports and all other relevant materials and assessing of the level of Tracked Category capabilities and any post-safeguards residual risks
- For each deployment under the Preparedness Framework, providing recommendations on potential next steps and any applicable risks to OpenAI Leadership, as well as rationale
- Making other recommendations to OpenAI Leadership on longer-term changes or investments that are forecasted to be necessary for upcoming models to continue to keep residual risks at acceptable levels
- For the avoidance of doubt, OpenAI Leadership can also make decisions without the SAG's participation, i.e., the SAG does not have the ability to "filibuster"

SAG Membership: the SAG provides a diversity of perspectives to evaluate the strength of evidence related to catastrophic risk and recommend appropriate actions.

- The members of the SAG and the SAG Chair are appointed by the OpenAI Leadership.
- SAG members serve for one year terms. OpenAI Leadership may choose to re-appoint someone from previous years to ensure there is continuity of knowledge and experience, while still ensuring that fresh and timely perspectives are present in the group.
- The SAG Chair makes any final decisions needed for the SAG. This role is expected to rotate, as appointed by OpenAI Leadership.

OpenAI Leadership, i.e., the CEO or a person designated by them, is responsible for:

- Making all final decisions, including accepting any residual risks and making deployment go/no-go decisions, informed by SAG's recommendations.
- Resourcing the implementation of the Preparedness Framework (e.g., additional work on safeguards where necessary).

The Safety and Security Committee (SSC) of the OpenAI Board of Directors (Board) will be given visibility into processes, and can review decisions and otherwise require reports and information from OpenAI Leadership as necessary to fulfill the Board's oversight role. Where necessary, the Board may reverse a decision and/or mandate a revised course of action.

Updates to the Preparedness Framework. The Preparedness Framework is a living document and will be updated. The SAG reviews proposed changes to the Preparedness Framework and makes a recommendation that is processed according to the standard decision-making process. We will review and potentially update the Preparedness Framework for continued sufficiency at least once a year.

Fast-track. In the rare case that a risk of severe harm rapidly develops (e.g., there is a change in our understanding of model safety that requires urgent response), we can request a fast track for the SAG to process the report urgently. The SAG Chair should also coordinate with OpenAI Leadership for immediate reaction as needed to address the risk.

C Illustrative safeguards, controls, and efficacy assessments

This Appendix provides illustrative examples of potential safeguards, and safeguard efficacy assessments that could be used to establish that we have sufficiently mitigated the risk of severe harm. The examples aim to provide insight on our thinking, but many of the techniques require further research. The safeguards should not be construed as a definitive or comprehensive list of the safeguards we will or could apply to a given launch.

As a reminder, covered systems that reach High capability must have safeguards that sufficiently minimize the associated risk of severe harm before they are deployed. Systems that reach Critical capability also require sufficient safeguards during development.

C.1 Safeguards against malicious users

Several of the Tracked Categories pose risks via malicious users leveraging the frontier capability to enable severe harm, such as professional hackers automating and scaling cyberattacks or terrorists consulting a model to debug the development of a biological weapon.

Safeguards should sufficiently minimize the risk of severe harm associated with misuse of the model's capabilities. This can be done by establishing that all plausible known vectors of enabling severe harm are sufficiently addressed by one or more of the following claims:

- **Robustness**: Users cannot use the model to cause the harm because they cannot elicit the capability, such as because the model is modified to refuse to provide assistance to harmful tasks and is robust to jailbreaks that would circumvent those refusals.
- **Usage Monitoring**: If a model does not refuse and provides assistance to harmful tasks, monitors can stop or catch adversarial users before their misuse has achieved an unacceptable scale of harm, through a combination of automated and human detection and enforcement within an acceptable time frame.
- **Trust-based Access**: The actors who gain access to the model will not use it in a way that presents an associated risk of severe harm under our threat model.

The table below illustrates possible examples of safeguards and safeguard efficacy assessments we could consider to claim we have sufficiently minimized the risks of severe harm associated with a High level of capability under any of our Tracked Categories.

Claim	Safeguards that could support this claim	Potential safeguard efficacy assessments
Robustness	• Training the model to refuse to help with high-risk tasks or to otherwise produce low-risk responses • Unlearning or training-data filtering to erase specific risk-enabling knowledge from the model's knowledge-base • Interpretability-based approaches, like activation steering, that directly edit models' thinking at inference time • Jailbreak robustness, including through adversarial training, inference-time deliberation, and more	• Automated and expert red-teaming (identifying success per resources) • Prevalence of jailbreaks identified via monitoring and reports, in historical deployments • Results from public jailbreak bounties and results from private and public jailbreak benchmarks
Usage Monitoring	• Monitoring harmful user requests, model responses, and aggregate usage patterns using auxiliary models trained to detect harmful tasks and jailbreak attempts • Blocking unsafe user requests and model responses automatically when possible and escalating to human review and approval otherwise • Expanding human monitoring and investigation capacity to track capabilities that pose a risk of severe harm, and developing data infrastructure and review tools to enable human investigations • Blocking access for users and organizations that violate our usage policies, leading to potential permanent bans • Expanding escalation channels with law enforcement channels	• Automated and human red-teaming of each oversight and review mechanism (focused on detection precision/recall) • In-production detection and enforcement SLAs for human monitoring teams, and historical performance • Rapid remediation of known jailbreaks: Time to patching a new known jailbreak • Redteamer success creating a certain number of accounts without detection (success per resources)
Trust-based Access	• Iterated deployment to wider user groups • Measures for establishing user trustworthiness – Trusted customer relationships – KYC and identity verification to enable traceability – KYC robustness to prevent a banned actor from creating new accounts at scale (recidivism) • Actor-policy standards, including bespoke usage policies that limit access to higher capabilities to trusted actors • Access conditions including variation across actors, e.g., rate-limits, fine-tuning access, releasing weights	• Iterated deployment to wider user group, with careful monitoring to detect instances of abuse

Table 4: Claims, potential safeguards, and efficacy assessments for malicious users

C.2 Safeguards against a misaligned model

In addition to human misuse, several of the Tracked Category capabilities pose risks when models themselves are able to autonomously execute a severe harm, whether due to misalignment or subversion by an adversary.

Safeguards should sufficiently minimize the risk associated with a misaligned model circumventing human control and oversight and executing severe harms. This can be done by establishing that all plausible known vectors of enabling severe harm are sufficiently addressed by one or more of the following claims:

- **Lack of Autonomous Capability**: The model is not capable of carrying out tasks autonomously, including the risk of severe harm.
- **Value Alignment**: The model consistently applies human values in novel settings (without any instructions) to avoid taking actions that cause harm, and has shown sufficiently minimal indications of misaligned behaviors like deception or scheming.
- **Instruction Alignment**: The model consistently understands and follows user or system instructions, even when vague, and those instructions rule out pathways to causing severe harm.
- **Reliable and Robust System Oversight**: Effective AI and human oversight of model actions detects and prevents execution of harm or subversion of safeguards.
- **System Architecture**: The model can't take actions that cause harm because it lacks access to output channels or mechanisms to persist sufficiently to execute the harm, due to containerization and restricted permissions.

These misalignment claims may be further complemented by the safeguards applied towards the malicious-actor-oriented "Robustness" and "Usage Monitoring" claims, which similarly limit a misaligned model's use of its own or another model's capabilities.

The table below illustrates possible examples of safeguards and safeguard efficacy assessments we could consider to claim we have sufficiently minimized the risks of severe harm associated with High level of capability under any of our Tracked Categories.

Claim	Potential safeguards	Potential safeguard efficacy assessments
Lack of Autonomous Capability	• Not applicable, as this involves capability evaluation.	• Long-range Autonomy capability evaluations show the model cannot act autonomously as described in the threat model (Section 2.3)
Value Alignment	• Generalization from specified or trained values • Steering or interpretability to control misaligned behaviors like deception or scheming	• Red-teamers create both environments and adversarial requests to probe the limits of the model's value alignment. Evidence of taking harmful actions, including deception and scheming, would be evidence of misalignment • Internal and external user reports, monitoring, or manual inspection for historical model deployments show no significant evidence of unexpected value alignment failures
Instruction Alignment	• Instruction following, instruction hierarchy and adherence to complex Model Specs • Model spec defined to comprehensively rule out risk pathways • Training user models and training AI to collaborate with users • Build risk models and training to avoid unnecessary risk • Training for uncertainty estimation and reliable answers	• Large scale usage or testing in realistic settings should show the model consistently takes action aligned to the user's instructions or goals (while maintaining value alignment and avoiding harm). Examples against would include trying to escape a container, lying to the user to appease them, etc.
Reliable and Robust System Oversight	• Monitor AI supervises agent actions to detect agent in the case of trying to take harmful actions, scheming, or similar • System forces human approval of agent actions (either by default, or by escalation of the monitor) • Model's reasoning is provided faithfully and interpretably to humans to review • Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm	• Red-teamers can train an arbitrarily adversarial agent or can pretend to be an agent in our setup and we should ensure that harmful behaviors are detected by monitors with a high recall rate • Monitor escalations are either outright blocked or, in empirical studies, users with the relevant level of expertise were shown to reliably understand and act on monitor flags before significant harm occurred • SLAs for asynchronous monitors to analyze and detect harmful agent actions post-hoc
System Architecture (including containerization and permissions)	• Limiting internet access and other tool access • Limiting credentials • Limiting access to system resources or filesystem (e.g., sandboxing) • Limiting persistence or state	• Testing and red-teaming of the system protections • "Canary evaluations" which test model capabilities to bypass less complex, easier-to-exploit versions of our security controls, establishing that our implemented controls are robust

Table 5: Claims, potential safeguards, and efficacy assessments for misaligned models

C.3 Security controls

Robust security practices and controls are essential to mitigating risks associated with the deployment of models reaching High capability thresholds. Our security approach integrates industry best practices with OpenAI-specific innovations tailored to manage the unique threats posed by frontier AI technologies. Our security practices are designed to protect against external and internal adversaries and align with established frameworks such as ISO 27001, SOC2, NIST SP 800-53, and FedRAMP. While nascent, AI-specific security standards may also provide frameworks for addressing security and regulatory standards that address the unique challenges of securing AI systems. This may include efforts emerging from the Cloud Security Alliance's AI Safety Initiative or the NIST SP 800-218 AI updates.

We will require the following practices for High capability models:

- **Security Threat Modeling and Risk Management**
  - Comprehensive Security Threat Models: Ensure OpenAI employs security threat modeling that systematically identifies and mitigates risks posed by adversaries. These models address both external and internal threats, explicitly mapping vulnerabilities and attack vectors relevant to frontier AI model access or misuse.
  - Continuous Review and Iteration: Ensure security threat models are regularly reviewed and updated. This includes updating the Security framework, policies, and controls as technologies and threats evolve.
  - Continuous Monitoring and Validation: Ensure security threat models and updates inform where security and data privacy controls should be implemented, improved, and monitored to further reduce risk. Internal and external assessments to validate these controls are conducted regularly and reports are provided to OpenAI leadership.
- **Defense in Depth**
  - Layered Security Architecture: Adopt a layered security strategy, ensuring robust protection through multiple defensive barriers, including physical and datacenter security, network segmentation and controls, workload isolation, data encryption, and other overlapping and complementary security controls.
  - Zero Trust Principles: Embrace Zero Trust principles in the design and operation of infrastructure, networks, and endpoints. Leverage hardware-backed security and modern authentication standards and controls for accessing critical resources and performing operations.
- **Access Management**
  - Principle of Least Privilege: Ensure access to systems and data is limited based on job functions, need-to-know, and operational requirements in alignment with the principles of least privilege and separation of duties. Ensure all access is regularly audited and reviewed. For all models, especially those presenting High capabilities, access must be strictly limited, protected with strong multi-factor authentication, and may require additional approvals for access.
  - Identity and Device Management: Employees must authenticate using multi-factor authentication (MFA) and managed devices meeting security baselines. Access must be logged and reviewed for detection and investigative purposes.
- **Secure Development and Supply Chain**
  - Secure Development Lifecycle: Integrate automated code analysis, formal security reviews, and penetration testing in engineering processes. Apply security reviews and validation to higher-sensitivity critical components prior to deployment.
  - Change Management: Establish and maintain a formal change management process to ensure that all modifications to systems, applications, and infrastructure are properly authorized, documented, tested, and approved before deployment. Changes to critical infrastructure should undergo multi-person approval and review. This process should include maintaining comprehensive records of changes and implementing rollback procedures to revert to previous states.

  - Software Supply Chain Integrity: Enforce supply chain security measures, including sourcing hardware and software from reputable sources, and continuous vetting and monitoring of third-party suppliers and software libraries.

- **Operational Security**
  - Monitoring and Incident Response: Monitor security and event logs continuously to detect, triage, and respond to security incidents rapidly by 24x7 on-call staff.
  - Vulnerability Management: Ensure known vulnerabilities are addressed through consistent patching and corrective actions, minimizing exploitation opportunities.
  - Adversarial Testing and Red-Teaming: Conduct adversarial testing and red-teaming exercises to proactively identify and mitigate potential vulnerabilities within corporate, research, and product systems, ensuring resilience against unknown vulnerabilities and emerging threats. Encourage reporting of good-faith security research through bug bounty programs.

- **Auditing and Transparency**
  - Independent Security Audits: Ensure security controls and practices are validated regularly by third-party auditors to ensure compliance with relevant standards and robustness against identified threats.
  - Transparency in Security Practices: Ensure security findings, remediation efforts, and key metrics from internal and independent audits are periodically shared with internal stakeholders and summarized publicly to demonstrate ongoing commitment and accountability.
  - Governance and Oversight: Ensure that management provides oversight over the information security and risk management programs.