OpenAI

An Update on our Voluntary Commitments and Safety Frameworks

Ahead of the Paris AI Action Summit, we have prepared an update showing our progress on the voluntary commitments made at previous summits, specifically those set forth at the AI Summits in Bletchley and Seoul. OpenAI remains committed to fulfilling the voluntary commitments made at previous summits and we look forward to sharing our progress on safety and engaging with world leaders and civil society groups in Paris.

We continue to advance our safety initiatives through rigorous internal processes, external collaborations, transparent reporting, and ensuring that our technology is used to benefit everyone responsibly.

OpenAI

Managing Frontier Risk

In December 2023, we published our Preparedness Framework, a living document that guides the safe deployment of our most advanced AI models. Grounded in science-driven assessments, iterative deployment, and continuous refinement, the Framework has shaped our approach to evaluating and mitigating frontier risks. Over the past year, we have gathered insights from real-world testing, expert feedback, and emerging research, and we are actively working on a revised version that we plan to publish later this year. This update will reflect refinements to our risk thresholds, mitigation strategies, and more.

The current Framework identifies and evaluates risks across several categories—including cybersecurity, chemical, biological, radiological, and nuclear (CBRN) threats, persuasion, autonomy, and potential "unknown unknowns"—by combining threat modeling, specialized capability elicitation, external expert reviews, and red-teaming to uncover worst-case scenarios. Each risk is evaluated through a dynamic Scorecard that measures outcomes both before and after safety measures are applied. This enables us not only to understand the severity of emerging threats in a worst-case context but also to verify that our interventions reduce risk to acceptable levels prior to any model release.

To address these risks, we use a variety of mitigation strategies. Containment strategies focus on limiting possession-related risks, such as compartmentalization and restricting access to trusted users. Deployment mitigations include measures like refusals, data redaction, usage policies, usage monitoring, enforcement, and alerting partners. Currently, only models with a post-mitigation score of "medium" or below can be deployed. Similarly, only models with a post-mitigation score of "high" or below can continue to be developed.

We have also established a governance structure to uphold procedural commitments and ensure effective risk management. This includes our preparedness team, which is focused on identifying, forecasting and quantifying frontier capabilities and potential catastrophic risks; our safety and alignment research teams, which ensure the safety, robustness and reliability of AI models and their deployment in the real world, and on researching scalable, trustworthy AI systems that consistently follow human intent. Our platform safety teams define our usage

policies, monitor usage and model behavior in production, take enforcement action when needed, and feed the learnings back to the upstream safety teams.

To provide structured oversight and decision making, we established the Safety Advisory Group, a cross-functional body that evaluates risk levels, thresholds, and mitigation strategies, making recommendations to leadership. In May of 2024, we also announced the formation of the OpenAI Board's Safety and Security Committee (SSC). The SSC performs oversight of major model releases and has the authority to delay a release until safety concerns are resolved.

As we continue to evolve our approach, we remain committed to refining our risk assessment and mitigation strategies. Our forthcoming update to the *Preparedness Framework* will incorporate what we have learned over the past year, ensuring that our models are developed and deployed with the highest standards of safety and responsibility.

Implementing Safety Best Practices

Beyond our work on frontier risks, we actively track, assess, and mitigate a wide range of non-frontier risks. To ensure our products remain safe and reliable, we implement a comprehensive set of safeguards, including external testing, rigorous evaluations, security and safety controls, usage policies, post-deployment monitoring, and provenance techniques.

External Testing

OpenAI has an extensive Red Teaming Network which consists of a diverse community of trusted external experts—including individual subject matter experts, research institutions, and civil society organizations—who help identify risks across cybersecurity, biological and chemical threats, and societal harms. We've collaborated with external partners like METR and Apollo Research to strengthen our frontier risk assessments, launched public bug bounties for probing vulnerabilities in cybersecurity, assessed our models on public jailbreak arenas to decrease model vulnerability to jailbreaks, and worked with various external red teamers to assess emerging risks from new modalities like voice and video generation. We also published new research on our approaches to red teaming to show how our external and automated red teaming efforts are advancing to help deliver safe and beneficial AI. In December 2024, we also invited safety researchers to apply for early access to our next frontier models.

Security and Safety Controls

We started investing in security years before ChatGPT or the API launched, ensuring that we stay ahead of evolving risks and continuously enhance our resilience as an organization. For example, we have made significant investments in cybersecurity and insider threat safeguards to protect our research infrastructure, including proprietary and unreleased model weights. We have also described high-level details about the security architecture of our research supercomputers, which enable us to deliver industry-leading models in both capabilities and safety while advancing the frontiers of AI.

As a leader in AI security, we are defining the security measures and implementing strict access controls, conducting internal and external penetration testing, and running a bug bounty program to safeguard training environments and high value algorithmic secrets. We believe that protecting advanced AI systems will benefit from an evolution of infrastructure security and are exploring novel controls to protect our technology. To empower cyber defense, we have funded third-party security researchers with our Cybersecurity Grant Program.

We are also dedicated to identifying, preventing and disrupting attempts to abuse our models for harmful ends and share threat intelligence with government, civil society, and industry stakeholders. We published our findings to amplify efforts to address these challenges. See state-affiliated threat actors (Feb 14, 2024); deceptive misuse by covert influence (May 2024); covert Iranian influence operation (Aug 16, 2024), deceptive AI misuse (October 2024). Last fall we also introduced 'omni-moderation-latest', a new GPT-4o-based model in our Moderation API that supports both text and image inputs, enhancing accuracy in detecting harmful content across multiple categories and languages.

Transparency and Information Sharing

System Cards: OpenAI regularly publishes system cards for new frontier AI systems, providing insights into safety evaluations, performance limitations, societal risks (fairness, bias), and adversarial testing results. These system cards aim to inform readers about key factors impacting the system's behavior, especially in areas pertinent for responsible usage, including limitations in performance that have implications for the domains of appropriate use, discussions of the model's effects on societal risks such as fairness and bias, and the results of adversarial testing conducted. Since signing the voluntary commitments, we have released system cards for DALL·E 3, GPT-4's vision capabilities, GPT-4o (August 2024), o1 (December 2024), Sora (December 2024), Operator (January 2025). We will share our safety insights and safeguards for deep research in a system card before we expand access.

Agreements with US and UK AI Safety Institutes: On August 29, 2024, OpenAI announced an agreement with the US AI Safety Institute at the U.S. Department of Commerce's National Institute of Standards and Technology that enables formal collaboration on AI safety research, testing and evaluation, as well as with its partner the UK AI Safety Institute. As part of this collaborations, these government partners have conducted capability evaluations on OpenAI's o1-preview and o1 models.

Frontier Model Forum: In July 2023, together with Anthropic, Google, and Microsoft, we announced the Frontier Model Forum, an industry body focused on advancing the safety and security of frontier AI models globally, and have since welcomed Meta and

Amazon. The Forum has since published on a range of technical best practices in safety and security, and helped develop AI safety public goods through the $10M AI Safety Fund.

User Awareness and Content Provenance

In May 2024, we joined the Steering Committee of Coalition for Content Provenance and Authenticity (C2PA), recognizing the societal value of a standardized approach to verifying digital content origins. As part of this effort, we are collaborating with industry partners to adopt, develop and promote an open standard that can help people verify the tools used for creating many kinds of digital content. Specifically, we've embedded C2PA metadata into all images generated by DALL·E 3 and videos produced by Sora, ensuring clear attribution. C2PA metadata retained by non-adversarial users serves as a trust signal, strengthening and fostering an online trust ecosystem. We detailed this research in a case study with the Partnership on AI and an August 2024 update. Beyond C2PA, we also incorporated audio watermarking into Voice Engine, our custom voice model, currently in research preview. To advance the science and develop stronger provenance techniques, we also allowed testing of our image detection tools through our Researcher Access Program.

Societal Risk and Impacts Research and Public Benefit Initiatives

Ensuring AI benefits everyone is central to our mission, and we support this through investments in scientific collaboration, community engagement, and ethical AI development. For example, OpenAI has partnered with institutions like Los Alamos National Laboratory to explore the safe use of AI in biosciences, leveraging AI's capabilities in complex scientific domains such as healthcare and quantum physics. The company also supported a hackathon focused on accelerating clean energy deployment, providing technical mentorship and resources to explore AI-driven solutions for sustainability challenges. Additionally, OpenAI launched OpenAI Academies to empower developers and organizations with AI tools.

We've also partnered with Common Sense Media to promote responsible AI usage among families and educators, offering educational materials and teacher training programs. Efforts to enhance AI accessibility include advancements in voice-enabled technology and collaborations with accessibility organizations to improve independence for users with disabilities. Late last year, we enabled wider access to ChatGPT through an experimental new launch of 1-800-ChatGPT.

OpenAI has established several funds and programs to jumpstart investment and support for research into societal resilience and social impact. These include, a societal resilience fund to drive adoption and understanding of provenance standards in partnership with Microsoft, a program focused on collecting and determining democratic methods to decide the rules that govern AI systems, and research on evaluating the performance of our models on first-person fairness. We're also proud to be Founding Partners of Robust Open Online Safety Tools (ROOST), a new non-profit entity designed to address the urgent need for accessible, high-quality Trust & Safety tools in the rapidly evolving digital landscape. OpenAI further fosters public engagement through case studies, forums, and storytelling initiatives that showcase AI's impact across various industries.