Report

Organization: **OpenAI**

www.openai.com

Publication date: Apr 22, 2025, 09:01 AM PDT

Reporting period: 2025

Section 1 - Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

As described in OpenAI's Preparedness Framework, OpenAI tracks several risk categories and classifies the levels of capability and post-mitigation risk in a scorecard. Models with a “High” capability score require safeguards that sufficiently minimize associated risks before they can be deployed, and models with a “Critical” capability score require safeguards that sufficiently minimize associated risks before they can be developed further.

Additionally, OpenAI’s safety systems teams undertake risk modeling to address a variety of potential risks and harmful content. Our organization classifies AI-related risks using our enterprise risk framework, which incorporates both general risk management principles (aligned with ISO 27001 and NIST) and AI-specific considerations drawn from our internal AI preparedness framework and associated readiness scorecards.

We categorize risks based on their type (e.g., compliance, security, reputational, operational), and evaluate them based on their likelihood and potential impact to users, the public, and OpenAI. These are assessed against defined risk thresholds to determine:

- Acceptable risks (within tolerance),
- Risks requiring mitigation, and
- Unacceptable risks (above tolerance and requiring redesign, escalation, or avoidance).

For AI systems, we pay particular attention to a variety of risks that may include, among other things:

- Model behavior: e.g., potential for misuse, unsafe outputs, or generation of false/misleading content
- Public safety and national security: e.g., ability to assist with cyber or chemical and biological attacks
- System oversight and control: risks arising from inadequate human-in-the-loop processes or unclear escalation paths
- Privacy and data use: e.g., training data sensitivity, inference risks, or regulatory exposure (GDPR, HIPAA, etc.)
- Fairness and bias: including differential performance across user groups or harmful stereotypes
- Transparency and explainability: how understandable the system is to users, developers, and stakeholders
- Deployment context and safeguards: ensuring model outputs are appropriate for the product setting and mitigated for known abuse vectors

- Third-party use and integration: risk of downstream misuse or unclear responsibilities in shared accountability models

We use structured scorecards to assess these categories at various stages of development and deployment, which helps us prioritize mitigation efforts and ensure alignment with both internal expectations and emerging legal standards.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

To identify risks across the lifecycle of advanced AI systems, including before deployment and placement on the market, our organization employs a multi-faceted approach:

Prior to deployment, OpenAI conducts a holistic assessment of potential risks that may stem from generative models. We use a combination of methods, spanning all stages of development across pre-training, post-training, product development, and policy. For example, during post-training, we align the model to human preferences; we red-team the resulting models and add product level mitigations such as monitoring and enforcement; and we provide moderation tools and transparency reports to our users. We also conduct evaluations to assess dual-use capabilities, including chemical, biological, cybersecurity, and model autonomy risks, as described in our Preparedness Framework. Evaluation methods include internal testing, red teaming, and external collaborations.

OpenAI continually invests in detecting misuse and emergent risks through deployment of classifiers, rules, content review systems, and manual and automated analysis. We take actions to respond to patterns of abuse or misuse as they emerge, including taking action within the platform to limit their impact, and incorporating lessons learned into model defenses and behavior to improve resiliency over time.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

OpenAI employs rigorous red-teaming processes to evaluate the performance of models and systems before deployment. OpenAI uses both human and automated red-teaming processes. Adversarial testing involves experts who attempt to exploit the model in ways that could cause harm or lead to undesirable outcomes, or who probe the model for unexpected or unwanted behavior in a variety of domains. This process is complemented by collaborations with external organizations, such as cybersecurity firms, to provide additional perspectives and expertise in identifying potential threats.

OpenAI's red teaming efforts leverage an external Red Teaming Network comprising a community of trusted and experienced experts that help inform our risk assessment and mitigation efforts. The Red Teaming Network draws on experts in a wide variety of domains, with diverse perspectives and lived experiences. A detailed description is available here: https://openai.com/index/red-teaming-network/

As an example of how we use red-teaming in practice, before launching GPT-4o we worked with more than 100 external red teamers, speaking a total of 45 different languages, and representing geographic backgrounds of 29 countries. Red teamers had access to various snapshots of the model at different stages of training and safety mitigation maturity for several months before deployment. A detailed description is available in the GPT-4o system card: https://openai.com/index/gpt-4o-system-card/.

Our model review process is also informed by results from testing carried out in collaboration with third party assessors. For example, we've worked with the US and UK AI Safety Institutes, and independent third party labs such as METR and Apollo to add an additional layer of validation for key risks. Where possible and relevant, we report on their findings in our systems cards, such as the scheming tests conducted by Apollo and autonomy and AI R&D tasks conducted by METR for o1 (https://openai.com/index/openai-o1-system-card/).

Red team assessments may also be carried out periodically, or due to underlying changes to infrastructure, application code, or in response to threat conditions. Responsible red teaming is permitted and encouraged by good-faith security researchers via our bug bounty program.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

We use a combination of quantitative (e.g., frequency of flagged outputs, output quality scores, performance deltas across demographics) and qualitative (e.g., SME judgment, ethical review outcomes, stakeholder feedback) metrics to assess AI-related risk. These are embedded in our AI readiness scorecards and risk evaluations. We describe some of these metrics, methodologies, and limitations in system cards and in research we publish, some of which can be found here: https://openai.com/news/research/

It is worthwhile to note that some risks (e.g., reputational or ethical harm) resist easy quantification, and we lean on expert judgment. AI system behavior can change over time (e.g., through fine-tuning or model evolution), so risk scores may need ongoing validation. Additionally, risk evaluations are context-sensitive — what is acceptable in one product setting may be high-risk in another. Risk evaluations are continually re-evaluated based on changes to the threat landscape, and in response to adverse, anomalous, or malicious activity observed on our platform.

OpenAI has mechanisms in place to receive reports of incidents and vulnerabilities from third parties. These mechanisms are part of OpenAI's commitment to safety and security, allowing external researchers, users, and other stakeholders to report issues that they may encounter.

For more detailed information, you can visit OpenAI's responsible disclosure page: https://openai.com/policies/coordinated-vulnerability-disclosure-policy/

OpenAI provides a model behavior feedback form (https://openai.com/form/model-behavior-feedback/) where users can submit reports when our models behave in unexpected or unwanted ways and maintains a bug bounty program through BugCrowd (https://bugcrowd.com/openai). In addition, OpenAI runs a Cybersecurity Grant Program to support research and development focused on protecting AI systems and infrastructure. This program encourages and funds initiatives that help identify and address vulnerabilities, ensuring the safe deployment of AI technologies.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

Yes. Please see response for, 1c) Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

OpenAI contributes to the work of the following standard development organizations:

- NIST's AISIC working groups and task forces on provenance, risk management, biorisk
- CSA's AI Standards
- CoSAI Security and Safety Standards
- C2PA provenance standards
- MLCommons safety evaluations
- ISO risk management
- FMF on frontier risk
- Coalition for Health AI

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

OpenAI collaborates with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks:

- Protecting children: A critical focus of our safety work is protecting children. We've built strong default guardrails and safety measures into our models to mitigate potential harms to children. We detect and remove child sexual abuse material (CSAM) from training data and report any confirmed CSAM to the relevant authorities like the National Center for Missing & Exploited Children (NCMEC) in the U.S. In 2024, we joined industry peers in committing to Thorn's Safety by Design principles (https://www.thorn.org/blog/generative-ai-principles/), which seeks to prioritize child safety at every stage in the development of AI. In 2025, we announced OpenAI as a founding partner of Robust Open Online Safety Tools (ROOST.tools), a community effort that will deliver free, open-source digital safety tools to public and private sectors globally, addressing critical gaps in online safety. We run Thorn's CSAM classifier to detect novel CSAM over all image uploads and generations, and we run a hash filter over all image uploads to catch known CSAM.
- Election integrity: We currently do not allow users to use our tools, including ChatGPT, for political campaigning. For example, ChatGPT is trained to refuse requests for targeted political persuasion. Universally we also disallow the categorization of individuals based on their biometric data to deduce or infer sensitive attributes such as political opinions. To prevent abuse, we don't allow users, builders on our API, or those creating shared GPTs to create tools that impersonate real people (e.g., candidates) or institutions (e.g., local government). We've also disrupted covert influence operations that sought to use our models in support of deceptive activity across the internet and continue to monitor and mitigate such abuses. We also have an opt-out process for public figures who don't want their likeness to be generated by our models. To improve transparency around AI-generated content, we implemented C2PA's digital credentials.
- Investment in impact assessment and policy analysis: Our impact assessment efforts have been widely influential in research, industry norms, and policy, including our work on measuring the chemical, biological, radiological, and nuclear (CBRN) risks associated with AI systems, and our research estimating the extent to which different occupations and industries might be impacted by language models. We also publish pioneering work on how society can best manage associated risks – for example, by working with external experts to assess the implications of language models for influence operations. (https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/; https://arxiv.org/abs/2303.10130)
- Partnering with governments: We partner with governments around the world to inform the development of effective and adaptable AI safety policies. This includes showing our work and sharing our learnings, collaborating to pilot government and other third party assurance, and informing the public debate over new standards and laws. (https://openai.com/global-affairs/our-approach-to-frontier-risk/) For example, in August 2024 we entered into voluntary agreements with the U.S. and UK AISIs to enable formal collaboration on AI safety research, testing and evaluation. As mentioned in the section on testing and red teaming, we partner with third party independent labs, academics, experts, and more to add an additional layer of validation for key risks.

Section 2 - Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

We address this question in part in section 1 and 2 of this report.

Evaluations are performed at multiple checkpoints throughout the AI lifecycle to ensure models are safe, effective, and align with OpenAI's objectives. In the pre-development phase, we begin with an initial risk assessment to identify potential hazards and define safety objectives, establishing a foundation for informed and responsible development. During the development phase, models undergo iterative testing to refine safety and performance as the model evolves. Before deployment, in the pre-deployment phase, we conduct a range of safety evaluations, as described in our Preparedness Framework, and to address other product-specific risks. In the deployment phase, we implement monitoring to detect unexpected behaviors or misuse, and we perform post-deployment evaluations using real-world feedback and performance data to identify emerging risks. Finally, in the post-deployment and maintenance phase, we update our evaluations based on new misuse patterns and advances in research.

We also run evaluations at multiple stages throughout the AI lifecycle to ensure the security and resilience of our systems. From the earliest phases of infrastructure and model development, we embed security directly into our design process, enabling proactive risk mitigation. As our models are deployed, we use AI-powered defenses to continuously monitor for and respond to threats in real time, effectively evaluating system behavior and integrity on an ongoing basis. We also engage in continuous adversarial red teaming with expert partners like SpecterOps, simulating realistic attacks across our corporate, cloud, and production environments to uncover vulnerabilities and improve our detection and response capabilities. For advanced AI agents, we conduct real-time evaluations through unified monitoring pipelines that track and enforce safe behavior. Additionally, in response to real-world threats—such as spear phishing attempts—we evaluate and adapt our defenses while sharing insights with other AI labs to strengthen collective security. More on our practices can be found here: https://openai.com/index/security-on-the-path-to-agi

b. How do testing measures inform actions to address identified risks?

Testing is a core part of our AI risk management process and directly informs mitigation decisions at multiple stages of the system lifecycle. Specifically:

We conduct pre-deployment evaluations using our internal AI preparedness framework, which includes structured scorecards that assess model behavior for the most significant safety risks. We also conduct other evaluations to assess product-specific risks, fairness, and other risk dimensions that AI presents. These evaluations surface risks that inform control design, model tuning, or deployment gating decisions.

Red teaming and adversarial testing are used to uncover failure modes, misuse potential, or unexpected behaviors. These findings often lead to downstream mitigation efforts — such as refining instructions, retraining, or introducing product-level safeguards.

Post-deployment, we rely on monitoring, user feedback, periodic security evaluations, and incident reporting to track risk emergence over time. Material findings can be fed back into model update cycles, safety classifiers, or policy adjustments.

The results of these testing activities are documented and reviewed by relevant stakeholders (e.g., model developers, product teams, risk reviewers), and actions are prioritized based on impact, feasibility, and alignment with risk thresholds.

c. When does testing take place in secure environments, if at all, and if it does, how?

Yes — testing occurs in environments which meet high security standards. We conduct certain types of testing in these environments, particularly when:

- Pre-release models are being evaluated that have not yet been exposed to a broader internal or external audience
- The testing involves sensitive data, such as simulated PII or partner-provided examples
- Red team exercises require a controlled environment to avoid leakage or unintended exposure

These environments are secured according to our internal standards, including restricted access to approved and authorized users, audit logging, and monitoring for security purposes. These environments support both automated evaluations and manual review workflows.

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

Prior to deployment, OpenAI assesses and mitigates potential risks that may stem from generative models, such as information harms, bias and discrimination, or other content that violates our safety policies. We use a combination of methods, spanning all stages of development across pre-training, post-training, product development, and policy. For example, during post-training, we align the model to human preferences; we red team the resulting models and add product-level mitigations such as monitoring and enforcement; and we provide moderation tools and transparency reports to our users.

We find that the majority of effective testing and mitigations are done after the pre-training stage because filtering pre-trained data alone cannot address nuanced and context-specific harms. At the same time, certain pre-training filtering mitigations can provide an additional layer of defense that, along with other safety mitigations, help exclude unwanted and harmful information from our datasets:

- We use our Moderation API and safety classifiers to filter out data that could contribute to harmful content or information hazards, including CSAM, hateful content, and violence.
- For image generation systems, we filter our image generation datasets for explicit content such as graphic sexual material and CSAM.
- We use advanced data filtering processes to reduce personal information from training data.

e. How does your organization protect intellectual property, including copyright-protected content?

We think about intellectual property throughout the AI lifecycle and deploy techniques at different stages to address intellectual property concerns, including copyright:

- Data collection: When collecting and curating data for pre-training, OpenAI filters out domains that we have identified as hosting pirated content, primarily aggregate personally identifiable information, or have content that otherwise violates our policies. This helps avoid exposing the model to unauthorized copyright-protected content. We also filter domains whose owners have opted-out of training our generative models as described further below.,
- Data deduplication: We take additional steps during training, such as removing duplicate copies of data, to prevent the model from seeing specific phrases, passages, or documents multiple times. This reduces the likelihood that the model will memorize or reproduce its training data without meaningful transformation.
- Preference signals for web content: We created a method for website owners to indicate preferences for how OpenAI uses their content. The method uses an existing industry standard, the Robots Exclusion Protocol (sometimes referred to as robots.txt), and allows website owners to indicate separate preferences for generative model training and their site being surfaced in response to user searches.
- Mitigating outputs: We take a variety of steps to prevent our models from outputting intellectual property beyond legal bounds. For example, we train our models to refuse requests to recite or generate copyrighted material. Additionally, we use output filters designed to block such material from being generated.
- Continuous evaluation and mitigation: OpenAI continuously assesses and mitigates risks related to copyright through various stages of model development. For example, we discuss copyright mitigations as part of our safety work in our recent system card for GPT-4o.

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

We want our AI models to learn about the world—not private individuals. We use training information to help our AI models, like those powering ChatGPT, to learn about language and how to understand and respond to it. We do not actively seek out personal information to train our models. Beyond training, we use personal data as described in our privacy policy.

We take steps at each stage of training and deploying our models to reduce the processing of personal information. For example, we apply filters and remove information from our training data that we do not want our models to learn from or output, such as hate speech, adult content, sites that primarily aggregate personal information, and spam. We then further train our models to reject requests for private or sensitive information, even where that information is available on the public web. We conduct a series of evaluations on our models to help ensure they align with these standards, implement robust usage policies to protect privacy, and monitor and respond to attempts to generate information in violation of these policies. In short, we implement a series of reinforcing privacy protections to train our models from the ground up to protect private and sensitive information.

We additionally give users of ChatGPT robust and easy-to-use data controls, including to review and delete their data, turn off features like ChatGPT memory, and to easily opt out of content being used to further improve our models. We also offer a Temporary Chat mode that allows users to have conversations with ChatGPT that are automatically deleted after 30 days. And when using ChatGPT data to improve our models, we take a number of privacy-protective steps to reduce the processing of any incidental personal information.

We're also increasingly relying on our AI models to enhance privacy, such as to improve the filtering of personal information and augment the use of synthetic data, and are excited about the possibilities presented by this research.

Also, please see response for, 2d) What measures does your organization take to promote data quality and mitigate harmful biases throughout the AI lifecycle, including in training and data collection processes?

g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- **i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?**
- **ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?**
- **iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?**
- **iv. How often are security measures reviewed?**
- **v. Does your organization have an insider threat detection program?**

OpenAI maintains a comprehensive cybersecurity, physical security, and insider threat program governed by an enterprise risk management process and overseen by the Safety and Security Committee (SSC) on our Board of Directors. Our approach integrates industry best practices with AI-specific innovations to manage the unique threats facing frontier AI systems. These practices are aligned with recognized frameworks such as ISO 27001, SOC 2, NIST SP 800-53, and FedRAMP.

We have published six key measures for securing advanced AI infrastructure that complement traditional cybersecurity practices:

1. Trusted computing for AI accelerators
2. Network and tenant isolation guarantees
3. Innovation in operational and physical security for datacenters
4. AI-specific audit and compliance programs
5. AI for cyber defense
6. Resilience, redundancy, and research

These are detailed in our public blog: Reimagining Secure Infrastructure for Advanced AI

(https://openai.com/index/reimagining-secure-infrastructure-for-advanced-ai).

We continuously assess cybersecurity risks across our systems and infrastructure. Our security controls are tailored to the risks of increasingly capable AI systems, and we adapt as we move closer to AGI. Security safeguards are reviewed regularly and refined through adversarial testing, red teaming, and ongoing research. Security controls and safeguards are also informed by changes in the threat landscape.

OpenAI treats unreleased model weights as core intellectual property. We enforce strict access controls, granting access only to individuals whose roles require it. These protections are supported by a robust insider threat detection program and secure research environments. Model weights and sensitive assets are handled exclusively within appropriately secured infrastructure designed to prevent unauthorized access or leakage. Technical controls in these environments include advanced physical security (for datacenters), cryptographic protections, real-time security monitoring, defense-in-depth network architecture, and supply chain security controls. For additional technical detail, see our public blog referenced above.

OpenAI maintains a mature vulnerability management lifecycle that includes:

1. Continuous vulnerability scanning
2. Industry threat intelligence monitoring
3. Third-party audits and assessments
4. Internal and external adversarial red teaming
5. Ongoing bug bounty programs
6. Annual independent audits of security practices

We act promptly to remediate identified risks and vulnerabilities and collaborate with partners and stakeholders when necessary. These risks are tracked through an enterprise risk management program with executive and board level reporting and oversight.

Security controls are reviewed continuously and iteratively—through ongoing red teaming, penetration testing, operational monitoring, third-party audits, and program-level reassessments. OpenAI engages in an independent third party audit on at least an annual basis which reviews our security measures, including our vulnerability management process.

OpenAI has a dedicated insider threat team. The program emphasizes prevention, detection, and response. This includes strong policy enforcement, employee education, access controls, monitoring mechanisms to detect anomalous or high-risk activity, and investigative capabilities.

h. How does your organization address vulnerabilities, incidents, emerging risks?

Identified risks are documented in our risk register or GRC platform, assigned to accountable owners, and tracked through to resolution. OpenAI disallows the use of our models and tools for certain activities and content, as outlined in our terms and usage policies. These policies are designed to prohibit the use of our models and tools in ways that cause individual or societal harm. We periodically update these policies in response to new risks and new information on how our models are being used. Access to and use of our models are also subject to OpenAI's Terms of Use and our Business Terms. We employ a combination of guardrails, tools, policies, and processes to detect and enforce against abuse of our models. In addition to usage policies, we employ automated systems (for example, our Moderation API) and AI-powered tools to detect and prevent harmful content, human reviewers and expert investigators to conduct detailed investigations of model behavior and platform abuse, and external input to help our teams iterate on and regularly improve our moderation, detection, disruption, and enforcement capabilities. We publish information related to our post-deployment monitoring, mitigation and enforcement activities as described in response to questions 3b and 4g below.

Addressing vulnerabilities, incidents, emerging risks and misuse often requires participation from a number of teams across the company, including where applicable, teams responsible for research; safety, security, and alignment; product; user or customer relationships and support; intelligence, investigations and teams supporting scaled detection and enforcement; security; policy teams; global affairs; legal and privacy; and communications.

Section 3 - Transparency reporting on advanced AI systems

a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- **i. How often are such reports usually updated?**
- **ii. How are new significant releases reflected in such reports?**
- **iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.**

We publish system cards for our frontier models and other significant launches, including GPT-4 and GPT-4o, GPT-4o Image Generation, GPT-4.5, Sora, o1, Operator and Deep Research. These cards provide an in-depth look at the models' capabilities, limitations, and safety evaluations. The system cards also discuss the domains of appropriate and inappropriate use.

We typically release system cards for new significant releases that advance the frontier of capabilities or that present novel net-new risks, and cover various risks described above such as safety, security, and societal risks.

OpenAI demonstrates transparency regarding its advanced AI systems through a variety of technical documentation, policies, and instructions that are publicly accessible. These resources include:

- Model Specifications: OpenAI publishes detailed documentation, such as the Model Spec, which outlines the desired behavior for its models. These documents serve as a guide for researchers and developers, providing transparency on how OpenAI shapes the behavior of its AI systems. It includes a set of core objectives, as well as guidance on how to deal with conflicting objectives or instructions. (https://model-spec.openai.com/2025-02-12.html)

- Usage Policies: OpenAI has established comprehensive usage policies to ensure that its technology is used responsibly. These policies cover a range of topics, including compliance with applicable laws, safeguarding privacy, and preventing misuse of AI for harmful activities. The policies are designed to provide flexibility for innovation while maintaining safety and ethical standards. (https://openai.com/policies/usage-policies/)

b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

OpenAI actively shares information with other stakeholders regarding the outcome of risk evaluations for advanced AI systems. Please see response in, 1h) How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

In addition, OpenAI both publishes certain insights derived from its detection and disruption of malicious use of its platform, and also shares certain insights with industry peers, industry groups, and relevant authorities to enhance our collective ability to identify and mitigate risks.

Please view our posts for more information:

- Disrupting malicious uses of AI by state-affiliated threat actors: https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/
- Disrupting deceptive uses of AI by covert influence operations https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/
- Disrupting a covert Iranian influence operation: https://openai.com/index/disrupting-a-covert-iranian-influence-operation/
- Influence and Cyber Operations: an update: https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf
- Disrupting malicious uses of our models: an update February 2025: https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf
- Interpretability research: https://openai.com/index/extracting-concepts-from-gpt-4/
- Deliberative alignment: https://openai.com/index/deliberative-alignment/
- Biorisk study: https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/

c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

OpenAI discloses privacy policies that address the use of personal data, user prompts, and the outputs of advanced AI systems. These policies outline how data is collected, stored, and used, emphasizing user privacy and data security. Please see the documentation referenced below.

- OpenAI's Privacy Policy: https://openai.com/policies/privacy-policy/
- OpenAI's Data processing addendum: https://openai.com/policies/data-processing-addendum/
- OpenAI's Business terms: https://openai.com/policies/business-terms/
- How ChatGPT and our language models are developed: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed
- How your data is used to improve model performance: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance
- For users to submit privacy requests: https://privacy.openai.com/policies
- Data Controls FAQ: https://help.openai.com/en/articles/7730893-data-controls-faq
- Trust Portal: https://trust.openai.com/

d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

We provide general information about our training data in our System Cards, which includes appropriate information about the sources of our data. That information is disclosed at a high level to balance the value of public disclosures with trade secret sensitivities.

e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

This question has been answered comprehensively in section 3 (a-d).

Any further comments and for implementation documentation

Link to our System Cards:

- DALLE-2: https://github.com/openai/dalle-2-preview/blob/main/system-card.md
- DALLE-3: https://openai.com/index/dall-e-3-system-card/
- GPT-4: https://cdn.openai.com/papers/gpt-4-system-card.pdf
- GPT-4V: https://openai.com/index/gpt-4v-system-card/
- GPT-4o: https://openai.com/index/gpt-4o-system-card/
- o1: https://openai.com/index/openai-o1-system-card/
- Sora: https://openai.com/index/sora-system-card/
- Operator: https://openai.com/index/operator-system-card/
- Deep research: https://openai.com/index/deep-research-system-card/
- o3-mini: https://openai.com/index/o3-mini-system-card/
- GPT-4.5: https://openai.com/index/gpt-4-5-system-card/
- Addendum to GPT-4o - Image Generation: https://openai.com/index/gpt-4o-image-generation-system-card-addendum/

Section 4 - Organizational governance, incident management and transparency

a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

OpenAI has implemented several AI risk management and governance policies, as described in System Cards, our Preparedness Framework and our Safety Center (https://openai.com/safety/). For example, before deploying frontier models and other significant launches, safety evaluations and mitigations are reviewed by an internal cross-functional group of senior company leaders, with recommendations to management. Additionally, the Safety and Security Committee of the Company's Board of Directors has oversight of the Company's risk management decisions, and, in certain circumstances, launches are reviewed by a Deployment Safety Board established among OpenAI and Microsoft.

OpenAI regularly iterates on other policies and procedures to assess, measure, and mitigate risks throughout the development and deployment lifecycle.

b. Are relevant staff trained on your organization’s governance policies and risk management practices? If so, how?

Yes. We provide role-specific training to ensure relevant employees understand their responsibilities under our governance and risk frameworks. This includes:

- Onboarding training for new hires, covering security, privacy, and acceptable use
- Annual compliance refreshers for all staff
- Guidance documents and FAQs that translate policy into actionable steps, especially for product development and AI deployment contexts

Governance-related training content is regularly updated to reflect policy changes and evolving external expectations.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Yes — we communicate selected risk management and governance practices publicly through:

- Our public policies (https://openai.com/policies), including the Security, Privacy, and Acceptable Use policies
- System cards and model evaluations that explain how risks are identified and mitigated
- Transparency reports, blog posts, and safety updates covering AI deployment practices and changes to models or safeguards
- Engagement with industry forums and standards bodies to contribute to evolving best practices
- Where feasible and appropriate, we publish risk-management policies such as our Preparedness Framework.

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

Yes. We log and track significant reported incidents through our internal incident response processes, which include:

- Triage and investigation logs
- Post-incident reviews (including root cause analysis and corrective actions)
- Lessons learned documentation shared with relevant internal teams
- Incident records are retained and reviewed as part of internal audits and ISO 27001 / SOC 2 evidence gathering.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

We share information through multiple channels, depending on severity and stakeholder needs:

Internally:

- With relevant security, privacy, legal, and product teams to inform mitigation and product adjustments

Externally:

- With partners or customers via direct communication channels (e.g., through Enterprise support)
- Through coordinated disclosure practices for vulnerabilities
- Occasionally via public blog posts or transparency statements (e.g., postmortems or model updates)

We also participate in information sharing with industry groups, civil society partners, and AI research forums when appropriate.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

Yes. When appropriate, we share information about AI-related incidents through:

- Direct stakeholder communication (e.g., with enterprise customers, partners, or researchers)
- Regulatory or legal disclosures, where required
- Public posts or transparency disclosures for incidents with broader impact, such as major system updates or safety concerns
- Participation in voluntary initiatives (e.g., red team result sharing or post-release evaluations with research partners)

We assess the appropriate level of disclosure, balancing transparency with security and confidentiality obligations.

g. How does your organization share research and best practices on addressing or managing risk?

- Protecting Children and Vulnerable Groups: OpenAI has partnered with organizations such as Thorn and the Tech Coalition to detect, review, and report child sexual abuse material (CSAM). By adopting comprehensive Safety by Design principles, OpenAI and our peers are ensuring that child safety is prioritized at every stage in the development of AI. To date, we have made significant effort to minimize the potential for our models to generate content that harms children, set age restrictions for ChatGPT, and actively engage with the National Center for Missing and Exploited Children (NCMEC), Tech Coalition, and other government and industry stakeholders on child protection issues and enhancements to reporting mechanisms. (https://openai.com/index/child-safety-adopting-sbd-principles/). In addition, OpenAI has partnered with Common Sense Media to collaborate on AI guidelines and education materials for parents, educators and young people, as well as a curation of family-friendly GPTs in the GPT Store based on Common Sense ratings and standards. See the Common Sense Media and OpenAI Partnership for more details: https://www.commonsensemedia.org/press-releases/common-sense-media-and-openai-partner-to-help-teens-and-families
- Upholding Democratic Values and Respecting Human Rights: OpenAI conducts research on the societal impacts of AI, focusing on ensuring that AI aligns with democratic values and human rights. This includes studying how AI can be governed to reflect democratic input and avoid exacerbating inequalities. (https://openai.com/index/democratic-inputs-to-ai/)
- Avoiding Harmful Bias: OpenAI conducts and has published research to identify and mitigate bias in AI systems. This includes evaluations to prevent the reinforcement of social biases and stereotypes, and developing methods to improve fairness in AI outputs. This is published in OpenAI's System Cards, and we recently published a research paper on fairness (https://openai.com/index/evaluating-fairness-in-chatgpt/)
- Content Provenance and Abuse Prevention: OpenAI is actively involved in the development of content provenance standards, such as C2PA, to help track the origin of digital content and prevent the spread of misinformation. This effort is part of broader initiatives to enhance the integrity and transparency of AI-generated content. See responses in Section 5 related to Content Authentication and Provenance Mechanisms. In addition, OpenAI is committed to enforcing policies that prevent abuse and to improve transparency around AI-generated content, as outlined by recent reports related to disrupting deceptive uses of AI by covert influence operations:
- https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/
- https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/
- https://openai.com/index/disrupting-a-covert-iranian-influence-operation/
- https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf
- https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf
- For more information on our safety practices see:
- https://openai.com/safety/
- https://openai.com/safety/how-we-think-about-safety-alignment/

h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes. Our governance and risk management programs align with international frameworks and standards including:

- ISO/IEC 27001 and 27701 (Information security and privacy management)
- SOC 2 (Trust Services Criteria, including Security and Confidentiality)
- NIST AI Risk Management Framework
- OECD and EU AI guidelines (including definitions of risk levels and governance expectations)
- ISO/IEC 42001 (AI Management Systems) — currently under review for future alignment

These standards inform both the design of controls and the structure of our governance documentation and training programs.

Section 5 - Content authentication & provenance mechanisms

a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

Please see our answer to question 5b).

Additionally, our usage policies require developers building on our platform to ensure that automated systems (e.g., chatbots) disclose to people that they are interacting with AI, unless it’s obvious from the context.

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

OpenAI is working to ensure the authenticity and transparency of digital content. Here our key efforts:

- C2PA: OpenAI is a steering committee member of the Coalition for Content Provenance and Authenticity (C2PA). OpenAI adds content credentials into images and videos generated by DALL·E 3, GPT-4o Image Generation, and Sora. Content credentials (metadata cryptographically signed to provide an attestation of the content's origins) provide information about how and when a piece of content was created or modified by our platform. Content credentials are metadata-based approaches to content provenance that require adoption by different players across the AI ecosystem to build trust in media online. We believe that C2PA is working to build such an ecosystem, and from our steering committee seat, we are helping to advance this work.
- Tamper-Resistant Watermarking: OpenAI has implemented watermarking to trace the origin of audio generated by Voice Engine, OpenAI's custom voice offering. These watermarks are invisible signals that are hard to remove, helping verify the authenticity of digital content. We have also developed visible watermarks for Sora, our text-to-video generation model.
- Societal Resilience Fund: OpenAI and Microsoft have launched a $2 million fund to support AI education and understanding. This fund supports organizations like AARP, International IDEA, and the Partnership on AI
- NIST 2.1 Taskforce: OpenAI has joined the NIST 2.1 Taskforce on Synthetic Content to help develop and contribute to global content provenance standards.

Section 6 - Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

We advance the state of AI safety and trustworthiness through a combination of internal research, external funding and partnerships, and product integration of emerging insights. Our efforts span across key domains such as robustness, disinformation, fairness, and transparency. We publish our research and make it available at https://openai.com/research.

1. Internal Research and Applied Safety Work

We invest heavily in internal research focused on the safe deployment of frontier models, including:

- Robustness and Red Teaming: Dedicated teams probe models for failure modes, including adversarial behavior, prompt injection, and content safety risks (see: https://openai.com/index/red-teaming-network/).
- Transparency and Interpretability: Research includes techniques like model probing, behavioral evaluation, and feature attribution to improve transparency (see for example: https://openai.com/index/language-models-can-explain-neurons-in-language-models/ and https://openai.com/index/extracting-concepts-from-gpt-4/).
- Bias and Fairness: We conduct evaluations on demographic disparities and systemic bias, with findings informing model development and moderation systems (see for example: https://openai.com/index/evaluating-fairness-in-chatgpt/).
- Misuse and abuse prevention: We continuously study how models might be misused to generate harmful content, or violate our usage policies, and implement corresponding safeguards or disrupt such activity (see our reports on disrupting influence operations detailed in answer 4g).

Research findings are shared via system cards, peer-reviewed papers, and updates to our model usage policies.

2. External Research Funding and Collaboration

We support external work through:

- Grants and fellowships to researchers and institutions advancing AI safety, fairness, and interpretability.
- Partnerships with academic labs and nonprofits to co-develop methods, benchmarks, and evaluations.
- Challenge programs (e.g., for red teaming or adversarial testing) that incentivize broader participation in AI safety research.

- These efforts promote shared learning and raise the industry bar on responsible AI development.

Across all of this, our goal is to continuously improve the trustworthiness and safety of our AI systems, and to share lessons with the broader ecosystem when possible.

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

See response to question 5b).

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

See response to question 5b).

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

We have engaged in a wide range of academic discussions and joint research projects, including looking at the potential economic and social implications of advanced AI systems. While much of this work is in preliminary stages, our view is that robust engagement with a wide range of subject-matter experts, including independent academics, think tanks, and research organizations, will further our understanding of how AI can be used to benefit humanity, in line with our company's mission.

Since joining OpenAI in November 2024, our new Chief Economist has established new workstreams that will answer rigorous and relevant questions about the economics of AI including research on the economic effects of AI on individuals, organizations, and communities.

Section 7 - Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

See response from question 6d).

Recently we launched OpenAI Academy, a free, public online resource hub designed to improve AI literacy and provide tools, best practices, and insights for people from all backgrounds to use AI more effectively and responsibly (https://academy.openai.com/).

Through a series of Economic Tabletop Exercises across the globe, OpenAI worked to help policymakers envision how they can deploy the economic benefits of AI to advance their own economic goals, deliver better public services, and support vulnerable populations. We conducted these exercises in the US, France, Brussels, the UK, and India and plan to continue engaging in other parts of the world throughout the year.

We have also released a US and a EU Economic Blueprint that lay out our policy proposals for capitalizing on the benefits of AI and driving economic growth across communities. This is a program that we are working to adapt and roll out in other countries.

b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

OpenAI prioritizes education and training to help people understand the nature, capabilities, limitations, and impact of advanced AI systems. We recently launched the OpenAI academy, an AI literacy and education platform including in-person and virtual events and resources, designed to help people of all backgrounds unlock the benefits of AI. Additionally, OpenAI collaborates with Common Sense Media to provide resources that promote responsible AI use among families, ensuring that users of all ages understand the ethical and social dimensions of AI.

- OpenAI Academy: academy.openai.com
- Common Sense Media: https://www.commonsensemedia.org/press-releases/common-sense-media-and-openai-partner-to-help-teens-and-families
- Teaching with AI: https://openai.com/index/teaching-with-ai/
- OpenAI Educator FAQ: https://help.openai.com/en/collections/5929286-educator-faq
- Microsoft and OpenAI launch Societal Resilience Fund: https://blogs.microsoft.com/on-the-issues/2024/05/07/societal-resilience-fund-open-ai/

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

OpenAI’s charter aligns closely with the UN Sustainable Development Goals. As it stands we focus on:

- Infrastructure: Our environmental impact primarily depends on the energy and water use of our data centers. Azure is our primary partner here Azure has committed to being carbon negative by 2030, using 100% renewable energy by 2025, and achieving water positivity, zero waste, and net-zero deforestation by 2030 (https://www.microsoft.com/en-us/corporate-responsibility/sustainability)
- Organization Level Initiatives: We are in progress on industry scorecards (ISO 14001 and 26000; and OHSAS 18001/ISO 45001), are actively working on the related disclosures (CSRD, SFDR, SECR, TCFD) and are focused on research that could result lowering the carbon footprint of ML training.
- Projects: OpenAI is supporting the work of Turn.io with various climate and social impact organizations and GitLab Foundation's AI for Economic Opportunity (https://www.gitlabfoundation.org/our-journey/openai). OpenAI has also hosted an AI Hackathon to accelerate clean energy development as part of a broader effort to help make the US energy grid more sustainable and resilient. (https://www.businesswire.com/news/home/20240628518463/en/Crusoe-Lowercarbon-to-Host-AI-Hackathon-to-Accelerate-Clean-Energy-Development-OpenAI-Offering-Credits-and-Mentoring-Hackers-Department-of-Energy-to-Speak-at-Public-Workshop)

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

OpenAI's UN General Assembly (UNGA) activation in 2024 brought together global governments, civil society, and our nonprofit partners (Kytabu, Digital Green, Visao Coop, and 10bedicu.org) to discuss progress toward addressing the UN Sustainable Development Goals. Our CEO Sam Altman launched our OpenAI Academy program at UNGA to support global organizations advancing development and hosted fireside chats with the UN Deputy Secretary General and the UN Tech Envoy on addressing the world's pressing challenges with AI. At UNGA we also brought together representatives from USAID, UNICEF, UNDP, UNESCO, the Center for Global Development, International Rescue Committee, along with the governments of Togo, Nigeria, and Iceland, and our nonprofit partners to discuss the importance of preserving low-resource languages and the power of AI tools to help people solve hard problems being addressed by the UN.

While we may be the ones building the technology, the real-world applications are emerging from the ingenuity of our users. The initiatives below demonstrates AI's groundbreaking potential:

In India the World Bank estimates that there are 0.7 physicians and 1.7 nurses for every 1,000 people. 10BedICU (https://10bedicu.org/) is addressing this shortage by developing a number of GPT-4 powered tools, including an automated discharge summary generator, that frees up medical providers' time to focus on patient care. They are also using GPT-4 Vision to sync outdated monitors with their digital EMR system, increasing the accuracy of patient records.

Digital Green (https://digitalgreen.org/) is addressing the shortage of agricultural extension workers with Farmer.Chat (https://farmerchat.digitalgreen.org/), an AI assistant that delivers agricultural expertise tailored to a specific community and context. Digital Green is working with India's Ministry of Agriculture to deploy Farmer.Chat across 12 Indian states. This will enable extension workers to reach farmers with more relevant advice, ultimately increasing farmer income.

In Kenya, Kytabu (https://kytabu.africa/) is currently using GPT-4 to pilot Somanasi (https://kytabu.africa/somanasi/), a supplementary learning app integrated with the Kenyan curriculum that helps students reinforce classroom learning at their own pace with engaging, gamified tools. This is particularly helpful for students with learning disabilities, who can dig deeper into classroom content without fear of embarrassment. Somanasi is currently being tested in ten schools, with plans to eventually make it accessible to all 33,000 (https://www.education.go.ke/) primary and secondary schools in Kenya.

In Sierra Leone and Ghana, Rising Academies (https://www.risingacademies.com/) has built Rori (https://rori.ai/), a WhatsApp-based math tutor designed to support distance learning. Rori helps low-literacy, low-income students build their foundational math skills by making learning fun.

These initiatives are a small sample of the myriad of ways we are seeing people around the world deploy our technology to improve lives in their communities. The experts who develop these applications to meet local needs are integral to our efforts to build these technologies in ways that can benefit everyone.

OpenAI is also directly investing in innovative AI-powered solutions in service of the SDGs through its AI for Global Development accelerator program, in partnership with the Center for Global Development and the Agency Fund. We are providing technical advisory, cash funding, and API credits to eight nonprofits working in education, health, and agriculture across India, Kenya, South Africa, Botswana, Bangladesh, Indonesia, and Ethiopia. These include Rocket Learning, which provides personalized learning pathways for parents and daycare workers to provide early childhood education in India; Noora Health, which equips caregivers with knowledge to support the health of family members in India; and Jacaranda Health, which offers personalized prenatal health advice in local languages in Kenya.

We recognize that there is still significant work to be done to develop AI ecosystems that foster broadly distributed benefits of these technologies in the long-term. Significant resourcing will be needed to strengthen AI ecosystems, such as deepening low-resource language datasets, researching and adapting to potential workforce impacts, and rapidly expanding AI literacy.

More stories of our partnerships below:

- Digital Green: https://openai.com/index/digital-green/
- Be My Eyes: https://openai.com/index/be-my-eyes/
- Khan Academy: https://openai.com/index/khan-academy/
- 10BedICU: https://openai.com/index/10bedicu/
- AI for Global Development Accelerator: https://theagencyfund.substack.com/p/ai-for-global-development-accelerator